UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escola Tècnica Superior d'Enginyeria Informàtica

Recuperació d'informació basada en representacions vectorials denses

Treball Fi de Grau

Grau en Enginyeria Informàtica

AUTOR/A: Casamayor Segarra, Andreu

Tutor/a: Sanchís Arnal, Emilio

Cotutor/a: Hurtado Oliver, Lluis Felip

CURS ACADÈMIC: 2021/2022

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf

Escola Tècnica Superior d'Enginyeria Informàtica

Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

Recuperació d'informació basada en representacions vectorials denses

TREBALL FI DE GRAU

Grau en Enginyeria Informàtica

Autor: Casamayor Segarra, Andreu

Tutor: Sanchis Arnal, Emilio
Hurtado Oliver, Lluís Felip

Curs 2021-2022

Dedicatoria

A les persones més properes i que més m'han ajudat: a ma mare Encarna, a mon pare Juan Carlos, a la meua germana Paula i a la meua parella Anna.

Agraïments

Agrair als meus tutors Emilio i Lluís per la seua inestimable ajuda, consell i guia, que m'han brindat al llarg del desenvolupament, i que sense ella haguera sigut impossible portar a terme aquest projecte.

Resum

Hui en dia vivim en un món que ens sobrealimenta d'informació, dificultant el procés de trobar el documents, articles, etc. que estàs cercant. Els sistemes de recuperació d'informació (SRI) resolen el problema anterior per mitjà d'algoritmes que retornen una informació precisa i similar a la que estàs cercat. Tanmateix, cada volta hi ha mes informació, necessitant nous models i algoritmes eficients que siguen capaços d'administrar tantes dades i tornar resultats precisos. Aquest problema s'accentua quan parlem d'idiomes minoritaris.

En aquest projecte desenvolupem tres models SRI: un model clàssic (Booleà), un model més actual basat en representacions vectorials denses (Word2Vec), i el model més modern (STSB) basat en representacions vectorials denses contextuals, per a un idioma minoritari com és el català. Usant un corpus del grup d'investigació ELIRF-VRAIN (DACSA) i eines com *SpaCy, HuggingFace, SentenceTransformer i Whoosh* que ens proporcionen models preentrenats per altres grups d'investigació i el model STSB en català que ha sigut creat per nosaltres, hem desenvolupat els tres sistemes presentats en el projecte. Hem obtingut uns resultats satisfactoris segons els objectius marcats. Trobem que el model STSB millora la cerca d'articles, ja que manté la semàntica de la cerca i té en compte el context. En conclusió, observem l'evolució dels SRI en aquest projecte comparant vells models amb els models actuals.

Paraules clau: Sistemes de recuperació d'informació; representacións vectorials denses; embeddings no contextuals, embeddings contextuals; Word2Vec; Sentence to BERT

Resumen

Hoy en día vivimos en un mundo que nos sobrealimenta de información, dificultando el proceso de encontrar documentos, artículos, etc. que estás buscando. Los sistemas de recuperación de información (SRI) resuelven el problema anterior por medio de algoritmos que devuelven una información precisa y similar a la que estás buscado. Sin embargo, cada vez hay más información, necesitando nuevos modelos y algoritmos eficientes que sean capaces de administrar tantos datos y devolver resultados precisos. Este problema se acentúa cuando hablamos de idiomas minoritarios.

En este proyecto desarrollamos tres modelos SRI: un modelo clásico (Booleano), un modelo más actual basado en representaciones vectoriales densas (Word2Vec), y el modelo más moderno (STSB) basado en representaciones vectoriales densas contextuales, para un idioma minoritario como es el Catalán. Usando un corpus del grupo de investigación ELIRF-VRAIN (DACSA) y herramientas como *SpaCy, HuggingFace, SentenceTransformer y Whoosh* que nos proporcionan modelos preentrenados por otros grupos de investigación y el modelo STSB en Catalán que ha sido creado por nosotros, hemos desarrollado los tres sistemas presentados en el proyecto. Hemos obtenido resultados satisfactorios según los objetivos marcados. Encontramos que el modelo STSB mejora la búsqueda de artículos, puesto que mantiene la semántica de la búsqueda y tiene en cuenta el contexto. En conclusión, observamos la evolución de los SRI en este proyecto comparando viejos modelos con los modelos actuales.

Palabras clave: Sistemas de recuperación de información; representaciones vectoriales densas; embeddings no contextuales; embeddings contextuales; Word2Vec; Sentence to BERT

Abstract

Today we live in a world that overfeeds us with information, making it difficult to find the documents, articles, etc. that you are looking for. Information retrieval systems (IRS) solve the above problem by means of algorithms that return accurate information similar to what you are looking for. However, there is more and more information, requiring new models and efficient algorithms that are able to manage so much data and return accurate results. This problem is accentuated when we talk about minority languages.

In this project we developed three IRS models: a classical model (Boolean), a more current model based on dense vector representations (Word2Vec), and the most modern model (STSB) based on contextual dense vector representations, for a minority language such as Catalan. Using a corpus from the ELIRF-VRAIN research group (DACSA) and tools such as *SpaCy, HuggingFace, SentenceTransformer and Whoosh* that provide us with models pre-trained by other research groups and the STSB model in Catalan that has been created by us, we have developed the three systems presented in the project. We have obtained satisfactory results according to the objectives set. We found that the STSB model improves the search for articles, since it maintains the semantics of the search and takes into account the context. In conclusion, we observe the evolution of IRS in this project by comparing old models with current models.

Key words: Information retrieval systems; dense vector representations; non-contextual embeddings; contextual embeddings; Word2Vec; Sentence to BERT

Índex

Índex de figures

Índex de taules

CAPÍTOL 1
Introducció

En el món actual on la sobreinformació és un gran problema tenim els sistemes de recuperació d'informació (SRI), que tenen com a objectiu recuperar amb precisió aquells documents que nosaltres hem cercat amb l'ajuda d'una consulta, és a dir, ens ajuden a recuperar la informació que realment estem buscant. Un dels exemples més utilitzats en el món actual és Google, en concret el seu buscador, on per mitjà d'una consulta et torna un llistat prioritzat de pàgines web i documents que responen a la consulta,bé perquè contenen els termes de la cerca o perquè contenen termes de característiques similars.

L'ús de tècniques avançades procedents del Processament del Llenguatge Natural (PLN) en SRI ha aportat importants millores. Però, com també passa en altres aplicacions del PLN, moltes de les tècniques s'han desenvolupat per a llengües majoritàries. Moltes de les ferramentes desenvolupades estan dutes a termes per a ser usades en anglès, castellà, alemany, etc. Per a llengües minoritàries, com és el cas del català, ens trobem que el ventall d'eines de PLN és reduït.

Malgrat aquest gran problema que té el català en l'àmbit del PLN, la comunitat catalanoparlant està cada volta més activa i desenvolupant nous projectes que ajudaran a millorar la posició del català en l'àmbit del PLN, per tant, en la seua aplicació a Sistemes de Recuperació d'Informació.

1.1 Motivació

La meua elecció del tema principal d'aquest TFG (Recuperació d'informació basada en vector densos) és causada pels següents motius:

Per una banda, el tercer any de carrera vaig cursar l'assignatura de SAR on el projecte final era crear un sistema de recuperació clàssic sense l'ús de cap llibreria especialitzada, comportant una primera presa de contacte amb els SRI. Malgrat alguns problemes, la implementació i tot el procés van crear en mi una necessitat d'aprenentatge de les últimes tècniques i tecnologies en aquest camp. En aquest projecte se'ns planteja un repte algorítmic enorme, on no només hem d'estudiar, crear i implementar models de representació de les paraules en models vectorials densos, que són les últimes tecnologies utilitzades en el mercat, sinó que hem de resoldre problemes d'escalabilitat, és a dir, tractar amb una abismal quantitat de dates, documents, etc., la qual cosa comportarà problemes d'escalabilitat de les necessitats tant de memòria com de temps.

Per altra banda, sóc una persona nascuda i criada a la ciutat de València, d'ençà que vaig descobrir el món de la Informàtica he sigut molt conscient del poc ús del català en els aspectes més tècnics, com per exemple l'àmbit del PLN. Tot i que cada volta més

la comunitat catalanoparlant està més present i hi ha de nous projectes, no ens podem comparar a altres idiomes més presents en el món de la Informàtica.

En últim lloc, crec que l'àmbit del PLN està en ple auge i és un bon lloc per on començar a desenvolupar les meues habilitats i aprenentatges, comportant així un enteniment del funcionament de les ferramentes d'ús quotidià com es Google. El meu objectiu final com a informàtic es acabar treballant en alguna empresa o equip d'investigació relacionat amb algun tòpic que envolta el *"machine learning o IA"*, i el PLN és un bon punt de partida.

En conclusió, crec que tant personalment com acadèmicament aquest projecte satisfarà les meues ambicions, desenvoluparà les meues habilitats i aportarà una nova ferramenta al català.

1.2 Objectius

L'objectiu principal d'aquest TFG:

- **Creació d'un recuperador d'informació basat en representacions vectorials denses.**

Els següents punts mostren els subobjectius marcats per a l'obtenció de l'objectiu principal

- **Estudi dels recuperadors d'informació: Descripció, història, classificació, i tipus.**
- **Cerca i estudi de les diferents ferramentes i llibreries a utilitzar.**
- **Implementació d'un recuperador d'informació clàssic.**
- **Implementació d'un recuperador d'informació vectorial (Word2Vec), sistema basat en les representacions vectorials no contextual de les frases, utilitzant l'eina *Word2Vec*.**
- **Implementació d'un recuperador d'informació semàntic (STSB), sistema basat en representacions vectorials contextuals de les frases, utilitzant l'eina *Sentence To BERT*.**
- **Avaluació i comparació del funcionament dels tres SRI implementats.**

1.3 Impacte Esperat

Un dels primers avantatges que trobem és l'ús dels SRI creats per tal de recuperar documents en català. Com he exposat amb anterioritat hi ha un gran desproveïment de ferramentes en català, per tant, aquest projecte suposarà una nova eina per tal de ser usada per la comunitat catalanoparlant.

Per tal d'arribar a complir l'objectiu del treball és necessari obtindre unes representacions vectorials denses en català per dos dels models SRI, és a dir, sistemes Word2Vec i sistemes STSB. Aquestes representacions poden ser útils per a altres treballs PLN, per tant, els deixarem a lliure disposició en un repositori com *HuggingFace*.

1.4 Estructura de la memòria

La memòria estarà estructurada per capítols següents.

1. **Introducció:** On és introduirà el projecte a desenvolupar i el problema a resoldre.
2. **Estat de l'art:** On s'exposarà tota la teoria necessària per entendre el problema, el context i trobar la solució a implementar.
3. **Anàlisi del problema:** Detallem el problema a resoldre, quines són les especificacions del projecte i les possibles solucions trobades.
4. **Disseny de la Solució:** Especificació dels components del projecte i l'arquitectura creada. A continuació, s'expliquen les eines que s'han utilitzat.
5. **Desenvolupament i implementació:** Explicació del procés d'implementació i quins errors hem tingut.
6. **Proves:** Proves realitzades per comprovar el correcte funcionament.
7. **Avaluació dels resultats:** Avaluació dels resultats obtinguts pels SRI i una comparació dels sistemes.
8. **Conclusió:** Conclusions finals del projecte.

CAPÍTOL 2
Estat del art

En aquest capítol anem a fer un desglossament de tota la teoria que fonamenta aquest TFG i un estudi de les ferramentes necessàries per a aconseguir els objectius.

2.1 Definició dels Sistemes de Recuperació d'Informació

La pregunta inicial que ens hem de plantejar és "Què és un sistema de recuperació d'informació?"

Per poder contestar aquesta pregunta primer hem de resoldre qué significa "recuperar informació". Aquest terme admet diverses definicions i la comunitat científica està dividida, com va dir Rijsbergen "es tracta d'un terme que sol ser definit en un sentit molt ample" [8].

Un primer grup d'investigadors adverteix que la influència de les tecnologies han fet oblidar que la recuperació d'informació no només usa tecnologia, sinó que es pot fer de forma manual, és a dir, per part de l'esforç humà (Biblioteques, índex de llibres...), i proposen que la definició de recuperació d'informació siga un sinònim de la recuperació de dades. Per posar un exemple, tenim el "Diccionari Mac Millan de Tecnologia de la Informació"suggereix la següent definició "tècniques empleades per emmagatzemar i cercar grans quantitats de dades i exposar-les a disposició dels usuaris" [5].

Un segon grup exposa grans diferències entre els dos termes anteriors. Conclouen que la recuperació de dades té com a inici una pregunta formalitzada o cap, on el resultat en un *matching*[1] entre patrons de bits, que resulta ser les dades buscades. En canvi, la recuperació d'informació comença amb una pregunta difícil de formalitzar-se, on el resultat és un conjunt de documents que tenen certa informació útil per a l'usuari, és a dir, és una ordenació dels documents que satisfà el criteri desitjat per l'usuari. La forma d'avaluació depèn del grau de satisfacció mostrat per l'usuari. Com a exemple tenim a Pérez-Carballo i Strzalkowski: "una típica tasca de la recuperació d'informació és mostrar documents rellevants des d'un gran arxiu en resposta a una pregunta formulada i ordenar-los d'acord amb la seua rellevància" [1]. També exposen que les parts més importants d'un sistema de recuperació d'informació són l'estructurat del sistema d'emmagatzematge i els algoritmes de cerca.

Existeixen altres grups d'autors que no segueixen aquests corrents. El que busquen és, o eludir la definició de recuperació d'informació i se centren en l'emmagatzemament i recuperació com Korfhage, que defineix "un usuari d'un sistema d'informació l'utilitza

[1] aparellament, concordància...

de dues formes possibles: per emmagatzemar informació en anticipació d'una futura necessitat, i per trobar informació en resposta d'una necessitat" [6], o fan una definició genèrica, com Croft que defineix "el conjunt de tasques per les quals l'usuari localitza i accedeix als recursos d'informació que són pertinents para la resolució del problema plantejat." [4].

Una volta definit el terme de recuperació de la informació, en aquest projecte utilitzem la segona definició, podem concloure que un sistema de recuperació de la informació és un sistema que amb un conjunt de Documents degudament processats i emmagatzemats (DOC), una pregunta (REQ) i un algoritme de cerca de similituds pot tornar un conjunt de documents que satisfaça el requeriment fet per l'usuari, és a dir, que mitjançant un algoritme determine quins documents contenen la informació desitjada per l'usuari.

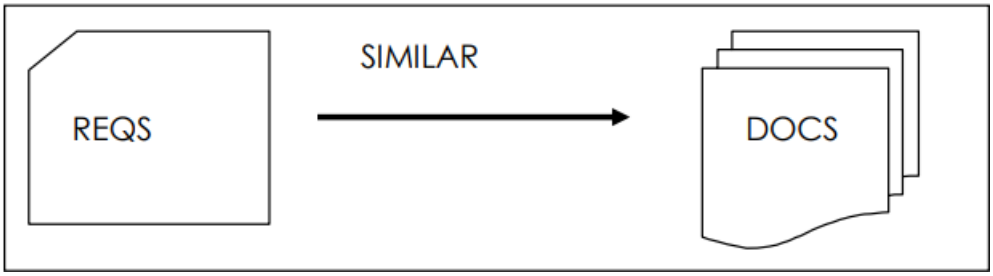


Figura 2.1: Esquema simple d'un SRI. Fuente Salton, G. and Mc Gill, M.J. Introduction to Modern Information Retrieval. New York: Mc Graw-Hill Computer Series, 1983.

Un sistema de recuperació d'informació realitzarà les següents funcions:

1. Analitzarà el contingut dels documents i els representarà en el format adequat

2. Analitzarà la petició de l'usuari per representar-la d'una forma compatible amb els documents guardats, i extraurà els elements determinants per poder fer la cerca.

3. Farà la necessària comparativa entre els requeriments i els continguts

4. Recuperarà la informació rellevant, pot ordenar-la o no, i la mostrarà a l'usuari.

2.2 Història dels Sistemes de Recuperació d'Informació

A continuació analitzarem l'evolució dels sistemes de recuperació d'informació al llarg de la història humana. Hi ha diferents autors que sintetitzen aquesta evolució, però ens centrarem en Baeza-Yates, qui ho fa en 3 fases diferents [9]:

1. Desenvolupament inicial, tornem al passat Egipte on ja existeixen mètodes de recuperació de papirs. Un altre exemple que trobem són l'índex dels llibres que s'han anat fent més complexos amb el pas dels temps o a mesura que la informació s'anava ampliant.

2. Les biblioteques, on primerament usaven un sistema molt rústic que utilitzava l'esforç humà per funcionar, cada llibre estava en una secció i tenia un número que l'identificava. Posteriorment, han anat actualitzant aquest catàleg en l'aparició de la tecnologia, on ja s'usen sistema de recuperació d'informació amb ordinadors, on la cerca és automàtica, no requereix esforç humà. On diferents empreses han clavat mà per poder desenvolupar-los.

3. La World Wide Web, és on els sistemes de recuperació més han brillat, on gràcies a ells el desenvolupament de la web ha sigut exponencial. Tant els usuaris com els SRI han rebut grans recompenses, com per exemple un abaratiment dels costos de producció, una enorme col·lecció de documents, ràpides cerques d'informació... On més han brillat ha sigut en el camps dels directoris i els recuperadors web.

4. El futur, aquesta evolució no ha acabat, podem dir que acaba de començar. Cada volta més els sistemes de recuperació són més necessaris donat el gran volum de documents actuals. Per tant, el SRI s'estan adaptant a totes les necessitats actuals i futures, s'adapten al medi, conseqüentment nous mètodes i algoritmes apareixen per fer-los més eficients, extens i ràpids. És un món en constant evolució.

2.3 Sistemes de Recuperació d'Informació

Una volta donada una definició i repassada la història dels SRI entrarem més en matèria. Aquesta secció es dividirà en dues parts, una primera explicant el processament necessari per poder usar els documents en un sistema de recuperació, i una segona on exposarem les característiques principals dels diferents models SRI.

2.3.1. Procés de classificació

2.3.1.1. Processament de Documents

Per poder fer un recuperador eficient i ràpid una part fonamental del sistema és el processament de la informació, és a dir, hem de processar tant la pregunta com el document.

En primer lloc, l'objectiu és aconseguir extraure del text original aquells fragments que millor el representen, és a dir, els fragments amb mes rellevància, per tal d'aconseguir-ho utilitzem el denominat processament de documents (Conjunt de tècniques especialitzades aplicades sobre el text original). En la **figure 2.2** hi ha una representació d'un possible procés complet.

En segon lloc, la pregunta pot ser vista com una expressió o requeriment que els documents han de complir, o podem representar-la com si fora també un document i buscar la similitud amb la nostra col·lecció.

L'objectiu final del procés és trobar les paraules que es poden usar per a formar l'índex (el que es coneix com a terme), per poder aconseguir-ho fem ús de la denominada tècnica de **preprocessament de documents**:

1. El primer pas consisteix en l'anàlisi textual del text per determinar el tractament dels guions, accents, noms propis, majúscules, etc. Ja que aquests són difícils de tractar i no aporten cap informació rellevant sobre el document a processar.

2. A continuació, realitzaríem l'eliminació de paraules buides, són totes aquelles paraules que no aporten cap o poca informació discriminatòria sobre el text i que el seu percentatge d'aparició en textos és gran, com per exemple *de*.

3. Seguidament, farem ús de la Lematització, és a dir, l'eliminació de les variacions sintàctiques de les paraules restants. Deixarem el Lexema, l'arrel de la paraula.

4. L'últim pas consisteix en l'elecció de les paraules que formaran part de l'índex, o de la frase depenent del que vulgues triar com a unitat. Naturalment, voldrem elegir aquestes paraules amb un pes més important, com podrien ser els substantius o verbs.

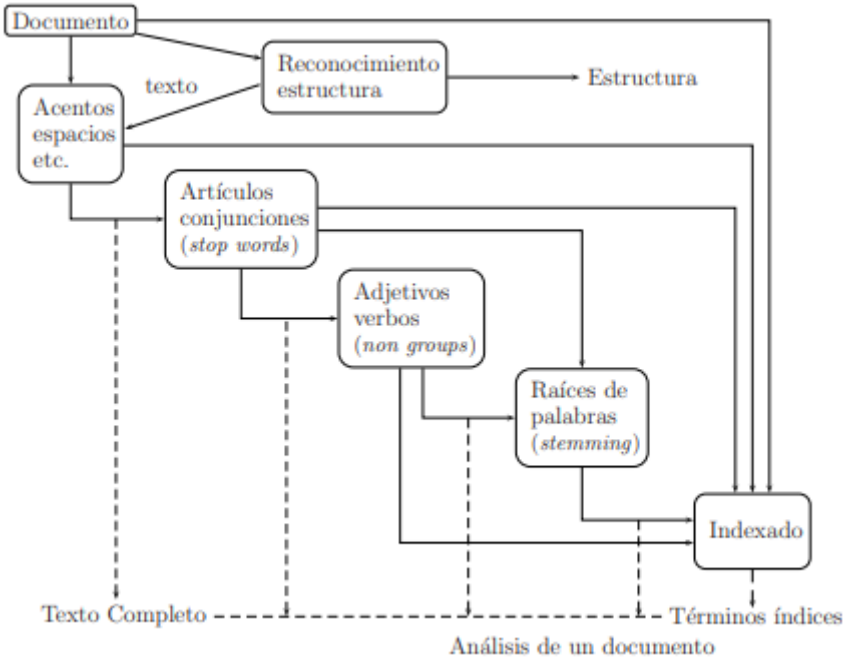


Figura 2.2: Processament de Documents.

Aquests passos mencionats són opcionals, és a dir, no són totalment necessaris per fer la indexació. Tanmateix, si la teua elecció és processar el text original, a l'hora de recuperar-lo en la consulta el resultat serà el text original sense cap mena de processat.

2.3.1.2. Vectorització del text

Procés usat en el Model del Espai Vectorial (VSM) o probabilístic, consisteix a representar el text, usant teoria de conjunts, models algebraics, models estadístics de freqüència, etc. L'objectiu és crear models que siguen capaços de transformar el text en vectors per a la seua posterior classificació i cerca de proximitat. Explicat amb més deteniment en l'apartat del VSM (ref. 2.3.3.1).

2.3.1.3. Extracció i selecció de característiques del text

Consisteix en la selecció d'elements característics i representatius del texts, i posteriorment el càlcul del seu pes. En altres paraules, l'extracció de les característiques del text per la seua posterior classificació, entenent-ho com l'ordenació/classificació dels textos depenent del contingut més característic. Els models més usats són:

1. Elecció manual de característiques més representatives de les característiques originals del text.
2. Les paraules que tenen major informació de classificació matemàticament parlant.
3. Ús de representacions vectorials denses (Word2Vec, BERT).

2.3.2. Classificació Models SRI

Els SRI es classifiquen segons els criteris següents:

1. Obtenció de les representacions dels documents i la consulta.

2. La mètrica de càlcul de similituds d'avaluació dels documents per trobar els més rellevants.

3. Mètodes per establir l'ordenació dels documents seguint la rellevància de la consulta.

En conseqüència l'autor Dominich classifica els SRI en 5 grans grups [2], encara que existeixen diferents propostes de classificació:

Models	**Descripció**
Models clàssics	Inclouen els tres més citats: booleà, espai vectorial y probabilístic
Models alternatius	Estan basats en la Lògica Fuzzy
Models lògics	Basats en la Lògica Formal. La recuperació de informació es un proces inferèncial.
Models basats en la interactivitat	Inclouen possibilitats de expansió del alcans de la búsqueda i fan us de retroalimentació por la rellevància de los documents recuperats.
Models basats en IA	Bases de coneixements, xarxes neuronals, algoritmes genètics y processaments del llenguatge natural.

Taula 2.1: Classificació dels Models de Recuperació d'Informació segons Dominich. Fuente: Dominich, S. 'A unified mathematical definition of classical information retrieval'. Journal of the American Society for Information Science, 51 (7), 2000. p. 614-624.

Citant altre autor, Baeza-Yates la seua classificació es basa en la interactivitat que fa l'usuari en primera instància [7]:

1. Recuperació: mitjançant un pregunta o equació de cerca inicial (retrieval) es recupera la informació

2. Navegació: Navegar per la col·lecció de documents fins a trobar la informació

Baeza-Yates subdivideix els models basats en recuperació en dos grans grups: Clàssics i Estructurats. El primer grup està constituït pels models booleans, vectorials i probabilístics i cadascun d'ells en una sèrie de paradigmes alternatius: Teoria de conjunts, Algebraics, Probabilístics. El segon gran grup corresponen a un llistat de termes sense coincidències i a proximitats de nodes.

Els models de Navegació es poden dividir en 3 grups: Estructura plana, estructura guiada i hipertext. El primer és un lectura sense context del document, el segon mitjançant una estructura ordenada, en classes i subclasses facilita l'exploració, l'últim basat a adquirir informació per mitjà de nodes i enllaços sense ser seqüencial.

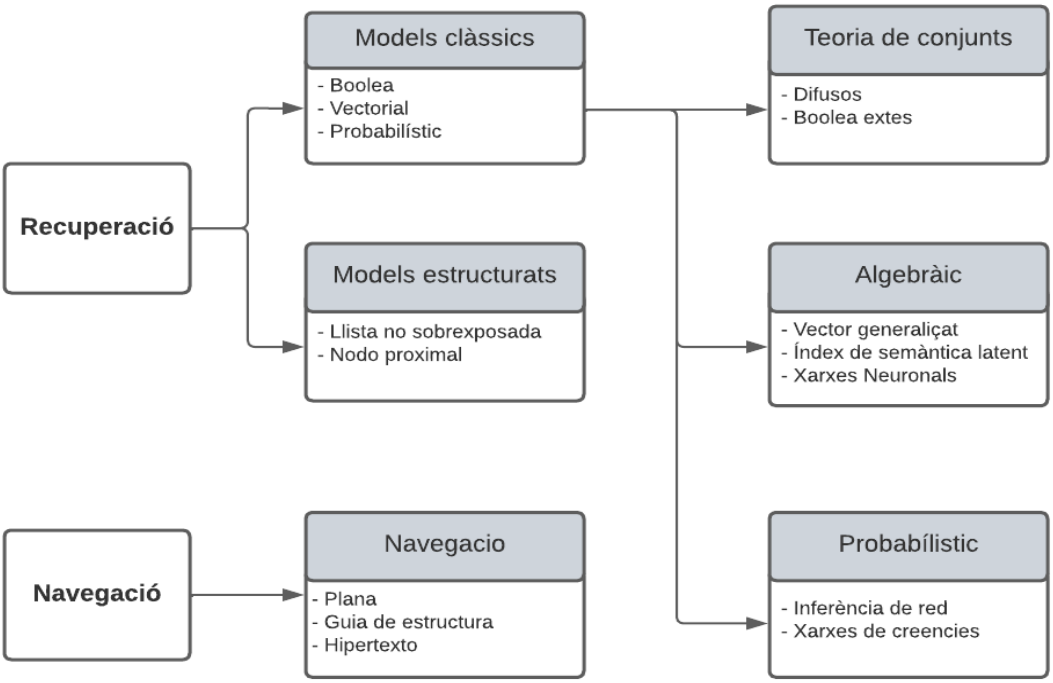


Figura 2.3: Classificació dels Models SRI

2.3.3. Models SRI

En aquest apartat explicarem els models clàssics que utilitzarem.

2.3.3.1. Models Booleà

Aquest model va ser uns dels primers a utilitzar-se, desgraciadament està en desús pels seus desavantatges, basat en L'Àlgebra de Bool i teoria de conjunts per recuperar la informació buscada.

Consisteix en una sèrie de documents (D) on cada document és preprocessat per poder treure les paraules més característiques, encara que és opcional. Per cada paraula anotarem en un vector de dimensions D si està (1) o no (0) en el document. És a dir, finalment obtindre un diccionari de totes les paraules que apareixen en la col·lecció de documents que tenen com a valor un vector que assenyala en quins documents estan. En la següent taula (ref: 2.2) podem observar com hi ha 5 documents i 5 termes, i l'aparició de cada terme en cada document.

	T1	T2	T3	T4	T5
D1	0	1	0	1	0
D2	0	0	1	0	0
D3	1	0	0	0	1
D4	0	0	1	0	1
D5	1	0	0	1	0

Taula 2.2: Taula exemple Model Booleà

Aquest model fa ús dels operants de la lògica de conjunt, com pot ser, la intersecció, la unió i negació. No sols individualment sinó que conjuntament a l'hora de fer-ne les preguntes i trobar el resultat.

La raó per la qual es denomina el model més simple és perquè a l'hora de retornar els resultats només té en compte si el terme buscat està o no en el document. És a dir, la cerca és binària, sols sap si el terme apareix o no en el document. Per tant, no té en compte les vegades que apareix, sinònims, etc.

Avantatges principals del model:

1. Implementació relativament fàcil i moltes ferramentes al teu abast per fer-ho
2. Consultes simples i fàcils d'entendre.

Desavantatges principals del model:

1. Discriminació nul·la entre documents
2. No té en compte les vegades que apareix un terme en el document
3. Recuperacions basades en decisions binàries no usa el matching parcial
4. La recuperació torna o massa documents o molt pocs
5. No retorna en rànquing de documents

En resum, és un model que va ser usat en les primeres dècades dels recuperadors d'informació, i encara s'usa en mails, biblioteques... Per cada vegada més està en desús pels seus desavantatges.

2.3.3.2. Model del Espai Vectorial (VSM)

En aquest models l'objectiu és realitzar una representació vectorial de termes per al conjunt de documents. Creat per G. Salton, C.S Yang i A. Wong en 1975, és un dels models de representació i sistema de recuperació més usat en el món.

VSM consisteix en la representació vectorial dels documents depenent de la freqüència d'aparició dels termes, aquests han de ser no buits, és a dir, paraules o termes que tenen un significat i aportar informació substancial al document.

Entrant mes en matèria, en VSM tenim les següents parts:

- Tenim un conjunt de Documents D_i
- T_j que són el termes del documents que serveixen per identificar el documents amb un pes associat a la importància entre '0' i '1'.

- Vector representatiu de cada document D:

$$D_i = (w_{i,1}, w_{i,2}, ..., w_{i,m}) \tag{2.1}$$

- $w_{i,j}$ que són la representació del j-èssim terme del i-èssim document.

A l'hora d'elegir el pes de cada termini és important fer-ho d'una forma coherent. La comunitat científica segueix buscant la millor tècnica per usar en aquest cas. Hi ha vàries tècniques usades per calcular el pes, les més conegudes són:

1. TF-IDF: la freqüència d'aparició dels termes en el documents
2. Representació booleana: Depenent de si el terme està inclòs o no en el document si li assigna el valor '1' o '0' respectivament.

El procés del VSM consisteix en un anàlisis individual de les paraules fonamentat principalment en un esquema de pesos i una mesura de similituds. En conseqüència, tenim una estricta comparació dels termes, és a dir, la comparació només admet el termes d'igual escriptura, i, per tant, opera estadísticament. En altres paraules, el VSM consisteix en un conjunt enorme de paraules que són analitzades individualment, on a cada document se li assigna un valor vectorial depenent del conjunt de paraules que li constitueixen.

El principal desavantatge d'aquest model és que no contempla el context de les paraules ni el seu significat. El concepte de la sinonímia o paraules en diferents significats no són analitzades seguint el correcte mètode.

TF-IDF

Principal tècnica de càlcul de pesos dins del Model del Espai Vectorial (VSM), usat per calcular la matriu de termes i documents, mitjançant el càlcul del pes de cadascun dels termes de la col·lecció de documents. Considerem les següents definicions:

- n = nombre de termes distints en D.
- tf_{ij} = voltes que apareix el terme t_j en el document D_i
- df_j = quantitat de documents que contenen el terme t_j
- idf_j = $\log D/df_j$

La representació vectorial de cada document consisteix en n entrades de valors en representació de cada terme que compon la col·lecció. El valor de les entrades es calcula mitjançant el pes del termes en la col·lecció de documents, depenent de la freqüència de la paraula en el total de la col·lecció i en un document individual. En conseqüència, podem deduir que el pes d'un terme augmenta si apareix molt en un document i disminueix si es repeteix en molts documents. Malgrat això, si en un document el terme no apareix el valor serà de 0.

El càlcul del pes es fa seguint aquesta formula:

$$w_{ij} = tf_{ij} \;\; x \;\; idf_j \tag{2.2}$$

On tf_{ij} informa sobre la importància en el document, tots els termes importen igual, en canvi, idf_j informa sobre la importància del terme en la col·lecció total de termes.

En conseqüència assignem un pes al terme tenint en compte la freqüència d'aparició del mateix en tots els documents i en un individual, per tant, assignem valors baixos a termes poc significatius.

Una vegada treta la matriu de termes i documents, podem efectuar les nostres consultes P. On primer, efectuarem el canvi a vector de la consulta, i per últim farem una comparació de similituds entre el vector pregunta i la nostra matriu, per extraure els vector amb major grau de similitud.

$$Similitud\_cosinus(P, D_i) \tag{2.3}$$

Acabem de representar la tècnica *tf-idf* més genèrica, però existeixen altres variants. No obstant això, totes estan fonamentades en el fet que el valor del pes d'un terme reflecteix la importància d'aquest dins d'un document i en tota la col·lecció. [10]

Similitud Cosinus

La Similitud Cosinus és un càlcul de similituds entre 2 vectors dins d'un espai vectorial donant com a resultat l'angle que es forma entre els dos vectors. Aquest resultat varia entre [-1, 1], 1 si tenen la mateixa direcció, 0 si l'angle format pels 2 vectors és de 90°, ja que s'anul·len els cosinus, i -1 si tenen direccions oposades.

Aquesta operació trigonomètrica s'utilitza per calcular la similitud entre dues representacions vectorials, és a dir, per calcular la similitud entre dues paraules o frases que han sigut transformades a vectors. En particular, estem tractant sempre d'espai positiu, per tant, el resultat varia entre [1,0]-[0,1].

La fórmula és:

$$similitud\_cosinus = S_c(A, B) := cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{||\mathbf{A}|| \cdot ||\mathbf{B}||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{2.4}$$

On A_i i B_i són elements dels vectors A i B.

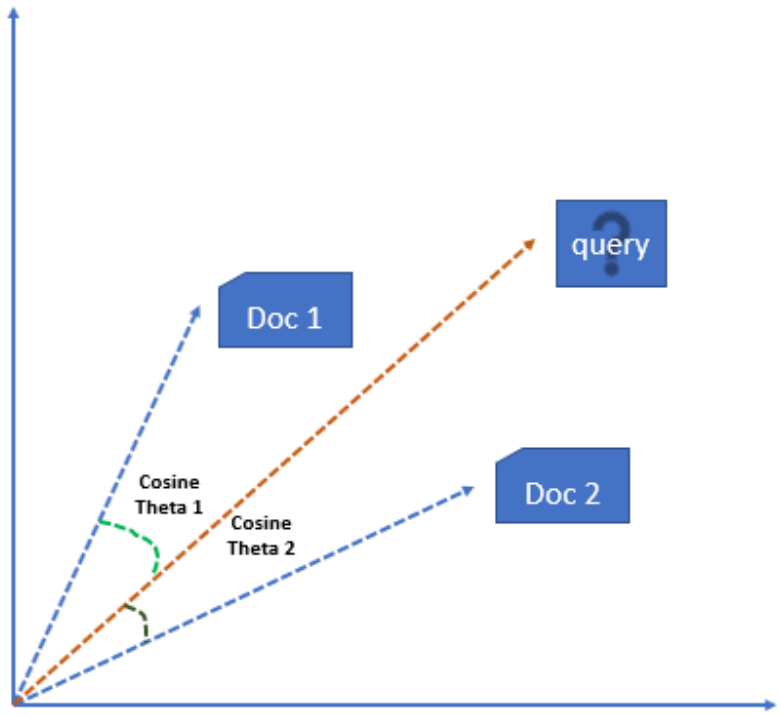


Figura 2.4: Similitud Cosinus

URL: https://2.bp.blogspot.com/-saTZSoc5RAA/WfghS_CMvJI/AAAAAAAAGBg/PcZvT0QNZCcPJq8fAv2v_cSwrnagdm9RgCK4BGAYYCw/s1600/cosine_similarity.PNG

2.3.3.3. Word Embedding

Abans d'entrar en matèria del Model Word2Vec (W2V) hem d'entendre que són els word embeddings. Perquè a conseqüència d'usar el model Word2Vec, haurem representat cada paraula o terme de la consulta i dels documents amb un vector dens associat. Usarem el terme 'embedding' per a nomenar aquests vectors.

L'embedding intenta captar les característiques de la paraula (semàntica, definició, context, etc.). Les representacions vectorials denses s'utilitzen en moltes aplicacions del PLN

Hi han diferents maneres de crear-los, per exemple usant xarxes neuronals, matrius de co-ocurrència, models probabilístics, etc.

Hi ha eines per a crear embeddings a partir de grans quantitats de text, entre elles tenim el Word2Vec. Word2Vec o altres assumeixen que paraules en un context similar tendeixen a tindre un significat similar i són usats per entrenar el models Word Embeddings. Els resultats obtinguts demostren una clara eficiència a l'hora de modelar els embeddings depenent del seu context, encara que trobem limitacions importants (no trobem significat precís, antònims i sinònims tenen un context similar, etc.)

2.3.3.4. Model Word2Vec (W2V)

A diferència del model vectorial que tracta els termes dels documents com si foren independents entre si, el Model Word2Vec (W2V) ha demostrat ser capaç d'extraure el context de les paraules i utilitzar-lo per a la seua classificació.

El Model Word2Vec (W2V) va ser creat en 2013 per Google, més concretament Tomas Mikolov i el seu equip [11], consisteix en una xarxa neuronal de dues capes amb la capacitat de processar documents. El funcionament consisteix d'un *input*[2] de documents, un corpus, on per a totes les paraules o frases més significatives de cada document del corpus extrau un vector dens, és a dir, crea una representació vectorial de les paraules o frases més importants del text. Aquest vector té l'especialitat de tindre en compte el context de la paraula. Gràcies a aquestes característiques és usat en diverses aplicacions.

Entrant mes en matèria, el model consta de 2 arquitectures fonamentals:

- *Skip-Gram*: la seua funció és predir el context a partir de la paraula.
- *Continuous bag-of-words* (CBOW): la seua funció és la de predicció de la paraula depenent del context on es trobe.

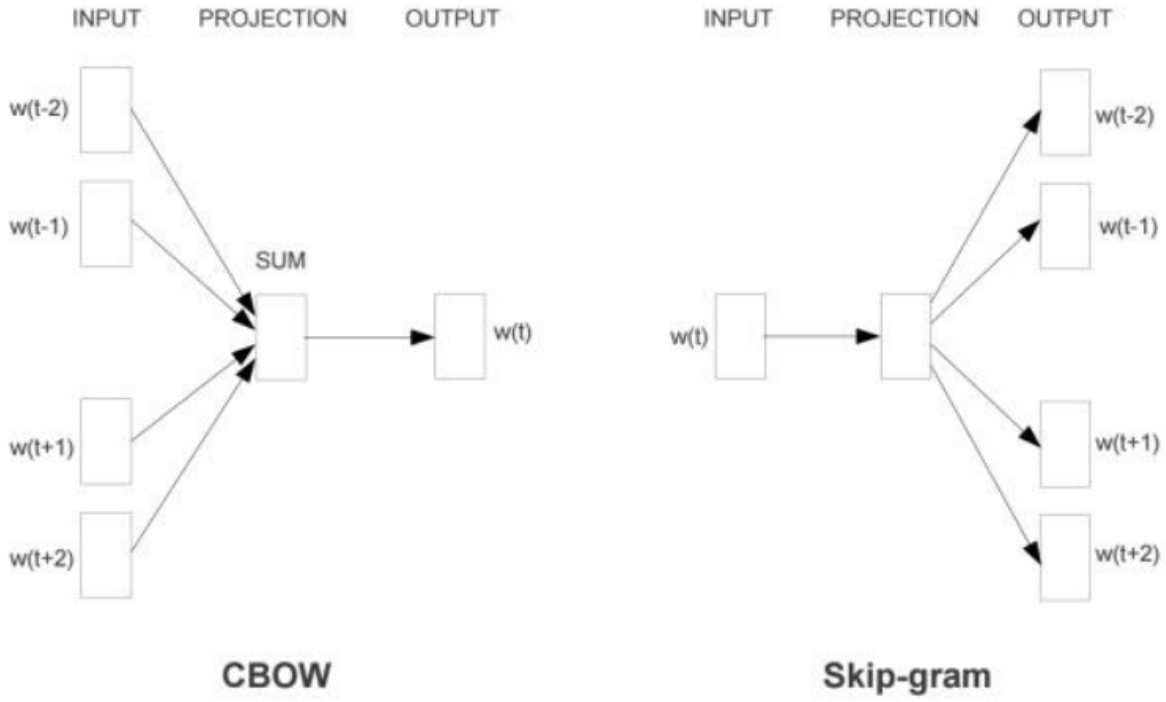


Figura 2.5: Les 2 arquitectures del model Word2Vec

Skip-Gram

És l'arquitectura més utilitzada hui en dia, s'usa per predir el context mitjançant una paraula d'entrada. És a dir, tenim la paraula objectiu en l'entrada i com a sortida les paraules que l'envolten. Per exemple, si tenim la frase *'I have a cute dog'*, l'entrada seria 'a', mentre que la sortida és *'I'*, *'have'*, etc. Tot suposant que tenim un ample de finestra de 5.

Concretant, aquesta arquitectura consisteix en una sèrie de capes d'una xarxa neuronal, 1 oculta de dimensions igual al *embedding* usat que és més xicoteta que el vector d'entrada i sortida. En el moment de fer el càlcul, les dades d'entrada i sortida han de tindre la mateixa dimensió i han d'estar *'one-hot-encode'*, sols hi ha un bit igual a 1 tots

[2]Components d'entrada en un mètode.

els altres en 0. En la capa final, la de sortida, s'aplica la funció d'activació de *softmax*, utilitzada per saber la probabilitat que té cada paraula en aparèixer en el context.

$$P(y = j|\theta^i) = \frac{e^{\theta^i}}{\sum_{j=0}^{k} e^{\theta_k^i}}$$
$$\text{on} \quad \theta = w_0x_0 + w_1x_1 + ... + w_kx_k = \sum_{i=0}^{k} w_ix_i = w^Tx \tag{2.5}$$

El *embedding* per a la paraula objectiu s'extrau de les capes ocultes de la xarxa després d'introduir-la en la representació *'one-hot'*.

En el següent gràfic es mostra les capes i el procés.

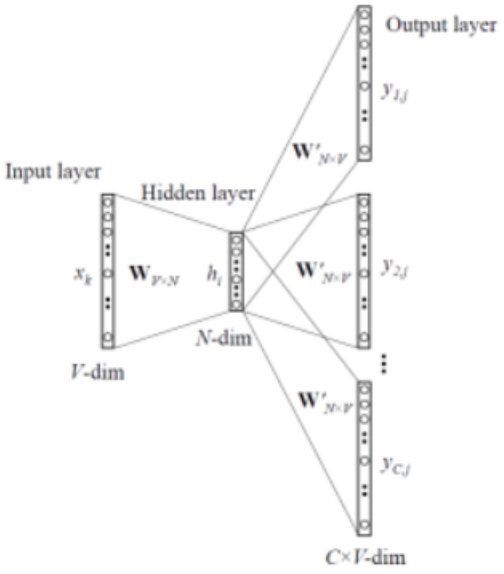


Figura 2.6: Capes i procés de l'arquitectura Skip-gram

Una de les característiques principals d'aquesta arquitectura és que les representacions en compte de mesurar la longitud del vocabulari (V) mesuren la de la capa oculta (N). En conseqüència, els vectors resultants són més significatius en les relacion interparaula. És a dir, en la resta de dos vectors ja calculats podem obtenir un tercer que tinga un significat semàntic similar, ja siga gènere, temps verbal, sinònim, etc. Tal com il·lustra la següent imatge.

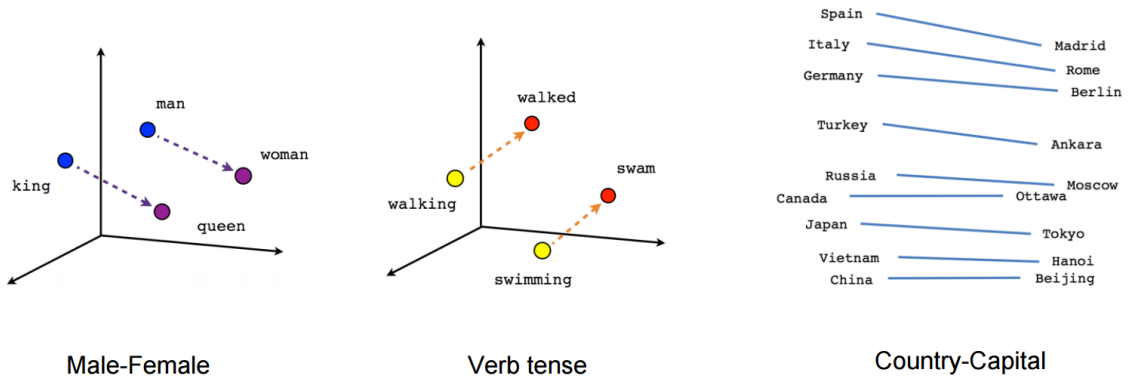


Figura 2.7: Relació linear de les paraules

URL: https://www.tensorflow.org/images/linear-relationships.png

Continuous bag-of-word CBOW

Continuous bag-of-word és l'arquitectura que donat un context com a entrada retorna la paraula objectiu. En altre paraules, donat un context inicial quina és la paraula que té més probabilitats d'aparèixer. En la següent imatge es pot observar les diferències principals en l'arquitectura Skip-gram.

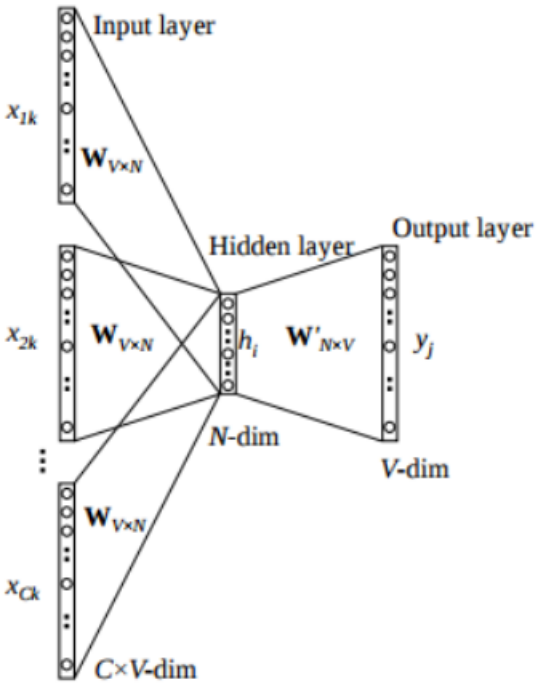


Figura 2.8: Capes i procés de l'arquitectura CBOW

Analitzant les dues arquitectures podem deduir que la diferencia més remarcable és la generació dels vectors de paraules. CBOW utilitza els contextos + la paraula objectiu per a alimentar a l'algoritme, i a més agafa la mitjana obtinguda en la capa oculta com a resultat. Per exemple, tenim 2 frases *'He is a nice guy'* i *'She is a wise queen'*, i la paraula objectiu és *'a'*, per tant, hem d'introduir en el model les dues frases sense la paraula *'a'*, és

a dir, *'He is nice guy'* i *'She is wise queen'*. Per últim, amb les dades de la capa oculta obtenir la mitjana per saber on apareixeria la paraula objectiu. On en la sortida es realitzarà una activació final per obtenir els resultats.

2.3.3.5. Model Sentence-BERT (STSB)

Aquest model es

> una modificació de la xarxa BERT utilitzant xarxes siameses i triplets que és capaç de derivar embeddings d'oracions semànticament significatives' [12]

Per tal d'entendre millor el model, primer hem de fer una explicació del model BERT.

BERT

Representacions de codificadors bidireccionals de Transformats o BERT és un model de representació del llenguatge creat l'octubre del 2019 per l'empresa Google. Produint un canvi en el panorama actual de l'àmbit dels SRI, ja que millora notablement els resultats de les consultes fetes pels usuaris.

BERT a diferència dels models anteriors utilitza un model bidireccional permetent conèixer el context que envolta a una paraula. És a dir, BERT entrena el seu model mirant tant a l'esquerra com a la dreta d'una paraula aprenent així el context. Aleshores, abans de codificar entén el context que la rodeja, permetent una representació semàntica de les paraules que té en compte el context en el qual apareixen. De fet, els embeddings proporcionats per BERT es coneixen com a embeddings contextuals.

A continuació un exemple:

1. Çarles està cursant una carrera"

2. Çarles està corrent una carrera"

En aquest exemple trobem 2 significat, el primer significa uns estudis Universitaris i el segon una competició, que per a un ésser humà aquesta clara la diferencia. Tanmateix, per a la màquina que utilitza els models anteriors a BERT entendria com carrera el fet d'una competició, ja que és l'ús més habitual.

D'altra banda, l'ús d'un model BERT ens proporcionaria un resultat diferent, ja que, com hem explicat abans, entén el context de la paraula i és capaç de fer una diferenciació semàntica.

A continuació explicarem detalladament el funcionament del BERT. L'ús de models BERT comporta dues fases d'entrenament:

1. **Pre-Training:** On el model compren el llenguatge

2. **Fine-Tuning:** On el model entrena per a feines mes especifiques.

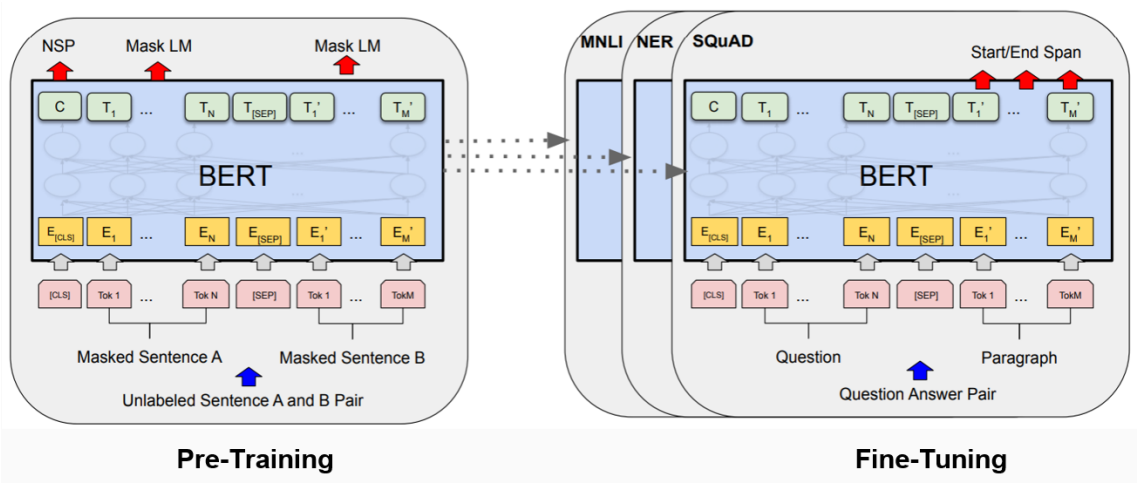


Figura 2.9: Proces de Pre-Training i Fine-Tunning

La fase de **Pre-Training** està dividida en dues etapes (MLM i NSP) per evitar que el model tracte dues voltes a una paraula en el context, ja que es bidireccional.

L'etapa **Masked Language Model (MLM)** consisteix a emmascarar un 15% els tokens[3] d'entrada. En aquest cas, l'emmascarament consisteix en un 80% de les paraules triades són canviades per **[MASK]** i un 10% per una paraula aleatòria per evitar problemes i l'altre 10% en la paraula original.

Input: The man went to the $[MASK]_1$. He bought a $[MASK]_2$ of milk .
Labels: $[MASK]_1$ = store; $[MASK]_2$ = gallon

Figura 2.10: Exemple de MLM

D'altra banda, l'objectiu de l'etapa **Next Sentence Prediction (NSP)** es aconseguir que el model siga capaç de percebre i entendre les relacions entre les frases, és a dir, que el model sàpiga quina frase continua a quina o quina va primer, etc. Per a obtenir el coneixement, l'etapa consisteix a tindre 2 frases (A i B) com a input[4] i predir si és successora o no, en un 50% dels casos B és la frase successora de A.

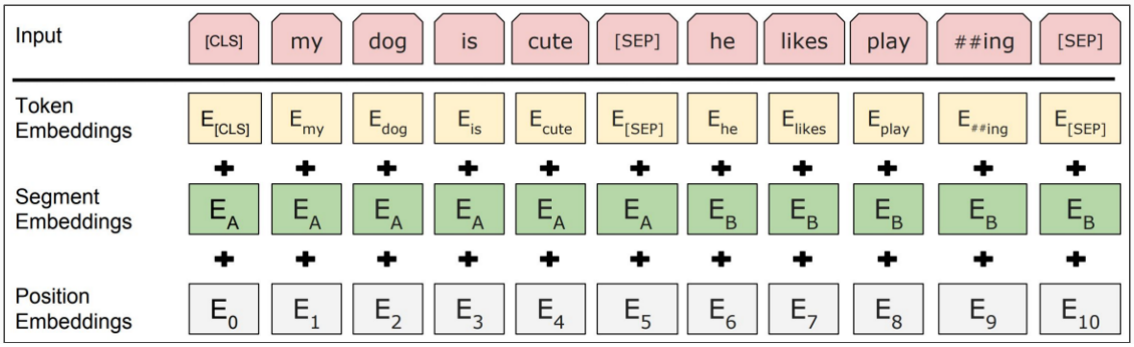


Figura 2.11: Exemple de MLM

La segona etapa **Fine-Tuning** és una etapa prou més senzilla, sols consisteix a entrenar al model amb la tasca especifica que tu necessites en el moment, en el nostre cas en

[3] Subsegments de paraules transformades en cadenes de caràcters
[4] Elements d'entrada a la màquina

un model especialitzat en **scoring similarity sentence**[5]

Els principals avantatges de model BERT són:

- Un model molt més precís que els anteriors
- Molt potent
- Una apropament més natural a la parla humana

Desavantatge:

- Un model prou més costos en termes de memòria i temps
- Similitud entre frases és una tasca costosa per aquest model, ja que ho fa d'una en una.

Sentence-BERT

Sentence-Bert és un model que ha aconseguit solucionar els errors del model BERT a l'hora de comparar 2 frases. Com diuen els autors en la següent afirmació:

> "Trobar la parella més semblant en una col·lecció de 10.000 frases requereix uns 50 milions de càlculs d'inferència (65 hores) amb BERT. La construcció de BERT el fa inadequat per a la similitud semàntica"[11]

En canvi, el model SBERT aconsegueix els mateixos resultats necessitant **5 segons** de funcionament.

SBERT aconsegueix aquesta reducció basant-se en el principi de baixar una capa d'abstracció, com a conseqüència obtenim dos nivells de codificació, com a paraula i com a frase, és a dir, codifiquem el significat de la paraula i de la frase.

Malgrat la poca eficiència que té BERT en la similitud de frases, no volem perdre el seu poder semàntic per això gastem com a nucli principal un model pre-entrenat BERT.

Entrant en més detall de la funcionalitat de SBERT, trobem que el model SBERT després d'utilitzar el BERT per fer una representació vectorial de la frase, la processa per mitjà d'una operació de *pooling*[6] obtenint un embeddings de grandària fixa. En aquest model utilitzem la versió de *pooling*: *MEAN*, que calcula la mitjana dels vectors resultats de BERT.

En ordre d'aconseguir uns embeddings que mantinguen el significat semàntic i siguen comparables utilitza xarxes siameses i triples per anar actualitzant els pesos del vectors. Les xarxes utilitzades estan constituïdes per una estructura *Regression Objective Function* que utilitza l'operació *Cossine similarity* en 2 frases.

[5] Comparació de dues frases per etiquetar quina similitud tenen i dotar d'un valor al resultat

[6] utilitzada per reduir la dimensió d'un vector, però mantenint les seues característiques per a ser classificat

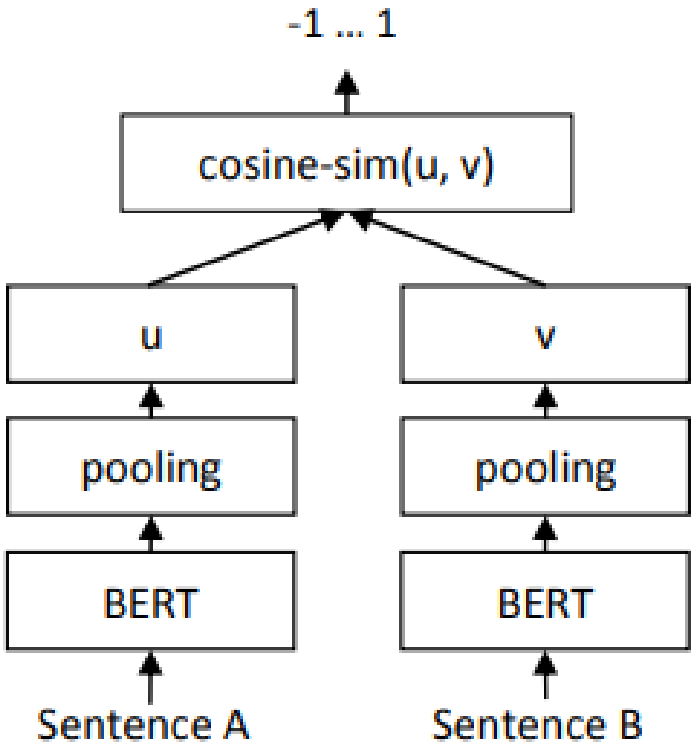


Figura 2.12: SBERT architecture at inference, for example, to compute similarity scores. This architecture is also used with the regression objective function. Fuente: [11]

Finalment, obtenint un vector que puga ser utilitzable per a *Semantic Similarity*.

Els avantatges del model són els següents:

- Precisió en l'operació de comparació
- Remarca el context i el seu significat semàntic
- Precís i potent

Desavantatges:

- L'enorme ús de recursos, memòria, CPU, etc. Encara que menys que el BERT
- Consum alt de temps en la indexació. Menys que el BERT
- Costos d'implementar i entrenar.

2.3.4. Exemples de SRI

En el mercat de hui en dia podem trobar diferents aplicacions que utilitzen els SRI. A continuació farem una llista dels més usats hui en dia.

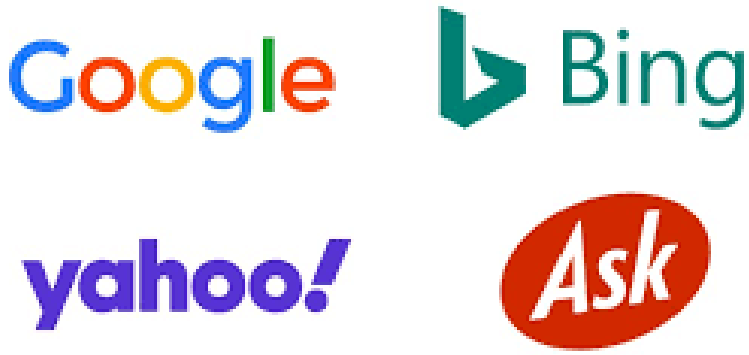


Figura 2.13: Exemples de SRI

1. Google: El motor de cerca de Google és el SRI més usat hui en dia, es pot dir que tot el món en internet l'utilitza. Va ser creat per Larry Page i Serger Brin, i la paraula se sol usar com a sinònim per fer les cerques. El que més destaca és la seua facilitat d'ús.

2. Yahoo, Bing: com Google són motor de cerca, la seua funció en buscar en internet pàgines, documents, etc. que concorden en la consulta.

3. Tesaures: Llistat terminològic controlat especialitzat sobre un àmbit específic del coneixement que guarden relació semàntiques i genèriques entre si. El termes estan ordenats jeràrquicament, per tant, aconseguint eficiència i precisió en la cerca.

4. Directoris: Llistes organitzades que guarden informació estructuradament i jeràrquica. Classificant els termes per categories i fent que l'usuari enllace dels més generals als més específics.

5. Índex: Llistat de termes que representen el contingut d'un document, curs, etc. estan normalitzats.

2.4 Ferramentes de desenvolupament dels SRI

En aquest apartat anem a estudiar les diferents ferramentes més utilitzades pel mercat actual a l'hora de tractar amb el Sistemes de recuperació d'informació. Com per exemple, els llenguatges de programació, llibreries de codi obert, etc.

2.4.1. Llenguatges de Programació

Figura 2.14: Logo Java i Python

URL: https://raygun.com/blog/images/java-vs-python/java-vs-python.png

2.4.1.1. Python

Python és considerat un dels millors llenguatges de programació per la seua versatilitat en tots el àmbits, com si es tractarà d'una navalla Suïssa. El seu llenguatge natural i coherent facilita sobre manera el seu aprenentatge, i al mateix temps gràcies al gran abast de paquets de codi i el seu llenguatge transparent el converteix en uns dels llenguatges de programació en millor predisposició per ser usat en l'àmbit del *Natural Language Processing*(PLN), més concretament per desenvolupar SRI. Malgrat això, la seua característica principal que ajuda són les grans quantitats de llibreries de codi obert que es poden utilitzar per gestionar diferents tasques en els SRI (parlarem d'elles en l'apart 2.4.1.3).

2.4.1.2. Java

Java és un dels llenguatge més utilitzats en l'ús comú del PLN, i, per tant, en el desenvolupament dels SRI. És un dels llenguatges en més llibreries i paquets creats per tota classe de feines, per tant, trobareu llibreries específiques per les tasques necessàries per a desenvolupar un SRI. Pareix que Python l'avance en el camp de la Ciència i en el PLN, encara que Java garanteix els mateixos avanços i característiques que Python gràcies a la seua comunitat tan gran que té i els equips que el suporten.

2.4.1.3. R

El llenguatge de programació R és ben conegut per l'ús en l'aprenentatge estadístic, tanmateix, té un gran ús en el PLN, per tant, en el desenvolupament de SRI. Com tots sabem en les tasques relacionades en els SRI hem de treballar en corpus plens de dades i ací és on entra en joc *R*, la seua especialitat és la d'analitzar i transformar grans grups de dades.

Hi ha d'altres llenguatge de programació, com Node.js, etc. Que serveixen per desenvolupar SRI, però en aquest treball el nostre llenguatge de programació utilitzat es Pyt-

hon, donades les seues característiques i la gran quantitat d'ajudes en llibreries o paquets que ens millora el nostre treball en SRI.

2.4.2. Llibreries i eines

Figura 2.15: Logo NLTK, spaCy i mes

URL: https://miro.medium.com/max/1200/1*qsRJFHCxC0edtLuZeoUi4g.png

2.4.2.1. Python

En Python podem trobar una gran quantitat d'eines i llibreries diferents que ens facilitaran diferents tasques necessàries en el nostre treball.

NLTK

El llenguatge natural Toolkit (NLTK) és una de les ferramentes més completes que podem trobar en el mercat actual. Entre les seues funcionalitats més destacades trobem la classificació, la tokenització, el derivat, l'etiquetatge, la segmentació i el raonament semàntic. Entre les seues característiques més importants està la gran flexibilitat, podem implementar diverses funcionalitats fent ús de diversos algoritmes i mètodes, a més fent gala de la seua compatibilitat en diferents idiomes, és a dir, que pots usar-la per a diferents idiomes en els teus projectes. Presenta diversos desavantatges, com per exemple l'expressió en cadena de les dades analitzades, útil per alguns contexts simples, però en els projectes avançats no és de gran ajuda. Un altre desavantatge és que en diferencia d'altres llibreries és lenta, la seua execució no és massa eficient. En conclusió, és unes de les ferramentes més utilitzades, a pesar dels desavantatges, i completes del mercat.

SpaCy

SpaCy és l'altra ferramenta fonamental per a Python si vols treballar en el camp PLN, concretament SRI. Es caracteritza per la seua interfície simple, unes funcionalitat simples

d'utilitzar i completes, un model de xarxa neuronal pel processament del llenguatge i el anàlisi complet de diversos components. En contraposició a NLTK, estem parlant d'una ferramenta ràpida i eficient, on les dades són representades com a objectes fent que el seu ús siga més simple. No obstant, no és compatible en tants idiomes, encara que en les últimes actualitzacions han posat nous. En resum, és una ferramentes ràpida i útil en el camp mencionat, ja que no requereix un algoritme específic per a cada funcionalitat i és molt compatible en l'àmbit científic.

2.4.2.2. Java

OpenPLN

OpenPLN és una ferramenta desenvolupada i administrada per la Fundació Apache, a conseqüència tenim una gran compatibilitat en altres projectes d'Apache (Flink, NiFi, Spark). Es caracteritza per ser una ferramenta molt completa, que conte una funcionalitat per tots els possibles components dels PLN, que poden ser utilitzats tant com comandos com en la seua forma de paquet i compatible en molts idiomes diferents. En resum, és una ferramenta ràpida, eficient i completa, tot el desitjat per treballar en l'àmbit dels SRI, donades totes les diferents funcionalitats que presenta.

Stanford CorePLN

Stanford CorePLN presenta un conjunt de ferramentes utilitzades en l'àmbit PLN, estan més orientades en l'aprenentatge i característiques basades en regles. Desenvolupat per l'àmbit científic per a ser utilitzada en l'àmbit científic, pot ser poc útil per a la producció. Considerada una ferramenta eficient, completa i flexible, ja que pot ser utilitzada amb diferents algoritmes i mètodes per a les seues funcionalitats. També té versió en Python, encara que la de Java és més completa. La seua llicència es doble, però en te de simples per l'ús comercial.

2.5 Crítica a l'estat de l'art

Una crítica que podem fer al mercat actual de sistemes de recuperació d'informació és la poca oferta d'eines i models en idiomes minoritaris, com és el català, és a dir, que a l'hora de desenvolupar el teu propi projecte no trobem tantes ajudes externes com ho faríem si estem desenvolupant-ho en castellà o Anglès. Aquest error el podem trobar encara en les universitats actuals on quasi tot el desenvolupament és produït en castellà i no en català/valencià, molt dels projectes on m'he basat per buscar idees o ajudes estan en castellà o directament utilitzen les eines, models en angles per facilitar-se el desenvolupament. Aquest paràgraf va orientat als models i eines de codi obert, és a dir, aquelles d'ús gratuït, ja que com tothom saps Google utilitza models en català, però no són de lliure disposició.

Un altre dels aspectes a millorar és la comparacions de models. No tothom és un especialista del PLN, per això és important fer un apropament cap a tota mena de persones. Per això, en aquest projecte intente fer una comparativa entre diferents models per trobar la millor solució en cada cas, és a dir, mitjançant la comparació podríem trobar quin model s'acobla millor donats diferents contextos en una posterior explicació per facilitar la seua comprensió.

En conclusió, el principal problema a resoldre és el poc repertori de sistemes de recuperació d'informació actuals, de lliure disposició, en idiomes minoritaris a les universitats i al món.

2.6 Proposta

La meua proposta no descobrix cap tecnologia nova, ni pot competir en Google, però actualment, no existeix cap model de similitud semàntica creat per ser usat en català de codi obert, és a dir, d'ús gratuït, per la qual cosa no hi ha cap possibilitat de desenvolupar projectes PLN en rapidesa i certesa sobre els resultats. Addicionalment, els sistemes de recuperació actuals desenvolupats en català de codi obert estan desactualitzats, utilitzant models anteriors (Booleà, vectorial, etc.) i no es poden comparar a escala de prestacions amb els utilitzats per empreses privades.

Com a conseqüència, presente un treball on es desenvolupa diferents models, alguns millors que altres, de SRI en català per resoldre els problemes presentats en els apartats de dalt. El meu treball va estar emmarcat en l'àmbit dels recuperadors d'informació en català. Aquest treball va estar caracteritzat per l'ús de les millors i més avançades tecnologies per al desenvolupament dels SRI, com pot ser l'ús de Python, i de llibreries com SpaCy o NLTK, que són les que estan guanyant la carrera tecnològica en l'àmbit del PLN. Estarà fonamentat en 3 models de SRI diferents per poder fer una comparativa millor, com són el Booleà, Word2Vec, STSB, per obtindre un coneixement sobre quin model s'adapta millor a diferents contexts.

Tanmateix, aquest treball té com a objectiu donar al món un nou model dels SRI de lliure ús. Un model de representacions vectorials contextuals (STSB) en català, que com hem explicat en l'apartat anterior (ref: 2.3.3.5) és un model innovador i que presenta millores substancials tant en el processament dels textos i l'anàlisi de les paraules o frases com en la recuperació d'aquestes. Encara que els temps d'indexació i processament puguen ser pitjor, la relació temps-resultat millora notablement a altres models SRI.

CAPÍTOL 3
Anàlisi del problema

3.1 Presentació del Problema

Com he explicat amb anterioritat les llengües minoritàries en l'àmbit del PLN estan molt poc desenvolupades. L'actual ventall d'eines i llibreries en català o d'altres és reduït, més concretament a la universitat politècnica tenim projectes centrats en la traducció automàtica de textos, resum automàtic o anàlisi d'emocions. Però cap sistema *SentenceBERT* de lliure ús creat per a català.

Presentant una nova oportunitat d'innovació, crear un SRI en català, que puga competir i innovar en el mercat actual, per poder utilitzar-ho posteriorment a la universitat o a nivell global. No obstant això, després d'una cerca detallada de la competència i teoria hem conclòs que hem de presentar un treball que puga complir els següents requeriments:

- Suficient capacitat per indexar un corpus[1] amb un gran nombre de documents.
- Eficiència a l'hora d'indexar i recuperar els documents.
- Competitivitat davant altres SRI (Temps, eficiència, etc.).
- Garanties de resultats correctes i comprovació d'aquests.
- Comparatives dels possibles models SRI creats.
- Codi net i organitzat per poder ser usat per altres persones.
- Documentació adjunta per poder entendre el codi.

Davant d'aquests requeriments trobem un gran problema, l'eficiència del projecte.

3.1.1. Anàlisi d'eficiència algorítmica

L'eficiència en l'àmbit de la informàtica sempre és un dels principals problemes, vivim en un món on la immediatesa en essencial a l'hora de buscar resultats i poder competir en termes de temps enfront de les empreses davanteres.

Un dels principals objectius d'aquest projecte és poder treballar amb un gran nombre de documents a indexar, estem parlant d'uns 100.000 o més, és a dir, la quantitat de dades a manejar és enorme comportant així un consum de temps a l'hora d'indexar

[1] La col·lecció de documents

gran. No només estem parlant d'un consum de temps alt, sinó que el tractament de tants documents comportarà un consum abundant en termes de memòria utilitzada.

Tenint en compte aquest greu problema, hem conclòs que un dels nostres punts forts haurà de ser l'ús de llibreries modernes i eficients que tinguen els algoritmes més avançats i eficients que el codi obert[2] pot oferir-nos. Per altra banda, els nostres esforços també aniran enfocats en reduir la quantitat de dades tractades i guardades en memòria per poder enfrontar-nos a un ús total d'aquesta i no poder continuar.

3.2 Identificació i anàlisi de possibles solucions

Coneixent els requeriments i principals problemes del projecte podem trobar que no existeix una única solució, sinó que podem solucionar aquest problema de diverses formes.

Els següents punts analitzaran diverses solucions possibles i exposaran els avantatges i desavantatges que presenten.

3.2.1. Solució amb un únic SRI implementat

En aquesta solució ens centraríem sols a implementar un SRI, tant siga Booleà, com Word2Vec o d'altres. Aquesta solució té una sèrie de pros i contres:

Avantatges:

- Ràpid d'implementar
- Rapidesa en treure resultats
- Avaluació ràpida dels resultats

Desavantatges:

- Cap comparació entre els models
- Poc abast del projecte
- Objectiu del projecte no complint, no es fa una diferenciació dels models
- Poc ambiciós.

En conclusió, seria un solució un poc pobre, ja que no faríem cap comparació dels models SRI i un dels objectius del TFG és poder comparar els models.

3.2.2. Solució amb model probabilístic o estructurals implementats

Solució basada en un implementació de models completament diferents dels explicats en l'estat de l'art (ref 2.3.2), són models basats en funcions lògiques o probabilitats d'ocurrència. En aquest cas, podríem observar el següent:

Avantatges:

- Ràpida implementació.
- Comparativa en els altres models.

[2] Llibreries, programari..., que no necessiten llicencia per ser usada o és gratis

- Avaluació d'uns resultats diferents.

Desavantatges:

- No són un models que s'adeqüen bé als objectius.
- Complicació a l'hora de l'avaluació dels resultats.
- Relació Temps-resultats mala en comparació a altres models.

En conclusió, a l'hora de buscar els millors models per poder comparar i implementar en aquest projecte, aquest tipus de model no encaixen bé, trobem que ens dificulta la feina en l'àmbit d'avaluar els resultats i treure unes conclusions correctes.

3.2.3. Implementació de 4 o més models diferents

La solució més completa i amplia de les possibles ja que implementem una gran varietat de model SRI. Malgrat que en ser la més completa trobem que és la més complexa. Per tant, si fem un anàlisi trobem:

Avantatges:

- Comparació total entre models
- Resultats avaluats amb precisió i comparació d'aquestos
- Compliment dels objectius
- Conclusió precisa i correcta

Desavantatges:

- Lenta implementació
- Lenta extracció de resultats
- Molta memòria consumida
- Molt d'abast

En conclusió, seria la solució més completa, ja que obtindríem un SRI per cada model possible, tanmateix, és una solució que consumeix molts recursos i temps, i els resultats podrien no ser tan detallats comparant-los en una solució que use menys models SRI.

3.3 Solució proposada

Estudiant totes les possible soluciones amb els seus avantatges i desavantatges hem conclòs que la solució òptima per poder complir els requisits del projecte i satisfer les nostres ambicions és la següent:

3.3.1. Implementació de 3 models de SRI

Aquesta solució està basada en l'estudi teòric de la cerca de la solució òptima i que millor s'ajusta al projecte, consisteix en la implementació de tres tipus de SRI i la seua comparació posteriors per trobar el model que millor s'ajusta a cada moment.

Models SRI elegits:

- Model Booleà (ref: 2.3.3.1) per la seua fàcil implementació i en ser uns dels primers models creats ens donarà una comparació en el temps i com ha evolucionat l'àmbit dels SRI en el temps.
- Model Word2Vec (ref: 2.3.3.4) Un dels models més utilitzat hui en dia on la quantitat d'eines eficients per al seu desenvolupament és enorme. I la posterior comparació amb els altres models és fàcil i molt concloent.
- Model STSB (ref: 2.3.3.5) hui en dia és un dels models punters en els SRI, ja que pot comparar semànticament frase o textos. Per tant, a l'hora de comparar ens dona una vista actualitzada de la realitat i com és el seu rendiment en contraposició dels 2 models anteriors.

Un dels motius principals de la tria d'aquest models és que en la WordWeb [3] la gran majoria de SRI implementats són models de representació vectorial (contextuals i no contextuals). Una altra característica fonamental per la tria és que són models que estan fonamentats en una teoria semblant, per tant, l'apartat de comparació de models i resultats és una feina més simple i fàcil d'entendre. Un altre aspecte a mencionar, són les facilitats a l'hora de la implementació, ja que, com tenen una teoria semblant una mateixa estructura d'índex i buscador pot servir per als 2 models, és a dir, no és necessari duplicar l'esforç a l'hora de crear el codi per crear els dos models SRI.

Tot aquest procediment l'implementarem tant en català com en castellà per poder fer una comparativa de com funciona en cada idioma, a més per assegurar-se del correcte funcionament dels models en català, ja que seran desenvolupats per nosaltres.

3.3.2. Model Conceptual

El model conceptual consisteix en una breu descripció del sistema, és a dir, de quin funcionament tindrà el nostre projecte quan estiga desenvolupat.

La idea base del projecte, explicada ja en els punts anteriors, és la creació de tres models SRI on cada models indexarà el mateix corpus, en el nostre cas tindrem dos corpus un en castellà i l'altre en català, per poder fer una comparació detallada dels resultats obtinguts de cada model respecte dels altres. Aleshores, en el moment en què la indexació està completa l'usuari farà una *query*[4] als 3 diferents índexs[5] i aquests retornen els documents que més semblants són a la *query*.

Per últim, per tal de poder fer una comparació real entre el models, es farà una sèrie de consultes per tal d'obtindre una resultat per cada model, aleshores, per mitjà d'una avaluació manual podrem comprovar el rendiment de cada model en cada una de les consultes.

Les següents imatges il·lustren el model conceptual del projecte:

[3] Google, Yahoo, etc. una paraula tècnica per referir-nos a Internet
[4] una consulta per trobar algun document
[5] Llista dels documents

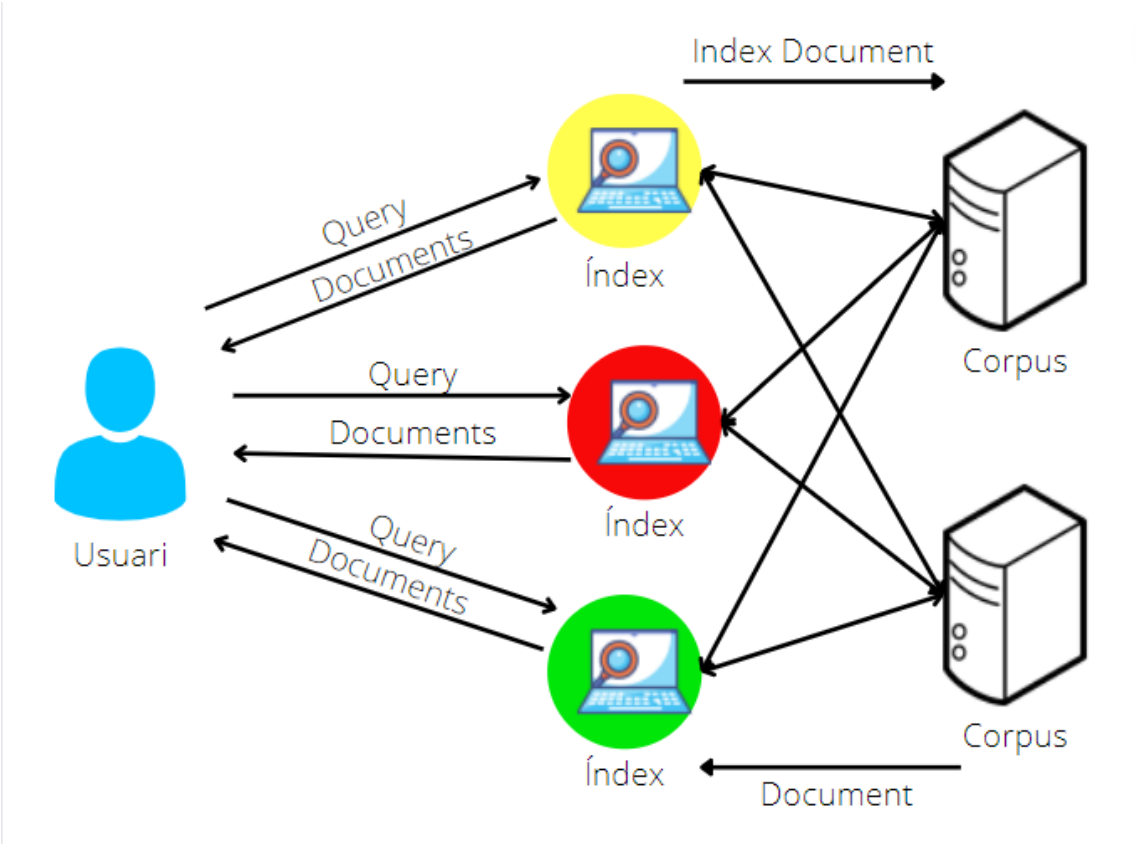


Figura 3.1: Model conceptual de la Solució proposada (part 1)

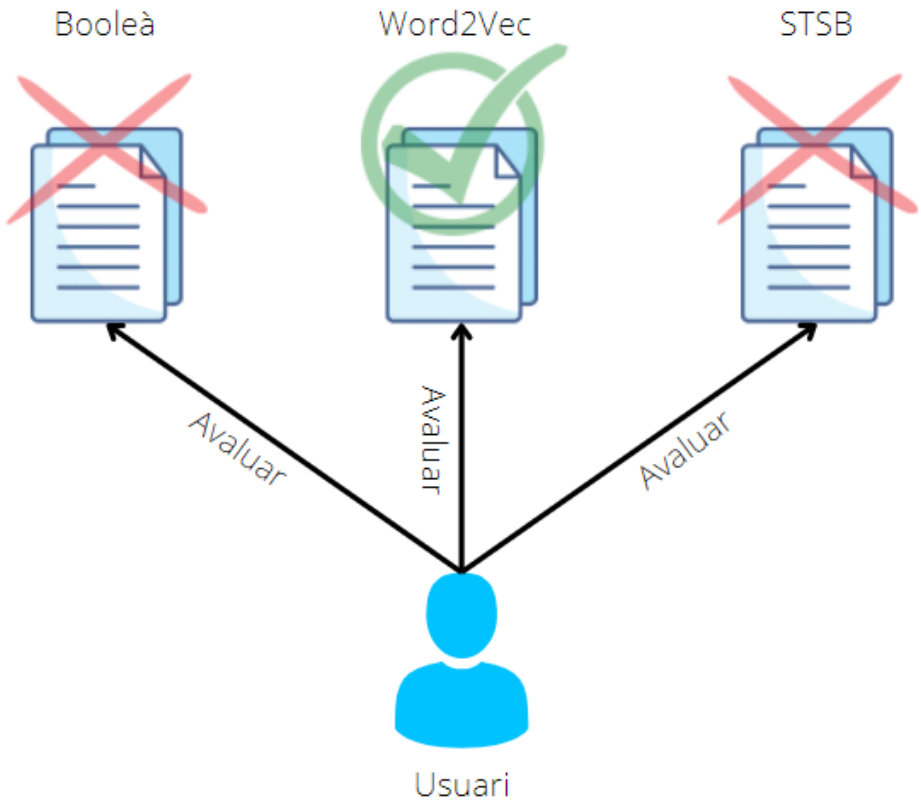


Figura 3.2: Model conceptual de la Solució proposada (part 2)

3.3.3. Pla de Treball

El pla de treball inicial estava dividit en 5 etapes principals a priori, al llarg del desenvolupament del projecte aquest pla va patir modificacions en detecta algun problema o alguna nova oportunitat.

3.3.3.1. Pla de Treball Inicial

En aquest apartat explicarem quin és el Pla de Treball esperat a l'hora de desenvolupar el projecte:

Etapa 1

L'Etapa 1 consisteix en l'estudi teòric dels SRI i l'estudi del mercat global. Primerament, consisteix en la lectura de múltiples documents que tinguen relació en el món del SRI, com per exemple, quins models existeixen, com es poden crear, quines característiques tenen, etc. Després farem una cerca de la possible competència i una estudi del mercat global, per entendre la situació actual. Seguidament de l'estudi teòric i del mercat, mitjançant una reunió amb els tutors, crearem una primera possible solució i els seus requeriments.

Temps Total: 2 setmanes de Treball

Etapa 2

A continuació de definir la solució teòricament, estudiarem les eines i llibreries més útils per a la posterior implementació. Una vegada decidit quines utilitzarem vindria el desenvolupament de la solució, és a dir, la creació del codi per tal d'anar creant el models i els índexs. Utilitzant un entorn de programació còmode i les eines correctes, podrem crear el codi necessari per portar a terme el projecte. En aquesta etapa tindrem especial cura en utilitzar els algoritmes d'indexació més eficients i models amb una tassa d'encert gran per obtindre uns resultats acordes els requisits.

Temps Total: 1 mes de Treball

Etapa 3

El treball d'aquesta etapa consistirà en la indexació del corpus per part dels diferents models SRI, cada model indexarà els dos corpus creant així un índex en català i l'altre en castellà. El treball d'aquesta etapa no necessitarà un gran esforç en el desenvolupament de crear codi, sinó que requerirà depurar i millorar eficientment el codi per poder indexar tot el corpus a utilitzar.

Temps Total: 2-3 setmanes de Treball

Etapa 4

A continuació decidirem quines són les consultes més deterministes per al nostre treball. El gran problema és determinar quines són les millors consultes, aleshores necessitarem un gran nombre de consultes per trobar aquelles que millor mostren les diferències entre els models. Com en l'anterior etapa, potser necessitem depurar el codi per fer-ho més eficient.

Temps Total: 2-3 setmanes de Treball

Etapa 5

L'última de totes les etapes és on el treball està concentrat en l'avaluació manual i subjectiva dels resultats obtinguts en l'anterior etapa. En aquesta etapa, analitzarem les possible variacions en els resultats d'una mateixa consulta per determinar quin model ha tret els millors resultats. Donat que no hi ha un conjunt de resultats de referència etiquetat haurem de fer-ho manualment, és a dir, llegir els trossos dels articles tornats i comparar-ho amb la consulta feta per tal d'analitzar la semblança entre les dues frases. Per últim i amb tots els resultats analitzar farem una conclusió explicativa dels resultats del projecte comparant els 3 models creats i escriure la memòria.

Temps Total: 3-4 setmanes de Treball

En conclusió, un projecte on el treball està ben repartit i cada etapa és important per poder desenvolupar-ho correctament. Donat que poden sorgir erros i més crec que el temps de treball pot quedar-se curt.

3.3.3.2. Pla de Treball Real

Una volta acabat el projecte puc fer un avaluació detalla de les hores reals treballades per portar-lo a terme.

Etapa 1

En primer lloc, el projecte anava a consistir en sols dos dels models creats, per tant, en haver de buscar informació d'un tercer el temps real va pujar.

Temps Total: 3-4 setmanes de Treball

Etapa 2

En aquesta etapa poden trobar una modificació substancial de temps de treball donada per la contínua actualització del codi a causa de certs factors limitants com la memòria i l'eficiència. Entrant mes en detall, l'ús d'un corpus gegant agregat a ser de les primeres voltes a desenvolupar un projecte d'aquestes magnitud va requerir un esforç major per tal d'aconseguir executar-lo, posteriorment varen sorgir problemes a causa de la falta de memòria en l'ordinador on desenvolupava el projecte. En conseqüència, va ser una lluita constant per millor el codi i fer-lo eficient. En conclusió, aquesta etapa va consumir més temps del pressuposat.

Temps Total: 6 setmanes de Treball

Etapes 3-4

Semblant a l'etapa anterior el treball va requerir un major treball respecte al pressuposat donat per limitacions de memòria a l'hora de guardar els índex o en el moment de la indexació. També ocorria en el moment de treure resultats, ja que els sistemes inicials eren poc eficients i tardaven molt a carregar l'índex.

Temps Total: 10 setmanes de treball

Etapa 5

Aquesta Etapa va ser la que més s'ajusta, només quedava analitzar les dades i comparar els resultats.

Temps Total: 3-4 setmanes de Treball

En conclusió, A causa dels problemes sorgits per la falta de memòria, les feines es van anar retardant. A part, en ser un treball on les etapes estan interconnectades entre elles, pot trobar-ne errors de l'Etapa 2 en l'Etapa 4 i retardar tot el desenvolupament.

Mesos totals Pressuposats: 3 mesos i 1 setmana
Mesos totals Real: 6 mesos

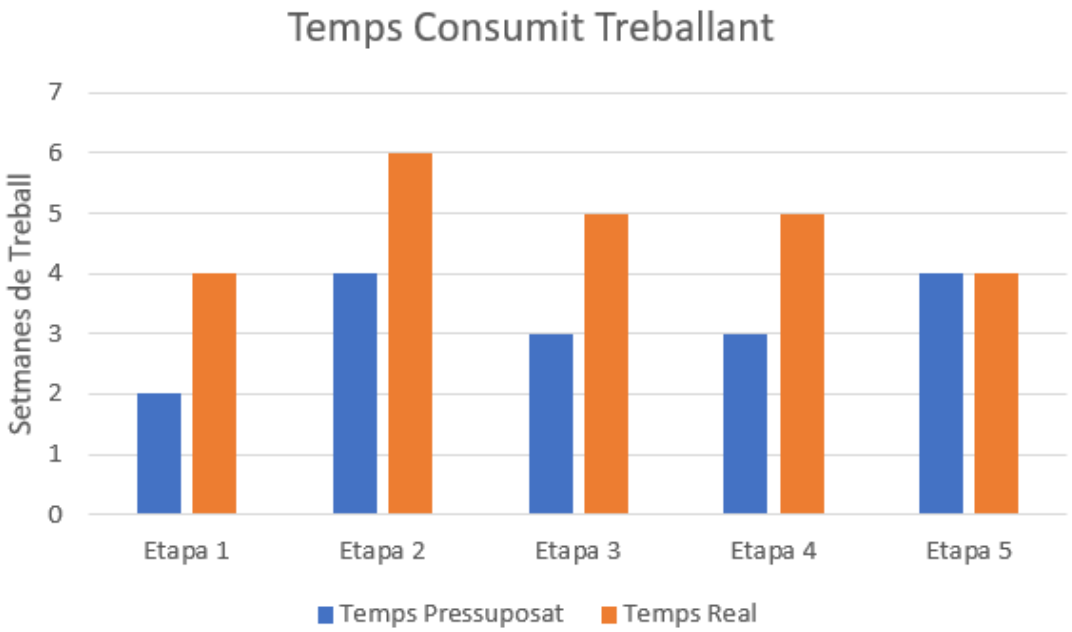


Figura 3.3: Taula del Temps de Treball consumit

CAPÍTOL 4
Disseny de la Solució

En aquest capítol entrarem en més detall en el disseny de la solució, i quins han sigut els mitjans seleccionats per tal de poder portar-la.

4.1 Arquitectura del Sistema

L'estructura general del projecte està basada en una arquitectura simple d'un SRI. En el nostre projecte tenim 5 elements essencials:

- **Col·lecció de documents (Corpus):** Com la mateixa paraula indica són un conjunt de documents escrits en llenguatge natural i són la part fonamental de tot SRI.
- **Indexador:** Component que indexa el corpus, és a dir, transforma el corpus en un índex perquè puga ser llegit per la màquina per a la seua posterior recuperació, crea una representació del corpus per poder comparar-ho en les futures consultes.
- **Índex:** Llista dels documents transformats per a la seua recuperació. En aquest projecte en particular tindrem un índex per cada idioma i model que utilitzarem. En total 6 índexs.
- **Interfície:** Component de SRI encarregat de rebre la consulta de l'usuari i passar-la als algoritmes de cerca.
- **Algoritme de cerca i rànquing:** Component encarregat de transformar la consulta per tal de poder recuperar el documents més rellevants i fer un rànquing a continuació. És a dir, verifica quins dels documents satisfan la consulta millor.

En la Figura 4.1 tenim un diagrama de components i com van relacionant-se entre ells.

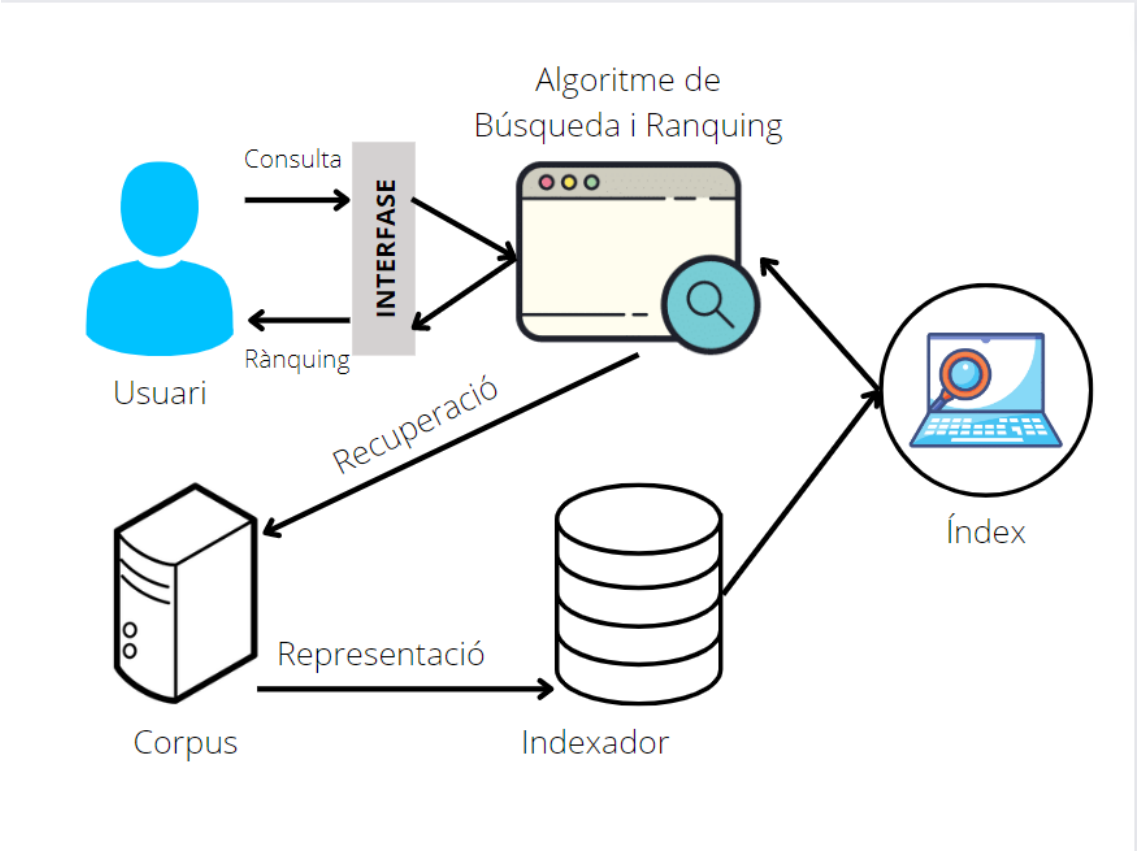


Figura 4.1: Diagrama de components

Com es pot observar en la Figura 4.1 els components estan interactuant entre si per tal de fer una recuperació precisa dels documents seleccionats. Sense entrar molt en detall, en primer lloc, l'indexador crearà una representació vectorial o booleana del corpus creant així l'índex per mitjà d'un algoritme de transformació explicat en el pròxim apartat (ref: 4.2). Una volta creat l'índex, l'usuari es connectarà a la interfície i interroga al sistema per mitjà d'una consulta que serà transmesa a l'algoritme de cerca i transformada en una representació comparable amb l'índex. Després, l'algoritme de cerca amb la nova representació cercara en l'índex quins documents són més semblants per a recuperar-los i, per últim, crearà una rànquing ordenant-los per rellevància. Finalment, aquest documents es recuperaran en el corpus per respondre a la consulta de l'usuari.

En el nostre projecte com a particularitat l'algoritme de cerca i rànquing pertany a la mateixa classe que l'indexador.

4.2 Disseny Detallat

En aquest apartat en centrarem a fer una descripció detallada dels components essencial d'aquest projecte. Partint de l'explicació anterior, anirem detallant els components i connectant-los uns amb els altres.

4.2.1. Col·lecció de documents (Corpus)

En primer lloc, anem a parlar de la col·lecció de documents utilitzada per tal de fer els SRI, és a dir, quins documents hem indexat per poder recuperar-los a posteriori amb una consulta.

DACSA, adquirit per investigadors del grup ELiRF adscrits a VRAIN, és la col·lecció triada per a indexar. "El corpus DACSA es va recopilar mitjançant un rastrejador web que va capturar més de 6 milions d'articles de notícies, prop de 2 milions d'articles publicats en català, i més de 4 milions escrits en castellà."[3] Després d'un procés de selecció i tractament, el corpus de DACSA està compres per 725,184 mostres en català i 2.120.649 mostres en castellà. Per tant, com en aquest projecte volem comprovar el rendiments dels diferents SRI tant per castellà com per a català DACSA és el corpus ideal. No sols ens proporciona un corpus en 2 idiomes, sinó que les notícies són d'actualitat i de tòpics interessant per als nostres resultats.

Cada mostra en DACSA conté informació d'una notícia publicada, on podem trobar diferents parts de l'article, ja pot ser, el *summary, url, id, article, etc.* (ref: 4.2) En el nostre projecte ens centrarem en el cos de l'article, és a dir, la notícia en si. Per tant, deurem centrar-nos en l'apart *article* que compren el cos de la notícia i és on trobarem les frases a indexar. Dividirem el cos de l'article en frases per poder fer la indexació, i tindre uns resultats més detallats i acordes a les consultes possibles, ja que s'ajusten més de mesura i detall.

Malgrat tot, hem hagut de treballar en un subgrup del corpus sencer per falta de memòria, el subgrup compren entre unes 50000-100000 mostres a indexar. a continuació els explicarem en més detall.

- **català:** Compren articles de 2 diaris diferents (**ara.cat** i *diari de girona*), encara que en el corpus total existeixen 9 diaris, utilitzant 90000 articles estem fent ús de la part d'entrenament que gasta DACSA. Els tòpics principals del són política, economia, internacional i societat.

- **castellà:** Compren articles de 2 diaris diferents, encara que hi ha 21 diaris en total al corpus, indexem vora 100000 mostres de la part d'entrenament de DACSA. I els temes més repetits són els mateixos que en català.

El corpus està emmagatzemat en memòria secundaria en un arxiu d'extensió *.jsonl* i la Figura 4.2 il·lustra l'estructura d'un document.


```
{"source": "ara",
   "id": "5e0ac8899309ce244c06851f",
   "url": "https://www.ara.cat/societat/escoles-garantir-docent-aules-durant_0_1902409751.html",
   "summary": "El ministeri de l'Interior fixa el seguiment del sector educatiu en un 31,5%.",
   "article": "Les escoles han de garantir un docent per cada sis aules a infantil i primària,
               segons els serveis mínims aprovats pel departament de Treball, Afers Socials i Famílies de la Generalitat per a la vaga general d'aquest dimecres.
               En educació especial, un docent per cada quatre aules, i per al servei de menjador i activitats complementàries,
               un terç de la plantilla, com a les llars d'infants. El secretari general tècnic del ministeri d'Interior,
               Juan Antonio Puigserver, ha manifestat aquest dimecres que el seguiment de la vaga general aquest matí ha tingut un \"escàs impacte\"
               en els centres de treball, a excepció del sector educatiu, amb un 31,5%. Segons el sindicat majoritari d'USTEC-STEs,
               el seguiment ha sigut del 45%. Pel que fa a les universitats, fonts de la Universitat de Barcelona (UB) asseguren que l'edifici del Raval està
               bloquejat per piquets i les facultats de Física i Química ho estan \"de manera parcial\". A la resta de facultats de la UB \"hi ha normalitat\".
               A la Universitat Pompeu Fabra (UPF) les classes s'estan produint pràcticament amb normalitat: al Campus del Mar s'han desenvolupat les classes de
               manera normal, al Campus del Poblenou la situació s'ha normalitzat cap a les 10.00 h i al Campus de la Ciutadella, a partir de les 12.00 h.
               Segons fonts de la Universitat Politècnica de Catalunya, el seguiment de la vaga ha sigut del 70% per part dels estudiants,
               d'un 4% pel personal docent i investigador (PDI) i un 5% pel personal d'administració i serveis (PAS).
               En la majoria d'escoles on hi havia exàmens aquest dimecres, les proves han sigut ajornades. El sindicat d'educació USTEC-STEs,
               la plataforma Universitats per la República, el Sindicat d'Estudiants dels Països Catalans (SEPC) i el Sindicat d'Estudiants (SE)
               donen suport a la vaga general d'aquest dimecres i a la manifestació convocada per Òmnium i ANC de dissabte, 11 de novembre.
               USTEC-STEs ha assegurat que comparteix els motius laborals de la vaga impulsada per Intersindical-CSC, ha rebutjat l'empresonament de
               consellers del Govern cessat i dels presidents de l'ANC i Òmnium, i ha defensat el sistema d'immersió lingüística de l'educació catalana
               davant l'article 155. La plataforma Universitats el SEPC i el SE fan una crida a \"buidar les aules\" aquest dimecres.
               La Fundació Escola Cristiana, patronal que agrupa els 434 col·legis religiosos catòlics de Catalunya, no fa cap crida a participar a la vaga,
               però ha deixat llibertat als centres perquè decideixin si la secunden o no. El sindicat majoritari a les escoles concertades,
               la USOC, no s'ha adherit a la convocatòria de vaga, cosa que sí que ha fet la USTEC, que és el sindicat majoritari a les escoles públiques.
               Algunes de les congregacions religioses, com les dominiques o les Escoles Pies, s'han mostrat més partidàries que d'altres de secundar la vaga.",
   "article_nwords": 454,
   "summary_nwords": 13,
   "similarity": 0.0,
   "html_detected": false,
   "too_short": false,
   "too_similar": false,
   "not_ended": false,
   "lang": "ca",
   "lang_prob": 1.0,
   "lang_ca_prob": 1.0}
```


Figura 4.2: Exemple d'un document en el corpus

4.2.2. Classe Llançadora:

Aquesta classe realment no representa cap component dels anteriors mencionats, sinó que és utilitzada per controlar el procés d'indexació.

Per poder ser llançat el projecte utilitzem aquesta classe. Amb el pas d'arguments per consola controlem diferents parts:

- El corpus a utilitzar.
- Els noms dels diferents índexs a crear.
- models dels indexador a utilitzar, és a dir, quins models SRI utilitzarem per crear els índex.

Aleshores per portar a terme la creació de l'índex fa una crida a la classe Indexador, utilitzant els mètodes d'indexar, save i vaciar (ref: 4.2.3).

Una de les principals funcions d'aquesta classe és calcular la demora, el temps d'indexació total, per fer un estudi posterior.

4.2.3. Classe Indexador:

Aquesta classe compren 2 dels diferents components que comprenen aquest projecte, l'Indexador i l'algoritme de cerca i rànquing.

4.2.3.1. L'Indexador

Per explicar aquest component ho desglossarem en 2 parts, quins són els sistemes de representació que utilitzarem amb els seus models i quin serà el mètode triat per indexar els documents.

Sistemes de Representació

A l'hora de parlar del sistemes de representació hem de parlar de quins models SRI volem utilitzar en el nostre projecte, aquest dilema ja ha sigut resolt en el capítol anterior (ref: 3.3), no obstant, entrarem en més detall sobre quins models hem utilitzat.

Booleà:

El model booleà fa una representació booleana de totes les paraules que componen el corpus, és a dir, a cada paraula li assigna en quins document apareix (ref: 2.3.3.1). Per implementar aquesta representació hem fet ús de la llibreria whoosh [17], ja que ens permet fer una ràpida indexació paraula a paraula.

Word2Vec:

Aquest model crea una representació vectorial de les frases que conté el document, és a dir, cada paraula de la frase és transformada en un vector de dimensió 300 i la mitjana d'aquestes és el vector resultant de la frase. Tot aquest procés és portat a terme per la llibreria especialitzada en PLN SpaCy [15], que ens proporciona ja models Word2Vec preentrenats tant per a català com per a castellà.

- **castellà:** *es_core_news_md* molt similar al model en català encara que serveix per al castellà. (ref: 4.3.3.1)

- **català:** *ca_core_news_md* model preentrenat per SpaCy, utilitzat per extraure embeddings Word2Vec en català (ref: 4.3.3.1).

STSB:

El model STSB crea una representació vectorial de les frases, en la particularitat que la dimensió del vector és de 768, ja que no sols conté informació del context com el Word2Vec sinó que també conté la informació semàntica de la frase. Per poder desenvolupar aquest models varen utilitzar *HuggingFace* (https://huggingface.co/) un conjunt de llibreries especialitzades en el PLN que guarda milers de models entrenats. Per poder fer ús d'aquest models es necessita la llibreria *Transformer*.

- **castellà:** *eduardofv/stsb-m-mt-es-distilbert-base-uncased* és el model triat per poder fer les transformació en castellà. *stsb_multi_mt*.

  > "Aquest model es va crear prenent distilbert-base-uncasedi entrenant-lo en una tasca de semblança textual semàntica mitjançant una versió modificada de l'script d'entrenament per a STS de *Sentece Transformers*. Es va entrenar utilitzant els conjunts de dades espanyols de *stsb_multi_mt*, que són els conjunts de dades *STSBenchmark* traduïts automàticament a altres idiomes mitjançant deepl.com."[14]

  Aquest model presenta una precisió d'un 75% a l'hora de calcular la similitud.

- **català:** Un dels principals reptes que presentava aquest projecte era el fet que no existia cap model STSB en català de lliure disposició. La solució va ser crear el model *driwnet/stsb-m-mt-ca-distilbert-base-uncased*.

  Aquest model està desenvolupat per nosaltres, basat en el model en castellà *eduardofv/stsb-m-mt-es-distilbert-base-uncased*, i pujat amb posterioritat a la pàgina de *HuggingFace*. Convertint-se en el primer model STSB en català de lliure ús en el món, aconseguint aportar nova tecnologia a l'àmbit del PNL de codi obert.

  A continuació, anem a fer un desglossament del procediment utilitzat per a la seua creació:

  – **Corpus**: El corpus utilitzat és la versió castellana de *stsb_multi_mt*.

    > "STS Benchmark comprises a selection of the English datasets used in the STS tasks organized in the context of SemEval between 2012 and 2017. The selection of datasets include text from image captions, news headlines and user forums."[13]

    Traduïda al castellà per mitjà de deepl.com i posteriorment utilitzant el traductor Salt i una revisió manual de les paraules marcades com no traduïdes.

```
Un home està tocant una gran flauta.,Un home està tocant una flauta.,3.8
Un home està untant formatge ratllat en una pizza.,Un home està untant formatge ratllat en una pizza crua.,3.8
Tres homes estan jugant als escacs.,Dos homes estan jugant als escacs.,2.6
Un home està tocant el violoncel.,Un home assegut està tocant el violoncel.,4.25
Alguns homes estan lluitant.,Dos homes estan lluitant.,4.25
Un home està fumant.,Un home està patinant.,0.5
L'home està tocant el piano.,L'home està tocant la guitarra.,1.6
```

Figura 4.3: Exemple d'una mostra en el corpus

Com es pot observar en la imatge, un sample és un col·lecció de 2 frases amb un valor de similitud.

- **Model:** És una modificació del model bàsic *distilbert-base-uncased* multilingüe, és a dir, agafem aquest model pre-entrenat i després farem el Fine-tuning per transformar-ho al català.
- **Fine-Tuning**: Utilitzant una modificació del scrip[1] d'entrenament gastat per al model en castellà (stsb_multi_mt.py) i el corpus en català s'entrena al model per a la seua similitud de paraua.
- **Resultat:** Un model de característiques semblant a l'utilitzat en castellà, on el percentatge d'èxit ronda un 74%.

Mètode d'Indexació

El mètode d'indexació consisteix en una sèrie de fases que comprenen des del tractament del text fins a la creació dels índex.

Tokenització i unitat bàsica:

- **Word2Vec i STSB:** En primer lloc, hem de decidir quina serà la unitat bàsica per la qual el nostre projecte indexarà, estudiant aquest projecte i tenint en compte els objectius hem decidit que la frase serà la unitat bàsica.

  Posteriorment, hem de ser capaços de transformar les paraules de les frases en tokens per poder fer una transformació explicada al següent apartat.

- **Booleà:** Hem tractat el text per eliminar paraules de poca rellevància i molt repetides. Per tant, la unitat bàsica serà la paraula.

Nova representació de la unitat bàsica

- **Word2Vec i STSB:** Una vegada tractat i dividit el text, obtenint la frase a indexar i la primera operació a fer és transformar la frase en una representació vectorial fent ús dels models carregats. En cada SRI, utilitzarem un model diferent obtenint un resultat diferent.

  En altres paraules, transformem la frase en un vector de dimensió fixa per a la seua posterior indexació.

- **Booleà:** No necessita una nova representació de la paraula. Donat que, ja ha sigut tokenitzada i es pot procedir a la indexació.

Indexació:

- **Word2Vec i STSB:** Obtenint la representació vectorial és moment de procedir a la indexació, agafant com a clau el vector i el valor el número del document utilitzarem un diccionari per fer d'índex. Obtindrem un diccionari on per fer la cerca tindrem com a clau la representació vectorial de la frase i com a valor l'ID del document que hem tractat.

  En un futur apartat (ref: [5.3.1]) analitzarem en més deteniment com es produeix la indexació, ja que no és trivial i cal utilitzar diversos índex.

- **Booleà:** Per aquest model farem ús d'una llibreria especialitzada a crear SRI de models Booleans, per mitjà d'un *schema*[2] assignarem els camps que volem que es guarden en el nostre índex, com per exemple l'ID, el **summary**, etc. aleshores en el moment d'indexar la paraula serà entesa com una clau i el *schema* com el valor. Com a conseqüència, podem obtindre una paraula en diversos *schema* assignats.

[1] codi que llança algun procediment
[2] Un esquema de com s'organitza el document

Transformació:

Aquest apartat sols fa referència als models **Word2Vec** i **STSB**, donat que en el model Booleà no és necessari. Una vegada creat l'índex agafarem les claus, és a dir, la col·lecció de representacions vectorials de les frases, i les transformarem en un *KDTree*[3] (ref: 4.3.4.2), és a dir, cada clau serà un node del KDTree. La gran particularitat d'aquest tipus d'estructura és la possibilitat d'utilitzar tantes dimensions com vulgues, és a dir, el KDTree permet fer cerques per distància cosinus eficientment.

Aquesta transformació és necessària per a ser capaços de comparar la representació vectorial de la consulta amb els nostres claus dels índex. La cerca dels documents més rellevants serà més eficient si utilitzem el KDTree, ja que ens possibilita de trobar els vectors claus més pròxims a la consulta.

Models	Unitat Bàsica	Representació	Indexació
Boolea	Paraula	Cap, sols tokenitzar	Schemas, whoosh
Word2Vec / STSB	Frase	Vectorial	Diccionaris i KDTree

Taula 4.1: Taula Resum de l'Indexador

4.2.3.2. L'algoritme de cerca i rànquing

Aquest complement sols s'ha implementat per al models Word2Vec i STSB, ja que són els únics que podem fer una comparativa del context o del significat semàntic de la consulta. El model Booleà no diferencia ni de context ni de semàntica, sols si la paraula està en un document o no. A més a més, aquest component està lligat a la classe buscadora, que és l'encarregada de tractar la consulta i llançar aquest algoritme. En aquest component també s'encarregarà de carregar l'índex.

Com hem explicat en l'apartat anterior, utilitzem el KDTree (ref: 4.3.4.2), una estructura de dates que permet configurar com a nodes els diferents vectors representatius de les frases. No sols ens permet això, sinó que també ens facilita les ferramentes necessàries per comparar distancies entre els nodes i trobar-ne els n veïns més propers. En altres paraules, gràcies a l'estructura del KDTree podem cercar el veïns més propers a un donat de forma eficient (ref: 4.3.4.3).

Com a conseqüència, obtindrem un resultat ordenat dels n veïns més proper a la consulta. En altres paraules, utilitzant el KDTree i l'algoritme *nearest-neighborhood* obtenim una llista ordenada de les frases més semblants a la consulta. I una vegada obtinguda la llista podem fer una recuperació completa dels articles que contenen aquestes frases i retornar el rànquing a l'usuari. Encara que primer de tot hem de processar la consulta per poder ser comparada en la resta del nodes de l'arbre.

4.2.4. L'Índex

Aquest component és creat per la classe Indexador, i està configurat de la següent forma:

Word2Vec i STSB:

1. Els objectes guardats són els diccionaris i el KDTree necessaris per fer l'índex (ref: 4.2.3.1).

[3] estructura de dades de k-dimensions en forma d'arbre

2. Guardat per mitjà de la llibreria *pickle*, capaç de guardar objectes sencers en un arxiu, i aquest arxiu és creat com a *.gz*, és a dir, que guardem els objectes en un arxiu comprimit per ocupar menys espai.

Booleà:

1. Escrivim en un document la matriu d'aparició dels termes en els documents, com a *schema* (ref: 4.2.3.1).

2. El document és guardat sense la necessitat de comprimir-lo.

4.2.5. Classe Buscadora

La classe encarregada de fer d'interfície amb l'Usuari, es usada per l'usuari per fer les consulta que siguen necessàries i rebre les respostes correctes.

4.2.5.1. Word2Vec i STSB:

Per mitjà d'uns paràmetres d'entrada l'usuari serà capaç de fer la consulta i rebre resposta:

- l'índex a utilitzar.
- Nom del model del SRI.
- Nombre de documents més semblants a recuperar.
- Tipus de model: 0 Word2Vec o 1 STSB
- Nom de l'arxiu on guardar els resultats
- Nom de l'arxiu on estan les consultes a fer
- *Path* del corpus en disc.

Després de tractar els paràmetres d'entrada, s'encarrega de cridar a la classe indexador i utilitzar els algoritmes de cerca i rànquing per tornar a l'usuari els documents correctes.

4.2.5.2. Booleà

Classe encarregada de carregar l'índex i processar la consulta per retornar una sèrie de documents com a resultat.

Com a entrada per consola rebrà els següents paràmetres:

- Nom de l'índex a usar
- Arxiu en les consultes a processar

Com a resultat mostra 4 articles que continguen la paraula o paraules cercades i el total de documents trobats.

4.3 Tecnologia Utilitzada

En aquest apartat farem una explicació detallada de les ferramentes i llibreries utilitzades per poder fer possible el projecte. On explicarem el perquè de l'ús de les ferramentes i quines han sigut les aportacions que han fet al projecte.

4.3.1. Entorn de desenvolupament

4.3.1.1. Visual Studio Code

Visual Studio Code (VSC) és un editor de codi font. Una eina que et proporciona un espai on escriure codi d'algun llenguatge de programació. Desenvolupada per Microsoft pot ser utilitzada en Windows, Linux o macOS. Visual Studio Code es tracta d'una ferramenta gratuïta. No sols ens facilita l'espai on escriure codi font, sinó que té altres funcionalitats essencials per desenvolupar un projecte, com per exemple, depuració suportada, finalització intel·ligent de codi i d'altres.

Visual Studio Code ens ha facilitat un entorn on poder escriure el codi necessari per poder desenvolupar tots els components dels SRI. VSC ha sigut utilitzada per escriure tot el codi dels diferents components. A més, ens ha facilitat el treball gràcies a la funcionalitat de depuració.

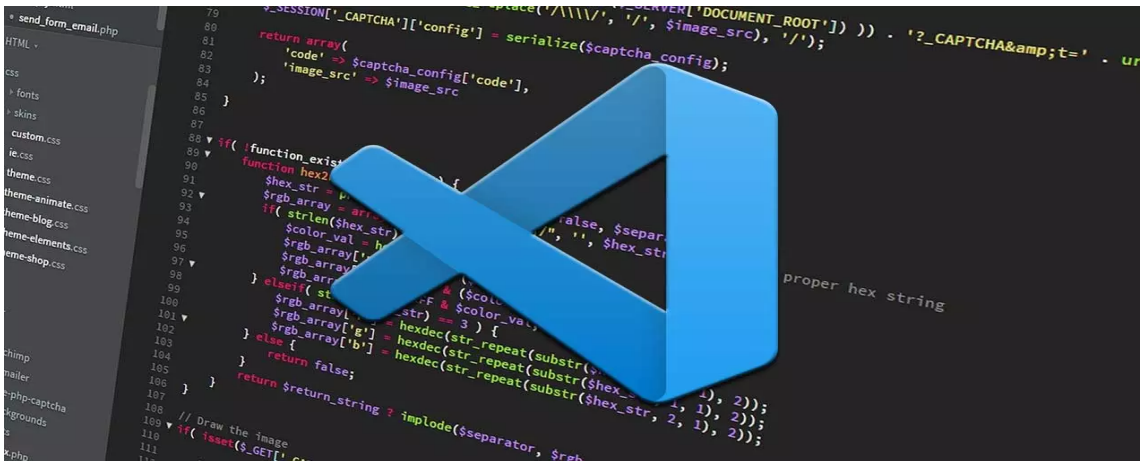


Figura 4.4: Visual Studio Code

4.3.1.2. Anaconda

Anaconda és una distribució dels llenguatges de programació Python i R, que ofereix serveix per poder simplificar la gestió i implementació de paquets. És a dir, proporciona uns serveixs mínims gratuïts que ajuden a l'organització i implementació dels paquets o llibreries que utilitzaràs en els teus projectes.

Anaconda Navigator

Anaconda Navigator és una interfície gràfica d'usuari (GUI) que està implementada en Anaconda. La gran diferència és que permet l'ús d'Anaconda i totes les seues utilitats sense la necessitat d'usar comando, és a dir, presenta una interfície des d'on poder llançar aplicacions i organitzar els paquets entre altres.

En el nostre projecte utilitzarem les dues opcions, en l'ordinador local on fer *testing*[4] utilitzarem Anaconda Navigator, ja que ens facilita el treball sense la necessitat d'utilitzar comandos, accelerant tot el procés. Per altra banda, a l'hora d'utilitzar les màquines del departament (ref: 4.3.1.3) que són més potents, però no tenen interfície gràfica, utilitzarem Anaconda per comandos. En els dos casos utilitzarem els mateixos paquets i llibreries.

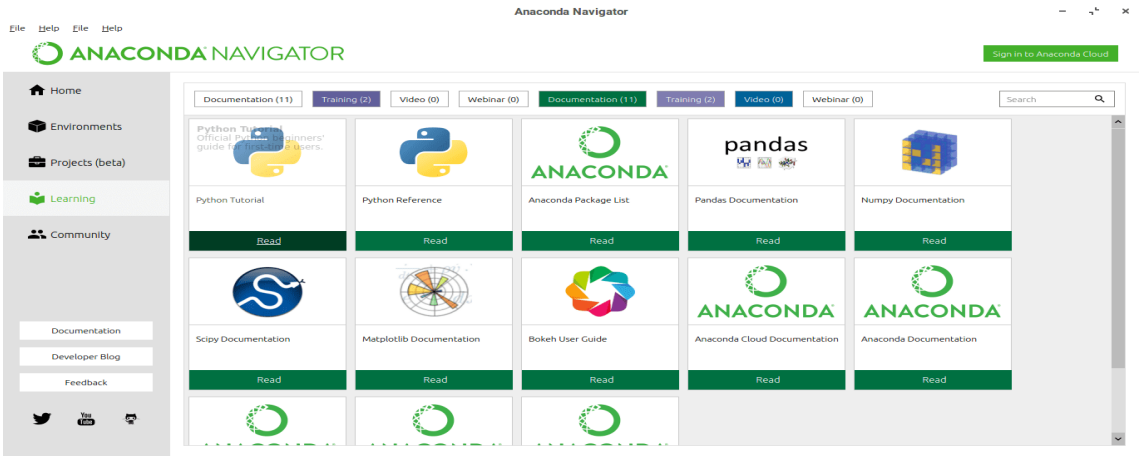


Figura 4.5: Anaconda Navigator

4.3.1.3. Tardis

Tardis és una computador del és del grup d'investigació ELiRF on s'està desenvolupant aquest treball. Tardis ens proporciona els recursos necessaris per poder tractar amb un corpus gran, i al mateix temps ens garanteix un procés eficient i ràpid en comparació a altres màquines.

Aquest projecte estarà dividit en dues etapes:

1. Treball local: Desenvolupament de tots els SRI en una escala reduïda comprovant el funcionament del treball, utilitzant un corpus de 10000 o menys documents. La màquina utilitzada és un ordinador propi amb una capacitat de recursos menors.

2. Treball en Tardis: utilitzant Tardis i amb la certesa del correcte funcionament del projecte indexarem una gran quantitat de documents fent ús dels grans recursos proporcionats.

4.3.2. Llenguatge de Programació

4.3.2.1. Python

Python és un dels millors llenguatges de programació per desenvolupar projectes PLN (ref: 2.4.1.1). Aquest projecte està desenvolupat en Python pels grans avantatges que proporciona, com per exemple, l'enorme nombre de llibreries relacionades en PLN, i com s'adapta als requisits del projecte. A més, Python ha sigut un dels llenguatges que més he treballat amb anterioritat obtenint així un gran domini del llenguatge.

Mòduls utilitzats de Python:

[4] provar els scripts creats i comprovar el funcionament

- ***Os:*** Proporciona versatilitat i simplicitat a l'hora de manejar funcionalitats del sistema operatiu. En aquest projecte ha sigut utilitzat per recórrer directoris i arxius, on hem manejat documents, escrivint, llegint-los, etc.
- ***Sys:*** Mòdul utilitzat per manipular parts de l'entorn d'execució de Python. La seua funcionalitat dins del projecte és la de proporcionar-nos els arguments d'entrada dels scripts.
- ***Json:*** Mòdul que proporciona les operacions i funcions necessàries per poder utilitzar arxius que tenen l'extensió *.jsonl*
- ***Pickle:*** Mòdul utilitzat per transformar objectes de Python en una seqüència de bytes. Per tant, proporciona la funcionalitat necessària per transformar els nostres diccionaris i arbres en bytes i guardar-los.
- ***Time:*** Mòdul de Python que ens ajuda a calcular el temps que transcorre mentre indexem o cerquem.
- ***Gzip:*** S'encarrega de la compressió dels pickles, és a dir, els arxius creats amb el pickles són comprimits pel mòdul gzip perquè no ocupen tant d'espai en memòria.
- ***Argparser:*** Mòdul que facilita l'escriptura d'interfícies de comandos fàcils d'utilitzar. Ens facilita la captació dels arguments que entren per sistema i en cas d'error la seua explicació.

4.3.3. Paquets i Llibreries

4.3.3.1. SpaCy

Llibreria orientada cap al desenvolupament de l'àmbit del PLN (ref: 2.4.2.1). En aquest projecte hem utilitzat les següents funcionalitats de la llibreria:

- **Models:** Hem fet ús de diferents models Word2Vec per poder crear el SRI:
  - ***es-core-news-md:*** Model creat i entrenat per crear embeddings Word2Vec en castellà. Compta amb diferents funcionalitats com per exemple *parser, lemmatizer, etc*. SpaCy exposa que té una precisió alta i per a 500000 paraules conte 20000 vectors únics. Va ser entrenat en el corpus en castellà d'Ancora. Els embeddings creats per aquest model tenen un dimensionalitat de 300, i per cada paraula o frase que es transforme obtindre un embedding que proporciona els pesos necessaris per a la seua futura indexació i comparació.[15]
  - ***ca-core-news-md:*** Model pre-entrenat per SpaCy amb la capacitat de crear embeddings en català de frases o paraules. Té les mateixes funcionalitats que el model creat en castellà. [16]
- *Mètodes:* Per poder utilitzar els diferents models hem hagut de fer ús dels següents mètodes:
  - ***spacy.load():*** Carrega el model a utilitzar passant-li com argument el nom i les funcions que vols que estiguen excloses o incloses. És a dir, s'utilitza per carregar el model amb tots els elements que vols inclosos.
  - ***encode():*** Com a tal no existeix el mètode escrit així, és una representació per ser entesa. S'encarrega de processar la unitat bàsica i fer la representació vectorial d'aquesta, creant així l'embedding. Aleshores com a argument és necessari passar-li la unitat bàsica siga una frase o una paraula.

4.3.3.2. NLTK

Llibreria enfocada en l'àmbit PLN (ref: 2.4.2.1). En aquest porjecte de les diferents funcionalitats que ofereix sols hem utilitzat la següent:

- ***sent_tokenize():*** Mètode que a partir d'un text passat com a argument crea una llista en totes les frases que el contenen. Per tant, la seua funcionalitat en aquest projecte és la de crear les frases a partir del cos de l'article a indexar.

4.3.3.3. HuggingFace i Sentence_Transformer

- **HuggingFace:** és un repositori de models PLN, on podem trobar tota mena de models o pujar els propis. Hem utilitzat aquesta repositori per usar els següents models:
  - **català:** *driwnet/stsb-m-mt-ca-distilbert-base-uncased* model STSB de pròpia creació partint d'un model pre-entrenat *distilber-base-uncased* (ref: 4.2.3.1).
  - **castellà:** *eduardofv/stsb-m-mt-es-distilbert-base-uncased* model STSB en castellà (ref: 4.2.3.1).
- **Sentece_Transformer:** Llibreria enfocada a l'ús dels models de *HuggingFace*, és a dir, proporciona les funcionalitats necessàries per poder utilitzar els models del repositori. Es tracta d'una llibreria enfocada a l'ús dels models BERT, i pot ser utilitzada en més de 100 idiomes diferents.
  - ***Sentence_Transformer():*** Mètode que carrega el model en memòria per al seu futur ús.

4.3.3.4. SciPy

Llibreria especialitzada en funcions matemàtiques, de ciències i de física. Proporciona diferents estructures, algoritmes i mètodes per ser utilitzades en Python, sempre des d'un punt de vista matemàtic. També conte funcions d'optimització, àlgebra, integració, etc.

- ***scipy.spatial.KDTree():*** Proporciona el mètode necessari per transformar un índex en un arbre k-dimensional per poder fer una cerca del veí més proper (ref: 4.3.4.2). També, inclou l'algoritme de cerca que utilitzarem en aquest projecte, el veí més proper (ref: 4.3.4.3).

4.3.3.5. Whoosh

La llibreria Whoosh va ser creada per Matt Chaput, es caracteritza per ser un motor de cerca ràpid i pur, creant índexs xicotets i eficients. Utilitza Python pur, és a dir, funcionarà en tots els llocs on s'utilitze Python.

La funcionalitat dins d'aquest projecte és la de creació del SRI Booleà, gràcies als seus algoritmes eficients i ràpids obtindre uns índexs de grandària reduïda i ben estructurats.

Utilitza els *schema* per estructurar l'índex a crear (ref: 4.3.4.4).

4.3.4. Estructures i algoritmes

4.3.4.1. Diccionaris

Els diccionaris són una estructura de dades capaces d'emmagatzemar diferents tipus de dades, com enters, reals, cadenes, llistes, etc. En la particularitat que ens permet identificar cada element per una clau (*key*) i assignar-li un valor (*value*). Els valors, com hem explicat abans, podem arribar a ser molt tipus de dades diferents. la seua estructura es veu de la següent forma:

$$\text{dic} = [\text{«}Key_1 : Value_1, key_2 : Value_2, ...\text{»}]$$

```
diccionario = {'nombre': 'Carlos', 'edad': 22, 'cursos':['Python','Django']}
```

Figura 4.6: Diccionari

Els diccionaris ens proporcionen una estructura de dades clau a l'hora d'indexar, ja que ens permet guardar relacionant cada representació vectorial de les frases amb el document d'on va ser extreta.

4.3.4.2. KDTree

KDTree és una estructura de dades, on l'espai és partit en diferent subconjunt que no se superposen, per organitzar diferents punt de k-dimensions.
Aquesta estructura està composta per diferents nodes, que representen punts de k-dimensions. Cada node que no es fulla suposa una separació de l'espai en 2 parts, per tant, l'espai es va anant partint fins a arribar als nodes fulles que estaran tots integrats en algun espai.

KDTree s'utilitza per cerques multidimensionals. Cerques que utilitzen algoritmes com el del veí més proper (ref: 4.3.4.3).

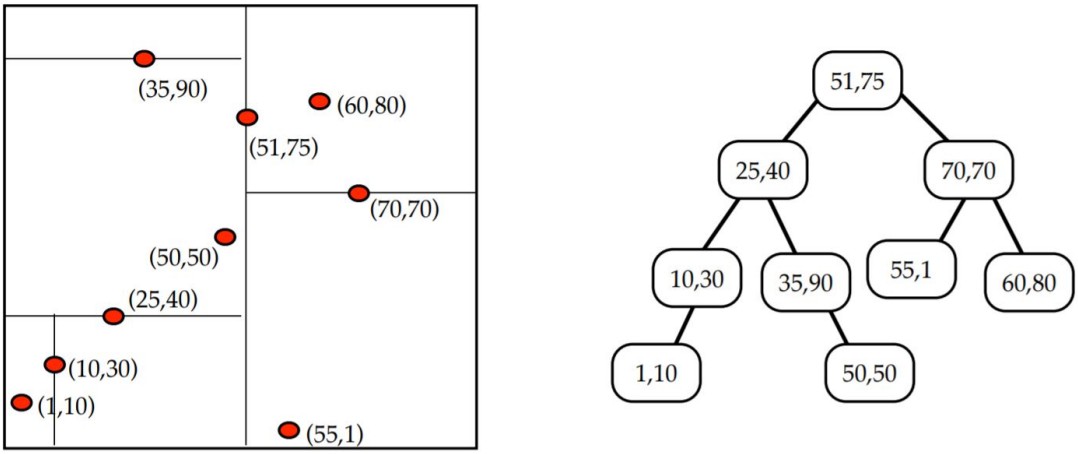


Figura 4.7: KDTree

KDTree ens ofereix una estructura de dades que pot transformar els diccionaris, o millor dit les claus dels diccionaris en un arbre on la comparació de distància entre punts

és simple i ràpida. Permetent-nos així, fer transformar la consulta en un vector de k-dimension i després comparant-lo amb els nodes de l'arbre.

4.3.4.3. Veí més proper

Algoritme que retorna el n nodes més propers a un donat. L'estructura del KDTree facilita a l'algoritme la cerca, ja que per com està estructurat l'arbre va separant el nodes entre regions (valor mitjà de les dades). Cadascuna de les regions està formada pels nodes que tenen característiques parelles o que estan més propers. En el moment de fer la cerca, la consulta estarà dins d'alguna regió i, per tant, sabent quins nodes estan dins de la regió la cerca és més simple. No obstant això, no sols cerca en la regió que està sinó en les del voltant per si hi ha algun node més proper.

L'algoritme fa ús de la **distància de Minkowski** per calcular la distància entre dos nodes de l'arbre.

Minkowski: es definix com la distància entre dos punts $X = (x_1, x_2, ..., x_n)$ i $Y = (y_1, y_2, ..., y_n)$ d'ordre *"p"* on:

$$dist\_Minkowski = (\sum_{i=1}^{n} |x_i - y_i|^p)^{1/p} \tag{4.1}$$

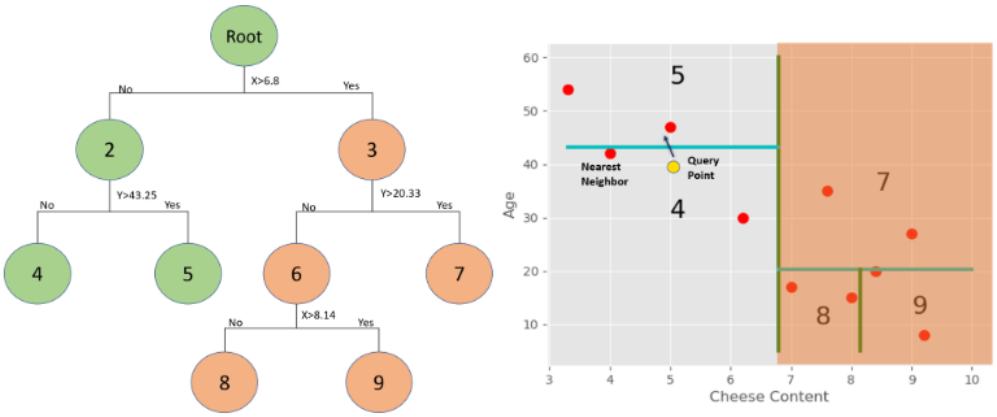


Figura 4.8: Exemple algoritme *Nearest-neighborhood*

4.3.4.4. Schema

És una estructura particular de la llibreria *Whoosh*, permet especificar i organitzar els camps dels índex, és a dir, els documents a indexar tenen camps (títol, resum, *URL*, etc.) i per mitjà del *schema* podem fer una tria de quins camps volem indexar i de quins no. Com a conseqüència, pots triar quins camps del documents es volen indexar en el document a guardar.
En el moment de la cerca, sols serà possible per als camps que has guardat, ja que els altres no pertanyen a l'índex creat.

```
schema = Schema(
    source= TEXT(stored = False),
    id= ID(stored = True),
    url= ID(stored=True),
    summary= TEXT(stored=False, analyzer=my_analyzer),
    article= TEXT(stored=False, analyzer=my_analyzer),
    article_nwords= NUMERIC(stored=False),
    summary_nwords= NUMERIC(stored=False),
    similarity = NUMERIC(stored=False),
    html_detected= BOOLEAN(stored= False),
    too_short= BOOLEAN(stored= False),
    too_similar= BOOLEAN(stored= False),
    not_ended= BOOLEAN(stored= False),
    lang= TEXT(stored = False),
    lang_prob= NUMERIC(stored= False),
    lang_es_prob= NUMERIC(stored= False),
    lang_ca_prob= NUMERIC(stored= False),
    path=ID(stored=True),
    line = NUMERIC(stored=True)
)
```

Figura 4.9: Exemple de *schema*

CAPÍTOL 5

Desenvolupament i implementació de la solució proposta

En aquest capítol detallarem quins han sigut els principals problemes que hem hagut de resoldre a l'hora de desenvolupar els projecte, i no sols això, sinó quines particularitats o punts crítics conté el codi del projecte. A més, parlarem de l'evolució del projecte al llarg del temps. En aquest capítol sols farem referència als model Word2Vec i STSB, ja que, el model Booleà no ha causat cap dificultat al ser molt simple d'implementar.

5.1 Problemes i dificultats

5.1.1. Eficiència i Memòria

Exposat en aparts anteriors un dels principals problemes que hem trobat en aquest projecte ha sigut el d'eficiència i memòria (ref: 3.1.1). Encara que l'eficiència ha sigut resolta utilitzant les llibreries més actualitzades del mercat que ens proporcionen els algoritmes més eficients, el problema de l'ús de la memòria ha sigut un mal de cap continu.

Solucions aplicades:

1. **Reduir el nombre:** Tot i que la màquina Tardis és potent en comparació als ordinadors habituals, no té la suficient memòria per a fer-se càrrec de la indexació total del corpus. Per tant, hem hagut d'anar reduint el nombre de documents a indexar:
   (a) **750000:** Capacitat total que podia indexar sense tindre en compte guardar l'arxiu, és a dir, es podia indexar fins a 750000, però si vols guardar l'índex era insuficient la memòria.
   (b) **500000:** Ocorria el mateix error que abans.
   (c) **200000:** Podia guardar-se l'índex per als models Word2Vec, en canvi, no tenia memòria suficient per als models STSB.
   (d) **100000:** Perfecta mida per als models Word2Vec i STSB, sols que s'havien de fer en 2 índex separat, és a dir, 50000 en un i en l'altre altres 50000 documents.
2. **Compressió dels arxius:** Una vegada solucionat el nombre de documents, per no ocupar tant espai en el disc, es va passar a comprimir els arxius creats per mitjà del mètode GZIP. Amb un nivell de compressió 3.
3. **Utilitzar tardis a hores poc concorries:** Donat que Tardis és utilitzat per altres persones que podien consumir recursos, vàrem prendre la decisió d'utilitzar Tardis a la nit, ja que és l'horari on menys gents l'utilitzava.

5.2 Implementació dels Components del SRI

Utilitzant Visual Studio Code i Python hem fet la implementació de totes les classes que teníem pensades (ref: 4.2). En total 5 classes:

- **SRI Boleà:**
  - **Classe indexador (indexador_word):** Classe implementada amb anterioritat en l'assignatura de SAR. Utilitza la llibreria whoosh i l'estructura de dades schema per indexador tota la col·lecció de documents. Guarda en un document quina ha sigut la demora total.
    1. **Indexador:** Mètode encarregat de la indexació del corpus fent ús de l'estructura schema per escriure en un documents quins camps dels articles volem guardar. Recorre totes les paraules del corpus per anar indexant una a una.
  - **Classe Búscadora (buscador):** Classe implementada seguint un projecte de SAR, prou simple d'implementar, utilitza l'índex anterior creat i el parser per fer la cerca de la paraula.
- **SRI Word2Vec i STSB:**
  - **Classe Llançadora (TFG_Indexador_nou):** Encarregada d'arreplegar els arguments per línia de comando i llançar les operacions necessàries per crear els índexs (ref: 4.2.2). Utilitza el mòdul argparse per agafar els arguments. Primer de tot crea l'indexador important-ho de la classe indexador i després llança les operacions indexar(), save() i vaciar() de l'objecte Indexador.
  - **Classe Indexador (Indexador_KDTree_nou:** Classe on es produeix la majoria de processos. Els mètodes que té son els següents:
    1. **__init__():** Mètode constructor on inicialitzem els diccionaris i el arbre a , i el model a NULL.
    2. **indexar (path de l'índex, model a usar, idioma, 0 = Word2Vec o 1 = STSB, document on escriure resultats):** Mètode encarregat de la indexació del corpus. Explicat en detall en l'apartat 5.3.1
    3. **save():** Guarda els diccionaris i arbres creats per mitjà del mòduls GZIP i Pickle. També calcula el temps de guardat.
    4. **vaciar():** Neteja els diccionaris, arbres i models per utilitzar-los de nou i que no hi haja contaminació de dades.
    5. **buscador (query, nom_model, num_doc = documents a retornar, opcio = STSB o Word2Vec, output):** Mètode que escriu en un document quins han sigut els resultats del processament de la consulta. Explicat en detall en l'apartat 5.3.2.
    6. **load (nom_index):** Carrega l'índex en memòria agafant el seu *path*.
    7. **tallar (text):** Transforma el cos de l'article en frases.
    8. **calcular_valor (frase, opcio):** Calcular la representació vectorial de la frase passada, l'argument opció és 0 per al Word2Vec i 1 per als STSB.
    9. **cargar_model(nom_model, opcio):** Carrega en memòria i en la variable *self.PLN* el model a utilitzar, opció Word2Vec o STSB.
    10. **eliminar_stopwords(frase):** s'encarrega d'eliminar les stopwords del input i retornar el resultat obtingut.

– **Classe buscadora:** Agafa per línia de comandos el nom de l'índex a usar, el model, el valor del veïns a retornar, STSB (1) o Word2Vec (0), el document on escriure els resultats i el document on estan les consultes, recorre les consultes resolent-les i escrivint el resultat en el document d'eixida.

5.3 Particularitats o punts crítics

En aquesta secció recalcarem el procés d'indexació i cerca, ja que les estructures creades per fer-ho són essencials per entendre com funciona el projecte.

5.3.1. Procés d'indexació

Per entendre com funciona el procés d'indexació primer hem d'entendre quines estructures de dades hem creat i perquè.

1. **valor_doc:** Per cada document que hem indexat hem assignat un valor únic, per poder dotar-li d'un id i en un futur poder recuper l'article correcte. Un id per a cada article.
2. **dic_doc:** Diccionari on les *keys* estan representades pel valor_únic dels documents i el seu valor assignat són el *path*[1] al document. Utilitzat per una vegada recuperat el valor_únic en la cerca saber quin és el document que estem buscant i on es troba.

$$\text{dic\_doc} = [\text{«}valor\_unic_1 : path_1\text{»}, \text{«}valor_unic_2 : path_2\text{»}...]$$

3. **dic_text:** Diccionari compost pels vectors representatius de les frases com a *keys* i el valor_únic del documents mes quina posició té la frase a indexar dins de l'article. Utilitzat per una vegada transformada la consulta en un vector i com a conseqüència la llista del n veïns més propers, recuperar en els vectors representatius els valors únics associats.

   $\text{dic\_text} = [\text{«}vector_1 : (valor\_únic_1,1)>>, <<vector_2 : (valor\_únic_1,2)>> ...]$

4. **tree:** KDTree que agafa els vectors representatius de totes les frases indexades i les transforma en els seus nodes. Simplificant-nos la cerca en un futur, ja que ens permet conèixer quins són els n veïns més propers a un vector donat.

Aleshores el funcionament és el següent (ref: A.1.1):

1. Anem recorrent directoris i subdirectoris fins a trobar l'arxiu .jsonl que conte tots el articles a indexar.
2. A cada article a indexar li assignem un valor, un id únic.
3. En el diccionari dic_doc indexem els id amb el *path* i la línia on es troba l'article.
4. Per cada article extraem les frases i les anem transformant en la seua representació vectorial.
5. Per cada representació li assignem el valor del id del document i quin número de frase és.

[1] especifica les rutes on estan els articles en el sistema

6. Seguidament, totes les representacions vectorials serveixen com a nodes en el KDTRee.

7. Per últim, guardarem els diccionaris i l'arbre en disc.

5.3.2. Procés de cerca

El procés de cerca compartix elements amb el procés d'indexació. El seu funcionament és (ref: A.1.2):

1. Primer de tot carregarem l'índex i el model a utilitzar. És a dir, carregarem en memòria els diccionaris i arbres que hem creat amb anterioritat.

2. Una vegada tinguem l'índex carregat anirem recorrent la llista de consultes, passarem a processar-les, és a dir, transformarem la consulta en una representació vectorial usant el model.

3. Seguidament utilitzant el mètode query del arbre KDTree podem extraure els n nodes més similars, és a dir, els n vectors representatius més propers a la consulta.

4. Recorrerem els vectors més propers, i per cada un d'ells extraurem el valor únic associat en el diccionari dic_text. Introduïm la clau, la representació vectorial, i obtenim el id del articles mes quin número de frase és.

5. Una vegada tenim el id, podem extraure el *path* de l'articles i carregar-lo en memòria.

6. Tenint el cos de l'article, podem tornar a processar-lo i com tenim el número de la frase, podem extraure-la.

7. Per últim, una vegada extreta la frase i recuperat l'article, podem tornar-ho a l'usuari.

En conclusió, hem fet una detalla explicació de com funcionen els dos processos més importants en aquest projecte, també hem contat com utilitzem les estructures de dades i com es relacionen entre si.

CAPÍTOL 6
Comprovació del funcionament

L'objectiu d'aquest capítol és l'avaluació del funcionament del sistema i del seu rendiment per a col·leccions de documents de diferents grandària. Per portar a terme aquest avaluació es varen fer proves amb 4 tipus de grandària de corpus diferents, mesurant el temps d'execució de cada procés i els resultats obtinguts.

Primer de tot, la realització de les proves ha sigut duta a terme per a 4 grandàries de corpus (169,1000,10000,90000/100000). 169, 1000, 10000, per comparar el rendiment de la implementació en totes aquestes i poder fer una comprovació del correcte funcionament. Una vegada comprovat el funcionament, utilitzar el corpus de 100000 per a castellà i 90000 per a català per tal de fer les proves finals. Ja que, l'objectiu final és extraure resultats amb els corpus de 100000 i 90000

Les proves consten de 3 fases per cada índex a crear:

1. Fase 1: Una comprovació del correcte funcionament de la indexació dels corpus i mesurar el temps d'indexació. És a dir, si la classe indexador funciona correctament, el preprocessament dels documents i el mètodes d'indexació i fer una comparació dels temps d'execució de cada model.
2. Fase 2: *Testing*[1] de la classe buscador per als diferents índexs creats.
3. Fase 3: Una comprovació manual dels resultats extrets, per tal de valorar si són adequats a la consulta realitzada.

Les dues primeres fases són una comprovació tècnica del codi i dels temps d'execució, és a dir, comprovarem si el codi implementat té un funcionament correcte. I l'última fase és on concloure'm si els models creats tornen els resultats esperats, o ha ocorregut un error en els nivells de ponderació de frases, la unitat seleccionada, o també un possible fallo de creació.

Un dels motius d'anar augmentant el nombre de documents a indexar és per extraure en quin tipus temps d'execució treballem (logarítmic, exponencial, etc.). Un altre motiu és que per a resoldre problemes d'implementació utilitzar un corpus menor comporta un reducció en el temps a gastar.

6.1 Indexació del Corpus

La secció constarà d'una comparativa entre els temps d'indexació i emmagatzematge dels diferents corpus amb els diferents models. En altres paraules, durant el procés d'inde-

[1] comprobació del funcionament

xació dels diferents corpus amb els diferents models hem calculat la demora de cada un, primer en la indexació i segon en l'emmagatzemament.

6.1.1. Temps d'indexació

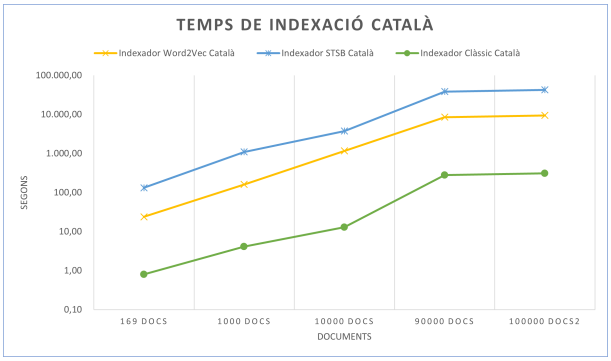
La taula 6.1 mostra el temps d'indexació dels diferents sistemes:

Temps d'indexació (segons)	169 docs	1000 docs	10000 docs	90000 docs	100000 docs
Word2Vec castellà	6,4	33,7	665,9	-	7543,7
STSB castellà	40	240,1	2206,7	-	26357,4
Clàssic castellà	0,38	2,2	9,1	-	252,4
Word2Vec Valencià	23,8	161,4	1163,3	8491,1	-
STSB Valencià	132,4	1096,4	3747,6	38406,6	-
Clàssic Valencià	0,8	4,1	12,9	279	-

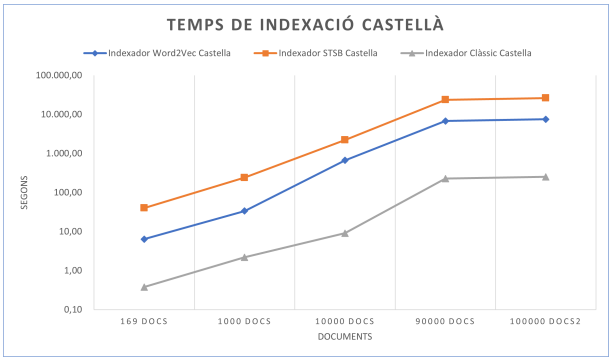
Taula 6.1: Temps en segons d'indexació dels diferents sistemes SRI

Com podem observar la taula detalla els temps d'execució a l'hora d'indexar el corpus en els diferents models. Com a particularitat, observem que els models en castellà no tenen un corpus de 90000 docs i els de català de 100000, ja que a causa d'un problema de memòria amb l'ordinador utilitzat en el desenvolupament, no hem pogut indexar més documents.

Analitzant la taula podem observar un comportament curiós en l'evolució del temps, com en talles petites el creixement pareix exponencial, però, en canvi, en talles grans el creixement pareix lineal.



(a) català

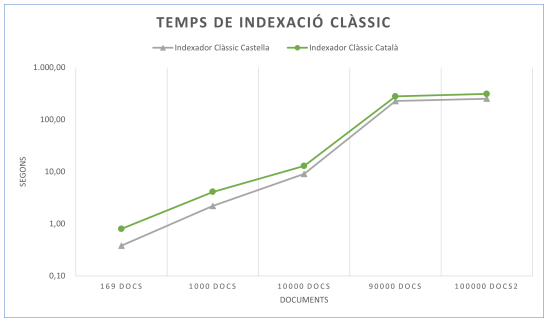


(b) castellà

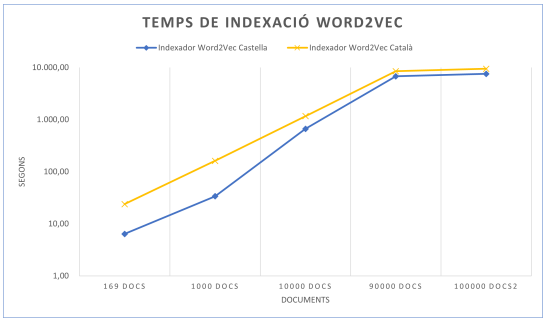
Figura 6.1: Gràfiques del temps d'indexació models en català i castellà

Aquestes gràfiques estan fetes utilitzant una escala logarítmica, per tal d'il·lustrar un creixement real del temps.

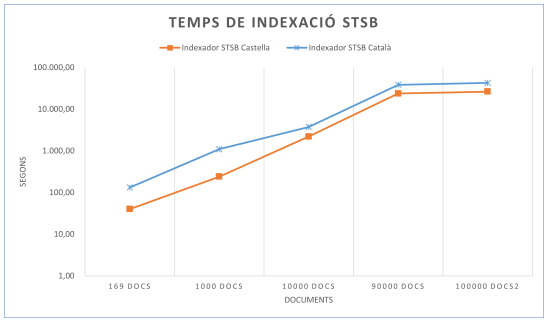
Les següents gràfiques ens mostren una comparativa entre els diferents idiomes utilitzant un mateix model i corpus.



(a) Booleà



(b) Word2Vec



(c) STSB

Figura 6.2: Gràfiques comparatives del temps d'indexació entre els models en català i castellà

Les primeres conclusions que podem extraure d'aquesta taula i gràfics són:

1. El temps d'execució creix linealment amb el nombre de documents a indexar, com es pot comprovar en les gràfiques, encara que, com hem comentat amb anterioritat, en talles menors el creixement pareix exponencial.

2. Els Indexadors més pesats són el que usen el model STSB, ja que, creen representacions vectorials de dimensió 769 que són més pesades que les representacions de 300 dimensions per al Word2Vec o paraules per al clàssic. Addicionalment, el model STSB és el més complex de tot. Trobem una gran diferencia entre els models que fan representació vectorial de la unitat i el que no ho fan, en termes de temps d'execució.

3. Les últimes gràfiques ens mostren com l'idioma català té un temps d'indexació superior en tots els models, a cause de diferents factors, com una eficiència menor en els models, uns documents en un nombre mitja de frases o paraules majors que el de castellà, etc.

Corpus	Català	Castellà	Increment
169 documents	26,69	11,11	140%
1000 documents	23,75	10,70	122%
10000 documents	20,21	11,08	82%
90000/100000 documents	17,05	13,67	25%

Taula 6.2: Nombre mitjà de frases per document

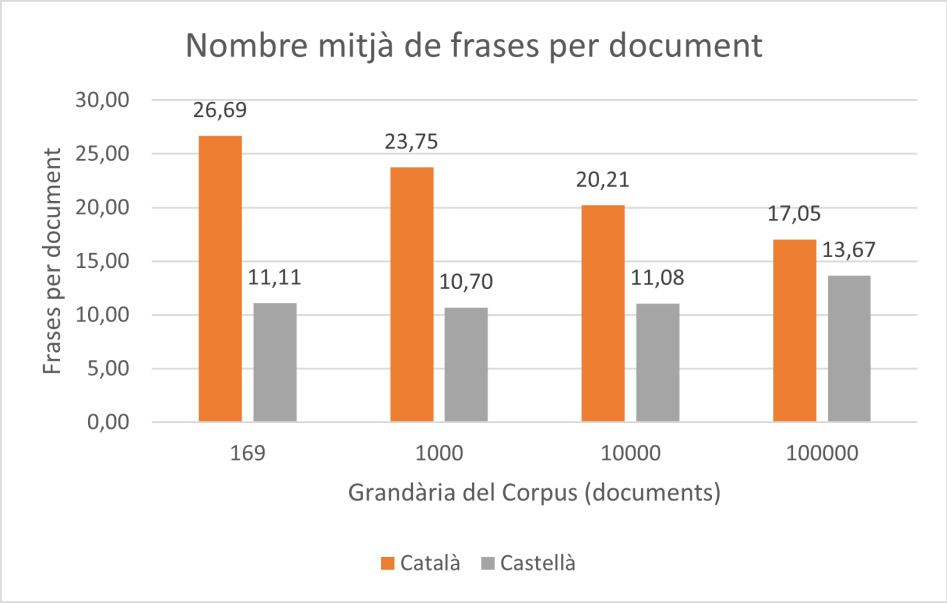


Figura 6.3: Comparació entre nombre mitja de frases per document

6.1.2. Temps d'emmagatzematge

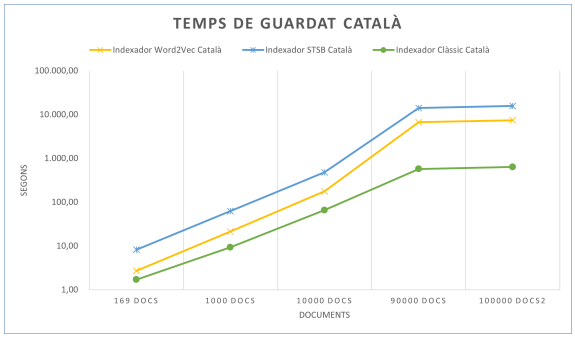
El temps d'emmagatzematge fa referència al temps transcorregut mentre els índex són guardats en disc, és a dir, el procés de transformació de l'índex en un fitxer en memòria secundària.

La següent taula 6.3 mostra el temps d'execució de l'emmagatzemament i les gràfiques comparatives, les primeres entre el mateix idioma i diferents models i les segones els mateixos models diferents idiomes:

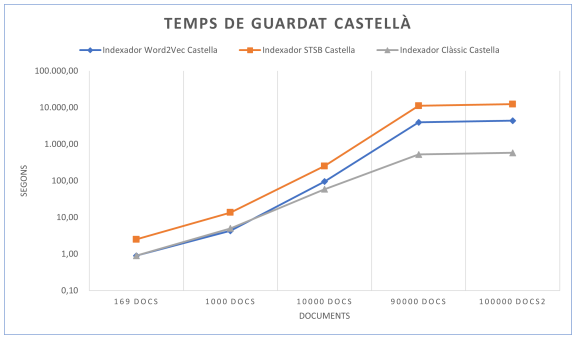
Temps d'emmagatzematge (segons)	169 docs	1000 docs	10000 docs	90000 docs	100000 docs
Word2Vec castellà	0,90	4,30	95,20	-	4.387,00
STSB castellà	2,50	13,60	251,00	-	12.399,90
Clàssic castellà	0,90	5,00	58,50	-	579,70
Word2Vec català	2,70	21,30	175,90	6.676,70	-
STSB català	8,20	62,30	479,50	14.092,60	-
Clàssic català	1,70	9,30	65,40	571,70	-

Taula 6.3: temps en segons d'emmagatzematge dels diferents models SRI

Deduïm de la taula que els temps d'emmagatzematge no depèn directament dels models o dels idiomes, sinó de la quantitat de paraules, vectors, etc. a guardar, ja que, la diferència de temps no es veu tan alterada com en la Taula 6.1 on els models havien de crear les representacions vectorials, sinó que en aquest part del procès l'única variable que canvia són el nombre de paràmetres a indexar.



(a) català

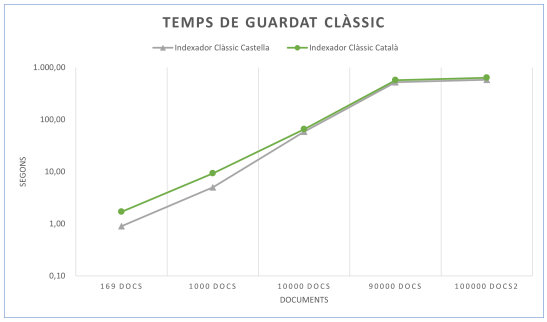


(b) castellà

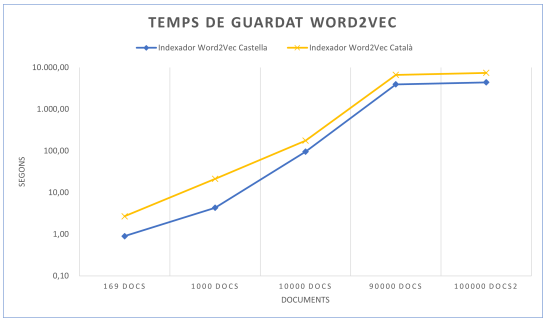
Figura 6.4: Gràfiques del temps d'emmagatzematge dels models en català i castellà

Com podem observar les gràfiques ens mostren que la deducció anterior és correcta, ja que en ser tan similars podem deduir que l'única variable d'importància és el nombre d'objectes a emmagatzemar. Els seguiments en les gràfiques mostren un clar augment de temps consumit en utilitzar models que necessiten una gran quantitat d'objectes (vectors) o de grans dimensions. Ja que, els models com STSB creen vectors de 768 dimensions, model Word2Vec de 300 dimensions i el Booleà indexa paraules.

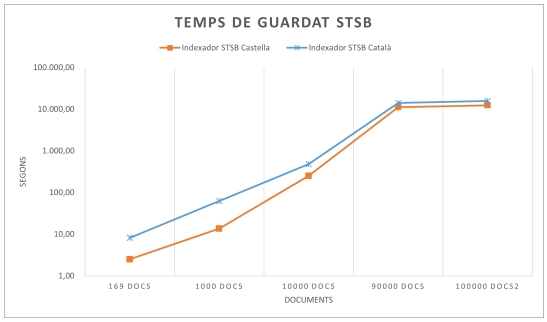
$$temps\_guardat(STSB) > temps\_guardat(W2V) > temps\_guardat(Boole)$$



(a) Booleà



(b) Word2Vec



(c) STSB

Figura 6.5: Gràfiques comparatives del temps de guardat entre els models en català i castellà

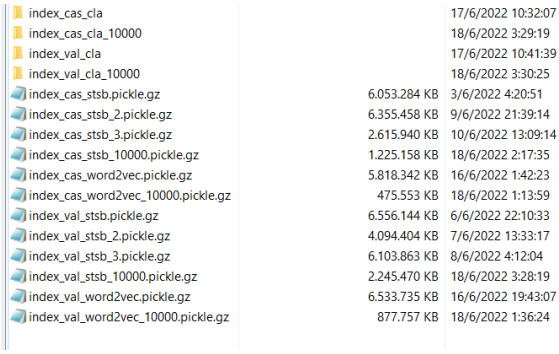
Aquestes últimes gràfiques ens demostren que les nostres deduccions anteriors són correctes, donat que les gràfiques dels diferents models mostren una similitud entre els dos idiomes. Es pot observar una xicoteta diferencia entre els temps de guardat dels dos idiomes per a cada model, això és causat pel fet que hi ha una mitjana de frases per document superior en el corpus de català.

6.1.3. Resultats de la indexació

Una vegada generats tots els índex, hem obtingut una sèrie de fitxers binaris comprimits en el format *gz*, un per cada. És a dir, hem obtingut els índex per cada un dels models en els idiomes corresponents.

Com a conseqüència hem obtingut els següents índexs:

(a) Índex 169 i 1000



(b) Índex 10000 i 100000

Figura 6.6: Exemples dels índexs creats per tots els corpus usats

En els cas dels índex del corpus de 100000 documents hem hagut d'indexar diversos subíndex per completar el final. Aquest fenomen està produït per la falta de recur-

sos a l'hora de tractar una quantitat enorme de documents. En detall, el fenomen és produït per mantenir en memòria principal l'índex creat i, al mateix temps, el fitxer binari on està emmagatzemant-se, que creix amb cada iteració d'emmagatzemament. Es pot observar en la Figura 6.6b els arxius *index_cas_stsb.pickle, index_cas_stsb_2.pickle i index_cas_stsb_3.pickle* formen l'índex *stsb* per a castellà complet, també ocorre el mateix en el cas per a català.

Concloent aquest apartat podem afirmar que el procés d'indexació funciona correctament, però com hem explicat al principi del capítol (6) no podem concloure que funciona correctament tot el projecte fins que no avaluem els resultats obtinguts. Per ara, sols podem afirmar que els índex estan creats i la implementació funciona amb eficiència.

6.2 *Testing* de l'obtenció dels resultats

Com en l'apartat anterior comprovarem la implementació del codi, per mitjà de la seua eficiència amb el temps d'execució i els resultats obtinguts.

6.2.1. Temps de Càrrega

El temps de càrrega referència el temps necessari per carregar en memòria els índexs creats amb anterioritat. Cal tindre en compte que utilitzem un mateix mètode per a tots, és a dir, utilitzem *load()* del mòdul *pickle*.

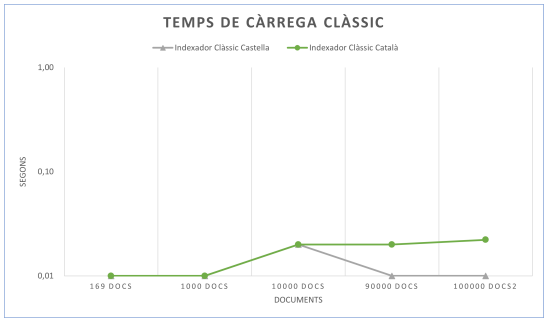
A continuació tenim la taula del temps de càrrega amb les posteriors gràfiques comparatives:

Temps de Càrrega (segons)	169 docs	1000 docs	10000 docs	90000 docs	100000 docs
Word2Vec castellà	0,01	0,97	14,91	-	712,46
STSB castellà	0,22	2,22	36,95	-	1.792,94
Clàssic castellà	0,01	0,01	0,02	-	0,01
Word2Vec català	0,31	3,55	27,87	1.262,68	-
STSB català	0,79	8,57	71,60	2.227,81	-
Clàssic català	0,01	0,01	0,02	0,02	-

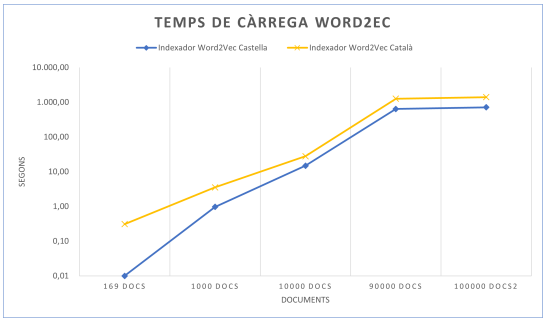
Taula 6.4: Taula dels temps de càrrega en segons de cada índex

(a) català **(b)** castellà

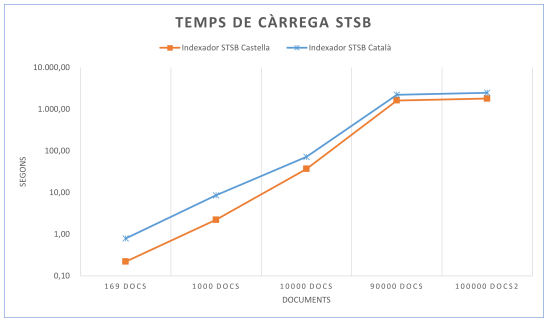
Figura 6.7: Gràfiques del temps de càrrega dels models en català i castellà



(a) Booleà



(b) Word2Vec



(c) STSB

Figura 6.8: Gràfiques comparatives del temps de càrrega entre els models en català i castellà

Aquest procés de càrrega resulta molt similar al procés d'emmagatzemament, per tant, podem deduir que el fenomen que produeix aquesta similitud és que la característica que determina el temps de càrrega són els paràmetres que hem de carregar en memòria.

En conclusió, el procès de càrrega es tracta d'un procés lineal que creix amb el nombre d'objectes.

6.2.2. Temps de cerca

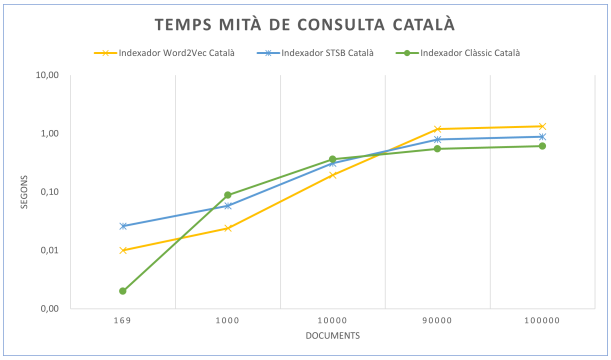
En aquest apartat tractarem l'eficiència del procés de cerca i rànquing. De cada índex creat hem calculat el temps d'execució mentre cerca el documents i crea el rànquing.

En la següent taula es mostren els resultats:

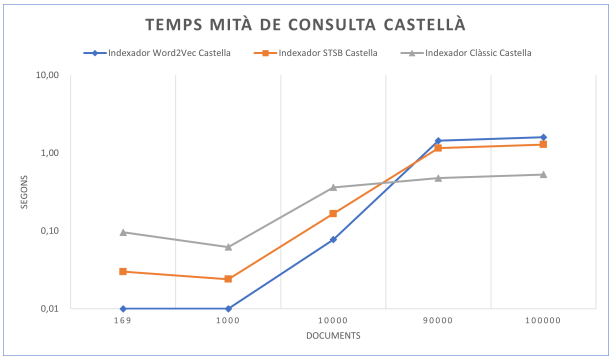
Temps de mitjà de consulta	169 docs	1000 docs	10000 docs	90000 docs	100000 docs
Word2Vec castellà	0,01	0,01	0,08	-	1,60
STSB castellà	0,03	0,02	0,17	-	1,28
Clàssic castellà	0,10	0,06	0,36	-	0,53
Word2Vec català	0,01	0,02	0,19	1,19	-
STSB català	0,03	0,06	0,31	0,79	-
Clàssic català	0,00	0,09	0,36	0,55	-

Taula 6.5: Temps, en segons, d'execució mitjà de la cerca per cada model

Primer la taula és una mitjana dels resultats obtinguts en cada consulta que podem observar-los en B.1. Segon, els resultats mostren un temps d'execució reduït i eficaç. En pocs casos estem parlant de mes d'un segons de demora.



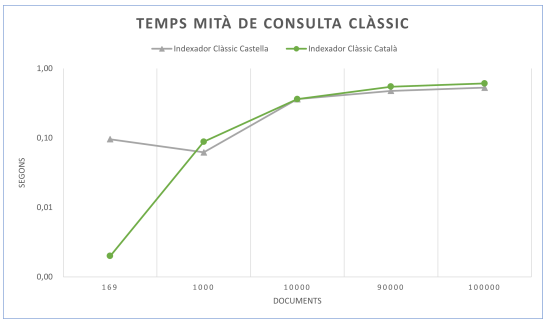
(a) català



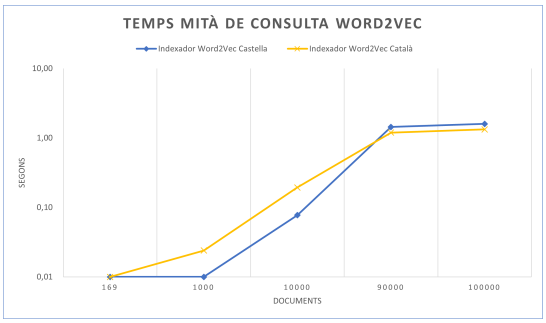
(b) castellà

Figura 6.9: Gràfiques del temps de cerca dels models en català i castellà

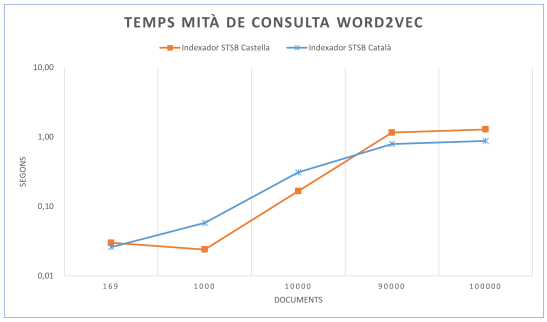
A diferència de les gràfiques mostrades amb anterioritat, aquest mostren en consultes en corpus reduïts els algoritmes de cerca dels Word2vec i STSB són més eficients que el Booleà, tenint en compte que els índexs creats per al Word2Vec i STSB són més grans. En canvi, quan tractem en corpus grans poden observar que el temps de cerca dels models amb representacions vectorials són superiors, donat que el nombre de vectors és superior al nombre de paraules indexades. Aquest fenomen és causat pel fet que l'algoritme de cerca del KDTree és logarítmic, per als sistemes Word2Vec i STSB, i, per al sistema Booleà, l'accés a les paraules que d'estan cercant és constant.



(a) Booleà



(b) Word2Vec



(c) STSB

Figura 6.10: Gràfiques comparatives del temps de consulta entre els models en català i castellà

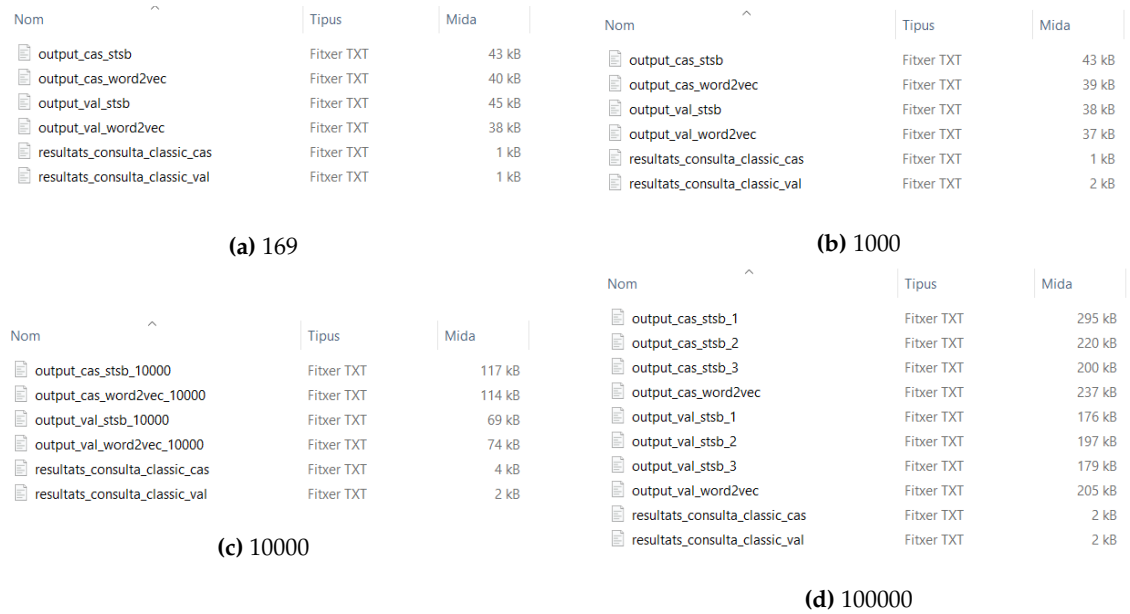
Les gràfiques ens mostren com l'idioma no és cap factor determinant en el procés de cerca, ja que els temps de cerca són molt similars en tots els casos.

En conclusió, podem deduir que els factors més determinants a l'hora de ser eficients en la cerca són els nombre de documents a recuperar i el nombre de vectors o paraules indexades. Com podem comprovar en els gràfics i la taula anterior tenim una alta eficiència

en la cerca dels documents, com a conjunt la cerca i el rànquing la demora principalment la trobem en carregar l'índex a memòria.

6.2.3. Resultats de la cerca

Detallarem quins han sigut els resultats de les consultes fetes, no entrarem en detall de l'avaluació de les consultes ni del significat, sols ens centrarem en si els resultats han sigut generats. En un futur apartat tractarem l'avaluació dels resultats.



(a) 169 **(b)** 1000 **(c)** 10000 **(d)** 100000

Figura 6.11: Gràfiques comparatives del temps de cerca entre els models en català i castellà

La Figura 6.11 ens mostra que per tots els índexs creats s'han creat un arxiu on es documenta tots els documents recuperats. L'estructura dels arxius és de la següent forma:

1. Quina és la consulta
2. Quin ha sigut el temps de demora
3. N documents recuperats depenent del nombre de veïns mes propers. Per cada article s'ha recuperat quin és el id que l'identifica, el resum, la línia més pareguda a la consulta i la distància entre la línia i a la consulta.

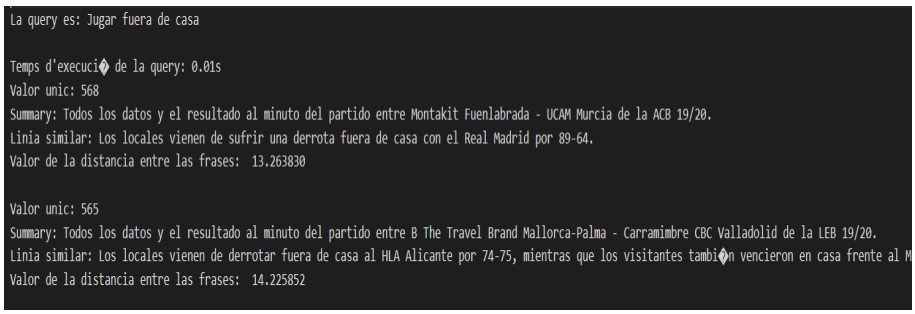


Figura 6.12: Exemple d'un resultat d'una consulta

6.3 Avaluació de la qualitat dels resultats

La secció consistira en una explicació detallada dels resultats obtinguts per cada corpus i models utilitzats. En cada resultat comprovarem el correcte funcionament del model i a posteriori una comparació dels models usats, així demostrarem el poder de recuperació de cada model.

Tanmateix, aquesta primera avaluació serà una comprovació del correcte funcionament del projecte, en un pròxim apartat tractarem de trobar consultes deterministes que mostren les diferències entres els models per a una futura comparació. És a dir, trobar consultes més similars o amb característiques més concretes que ressalten el poder semàntic del STSB, per exemple.

6.3.1. Consultes i configuració

Per produir els resultats hem utilitzat les següents consultes:

Corpus de 169 i 1000 articles:

- **català:**
  1. El partit d'esquerres va pedre les eleccions.
- **castellà:**
  1. Jugar fuera de casa
  2. Jugar en casa

Com es pot comprovar són consultes simples i enfocades als diferents corpus que teníem. Per a català utilitzem un corpus enfocat en articles polítics i econòmics. D'altra banda, per a castellà usem un corpus de notícies relacionades en el món esportiu.

Corpus de 10000 i 100000 articles:

- **català:**
  1. Només 100 vots a favor de la nova llei.
  2. L'economia creix en aquest darrer període.
- **castellà:**
  1. La economía crece en este último período
  2. El juez decreto fallo en la sentencia

Consultes molt genèriques que estan enfocades en temes polítics i econòmics.

La configuració fa referència a quins han sigut el nombre d'articles recuperat per consulta:

- **169 i 1000 articles:** hem utilitzat un nombre de 20 articles per notícia, és a dir, els 20 veïns mes propers. No sempre es recuperen 20 articles, ja que pot ser que alguns veïns siguen frases d'un mateix article.
- **10000 i 100000 articles:** Hem utilitzat un valor de 50, és a dir, els 50 veïns més propers a la consulta.

6.3.2. Resultats

Finalment, comprovarem el correcte funcionament del projecte observant els resultats. En aquest apartat ens centrarem en l'avaluació manual dels primers resultats seguint la següent directriu: si el seu funcionament s'ajusta al teòric. És a dir, si els resultats retornen documents que s'ajusten a les pautes de la consulta.

En l'apèndix B.2 trobarem una xicoteta mostra dels resultats obtinguts de les consultes anterior, on l'estructura és un rànquing dels 5 primers articles extrets per cada model en cada corpus indexat.

Ens centrarem en la comprovació dels resultats per al corpus de 100000 articles, encara que en l'apèndix s'inclouen resultats per a tots els corpus.

6.3.2.1. Model STSB castellà

Utilitzant la consulta *"El juez decreto fallo en la sentencia"* i analitzant els resultats obtinguts en la taula B.14, podem deduir un correcte funcionament del model, ja que els resultats obtinguts concorden adequadament amb la consulta. Exemple:

"El juicio quedó visto para sentencia

Analitzant-la, trobem que és una frase prou similar a la consulta feta, que el significat semàntic es té en compte (*juicio* $\approx$ *juez*) i l'aparició de la mateixa paraula (*sentencia*).

Per altra banda, la distància de similitud entre la consulta i línies que han fet *matching* són la característica principal a l'hora de fer el rànquing i com es pot observar a la taula B.14 i analitzant les frases concloem que l'algoritme de rànquing funciona correctament.

Com a conclusió, després d'analitzar els resultats podem afirmar un correcte funcionament del model.

6.3.2.2. Model Word2Vec castellà

El funcionament d'aquest model està comprovat indirectament gràcies al model que hem utilitzat de SpaCy, ja que, en la pàgina oficial exposa els grau d'encert i el seu correcte funcionament. [https://spacy.io/usage/linguistic-features#vectors-similarity]

No obstant, analitzant els resultats obtinguts de la consulta *"El juez decretó fallo en la sentencia* podem comprovar com les frases obtingudes són prou similars, exemple:

"El juez Ángel Patín dictará la sentencia"

Observem com el model ha tingut en compte el context de la consulta per extraure la frase, comprovant així el correcte funcionament.

Per altra banda, l'algoritme de rànquing és el mètode de veí més proper del KDTree, ja que retorna una llista ordenada del articles amb més semblança.

6.3.2.3. Model Booleà castellà

En aquest cas, en usar una llibreria especialitzada (*whoosh*) en aquest tipus de SRI i model, el funcionament està molt refutat.

6.3.2.4. Model STSB català

Aquest model és el que genera major interès de tots els creats, a causa d'haver sigut desenvolupat completament per aquest projecte, ja que no existia cap model STSB creat per a l'idioma en català.

Primer analitzarem els resultats retornats B.16 de la consulta "Només 100 vots a favor de la nova llei".

Aquest son els 2 primers resultats:

1. Només si aquestes esmenes prosperen acabarà votant a favor de la proposició de llei
2. El 'sí'-la proposta de Duran- va guanyar per 95 vots
3. ...

Analitzant-los podem observar com hi ha una concordança en el tòpic a tractar, una similitud semàntica en les frases ($proposicidellei \approx novallei$; $95vots \approx 100vots$; $afavor \approx el'si'$) i paraules iguals ($vots, afavor, llei, noms$), per tant, podem deduir una bona capacitat del model a l'hora de trobar similitud entre les frases i la consulta. No sols, té en compte característiques contextuals sinó que també semàntiques.

Com a conclusió, podem observar el correcte funcionament del model, ja que funciona tant la part d'extracció i comparació de similitud com la de rànquing.

6.3.2.5. Model Word2Vec català

Aquest model és molt paregut al model en castellà, ja que, utilitzem un model preentrenat per SpaCy, ca_core_news_md 4.3.3.1, per tant, podem suposar un correcte funcionament del SRI seguin un model Word2Vec.

La consulta és "Només 100 vots a favor de la nova llei"

Resultats:

1. Precisament, aquest dimecres la CUP donarà suport a través de dos vots a favorables a la llei de pressupostos.
2. La setmana passada, Jeremy Corbyn va ordenar als seus diputats de votar a favor de la llei de l'article 50.
3. ...

Fàcilment, comprovem com els resultats obtinguts tenen una similitud contextual amb la consulta feta. No sols això, si no és aquest cas podem observar una gran comparació com *vots favorables* $\approx$ *vots a favor*.

Finalment, l'algoritme de rànquing retorna una llista ordenada per la distància entre el nodes (les representacions vectorials transformades en nodes) i amb el correcte funcionament de l'algoritme de cerca i l'índex comproven un funcionament adequat.

6.3.2.6. Model Booleà català

Igual que en el cas del model booleà per a castellà, es tracta d'un model creat a partir de la llibreria *Whoosh* i, per tant, tot el procés ha sigut comprovat amb anterioritat, és a dir, els resultats són correctes i adequats.

6.4 Conclusió Proves

Una vegada analitzats els resultats dels diferents models en els diferents corpus, els temps d'execució dels diferents processos del projecte i comparar-los entre els idiomes hem comprovat que tots els SRI creats en aquest projecte funcionen adequadament.

L'objectiu de les proves fetes és la comprovació del correcte funcionament dels models en català per al corpus final (90000 documents). Per poder fer-la hem usat dues tàctiques fonamentalment:

1. Una comparació directa de l'eficiència i resultats amb el model castellà, ja que ha sigut creat i comprovat amb anterioritat.
2. Ús de corpus xicotets per comprovar el rendiments i resultats són adequats en tots els casos.

Gràcies a aquestes tàctiques, hem pogut demostrar com no hi ha una diferència plausible entre els models en castellà i català, ja que, estem parlant de termes similars quan analitzem l'eficiència i el rendiment. A més a l'hora de comparar els resultats podem observar com els resultats donats per model STSB en català i els resultats en castellà són semblant salvant les distàncies. Tanmateix, examinant els resultats obtinguts pel model Word2Vec reparem que alguns resultats se superposen amb els resultats obtinguts en el model STSB.

Per altra banda, disseccionant un poc els resultats obtinguts per model STSB, reconeixem fàcilment com mantenen el significat semàntic de la consulta, és a dir, cerca frases del mateix tòpic amb un significat semblant. No sols això, sinó que en algun resultats trobem sinònims de paraules incloses en la consulta. Exemple:

$$Partits\ \ d'esquerres = ERC$$

En conclusió, la funció portada a terme per els SRI és apropiada i eficaç.

CAPÍTOL 7
Avaluació resultats

Una vegada comprovat el correcte funcionament i rendiment dels models, en centrarem a portar a terme una avaluació detallada dels resultats obtinguts per als models en català. Tot i que el nostre objectiu final és la creació d'un recuperador d'informació semàntic en català, compararem els 3 sistemes creats per fer una valoració subjectiva del seu rendiment.

El principal problema a l'hora d'avaluar els resultats és que no tenim cap mètode objectiu, és a dir, no hi ha cap mena de referència on estiguen les solucions correctes per a una consulta. Aquesta avaluació serà subjectiva basada en el nostre judici sobre les consultes i els resultats obtinguts, en altres paraules, nosaltres determinarem quin ha sigut el grau de satisfacció en els resultats que hem obtingut. No obstant això, cal tindre en compte que hi ha un mètode de rànquing que llista els articles de més importants a menys, però pot ser que realment un article menys important siga més significatiu que altre més dalt en el rànquing.

Partim del coneixement del fet que un model Booleà retorne els articles que contenen les paraules cercades amb la consulta, hem fet l'hipòtesis com a punt de partida de l'avaluació que els resultats obtinguts pel model Booleà seran correcte i observarem quin és el nombre d'articles que se superposen amb la resta de models. Per als nostres models obtindrem alguns articles que comprenen les paraules cercades i d'altres que no, aleshores analitzarem detalladament els articles i entendrem el perquè de la selecció.

Des d'un principi podem deduir que molts dels articles obtinguts no se superposaran, ja que, els models de representació vectorial un entén el context i la semàntica i l'altre sols la semàntica, per tant, obtindrem molts articles que alguna de les seues frases tindrà més similitud encara que no estiguen totes les paraules de la consulta. I un altre problema que podem esperar és que les paraules cercades són en tot l'article i no sols en una frase.

En conclusió, que el mètode per avaluar no és totalment confiable, però l'utilitzarem com a punt de partida, ja que farem una valoració subjectiva de tots els resultats.

7.1 Consultes

Per poder fer la comparativa dels models i després d'analitzar un gran nombre d'articles en el corpus, hem extret les següents consultes que faran referència als principals tòpics presents al corpus (Política i economia).

- El president va ser expulsat del congrés.
- Només 100 vots a favor de la nova llei.
- L'equip local va guanyar per golejada.
- Nova llei aprovada al congrés.
- L'economia creix en aquest darrer període.
- votació d'abril del 2019.
- El jutge decreta error en la sentència.
- Els diputats votaran aquest dijous.
- Espanya es prepara per a una crisi.
- El partit d'esquerres guanyarà les futures eleccions.
- La mesa electoral va cometre dos errors.
- Els independentistes no accepten el tracte oferit pel Govern.
- El deute econòmic creix.

7.2 Avaluació

Consistira en una avaluació individual dels resultats obtinguts i amb una posterior comparació dels models creats. Com hem explicat abans el Model Booleà serà el punt de partida.

7.2.1. Resultats 1ª consulta

La consulta realitzada és la següent: *El president va ser expulsat del congrés*

Amb un caràcter polític i tres paraules claus (president, expulsat, congrés) la cerca va donar els següents resultats:

Model Booleà:			
45301	3332	9519	30680
39134	25586	64352	7396
77377	4851	66066	23885
7531	68450	23505	64707
45305	21852	5320	43534
6726	41421	60236	19328
28164	42720	19841	5018
6948	70129	9488	13275
65107	24629	47110	60469
66736	23359	22453	47677
68434	40735	10274	60889
17715	24101	44548	3298
40729	13254	2138	4604
Nombre d'articles:		52	

(a) Model Booleà

Model Word2Vec	Model STSB
2010	12625
48768	21708
28429	17940
32358	21707
64848	19838
64000	8261
17720	19838
40632	37228
63922	11181
9511	19565

(b) Rànqing Model Word2vec i STSB

Taula 7.1: Taules dels id dels articles resultats de la consulta "El president va ser expulsat del congrés"

7.2.1.1. Model STSB

La taula B.18 en l'apèndix mostra els resultats obtinguts. A continuació avaluarem els resultats:

1. La similitud és causada per la frase:

   ", relata la presidenta del Congrés".

   Com es pot observar és una similitud molt propera a dues de les paraules claus que hem cercat, i per això obtenim una distància prou xicoteta. Aquest resultat manté el significat semàntic (*president* $\approx$ *presidenta*) d'algunes paraules i d'altres fa *matching* perfecte (*Congres* $=$ *Congres*).

2. Frase similar:

   "El congrés del PDC va ser un èxit

   En aquest article podem observar com està relacionat en el congrés, fent un *matching* perfecte en una paraules. Mantenint el tòpic cercat, però amb una distància major que l'anterior, ja que, el seu significat és menys similar.

3. Frase similar:

   "El testimoni més emotiu va ser el del fill del president, Josep Tarradellas i Macià."

   Aquest similitud és causada pel fet de parlar del fill del president, fent coincidir la paraula president, encara i tot no manté una similitud molt gran amb la consulta, per això la distància és major.

4. Frase similar:

   "El president del Congrés es tria nominalment i en secret."

   Tenim una similitud prou gran en un troç de la frase, "El president del Congrés". És un cas un poc estrany, ja que hauria de ser superior a l'anterior frase.

5. Frase similar:

   "El Congrés es va congelar"

   En aquest cas estem parlant d'un *matching* entre la paraula Congrés. Aquest frase té una similitud menor que les anteriors, fent així que ocupe el 5é lloc, tot i que manté el tòpic.

6. Frase similar:

   "El va acompanyar, en representació de l'Estat, la presidenta del Congrés, l'exministra del PP Ana Pastor."

   Trobem un dels primers casos que hauria d'estar més amunt en el rànquing, ja que estem mantenint un significat semàntic i contextual molt similar a la consulta. Ja que, fem *matching* de congrés i presidenta, però en aquest cas la unitat semàntica és la frase i aquest fet provoca que el pes d'aquestes paraules siga menor en el conjunt.

7. Frase similar:

   Un minut de silenci ordenat per la presidenta del Congrés, Ana Pastor, va provocar la polèmica

   Un cas molt similar a mencionant anteriorment, una part de la frase té major similitud en la consulta, però el pes complet de la frase és menor que les altres.

8. Frase similar:

   El desembre del 2013 va ser el portaveu dels republicans al Congrés, Alfred Bosch, qui es va entrevistar amb Picardo durant la visita a la zona.

   Com es pot observar, a mesura que baixem en el rànquig les similituds van desapareixent, en aquest cas sols tenim semànticament (*portaveu* $\approx$ *president*) i Congrés.

9. Frase similar:

   El president del Congrés, o del Parlament, és l'única autoritat que ordena com es desenvolupa l'acte.

   Una frase prou similar a la consulta feta, ja que, fa *matching* amb *El president* i *el Congrés*. sols que la resta fa llevar-li pes.

10. Frase similar:

    El portaveu d' ERC al Congrés, Gabriel Rufián, troba "dramàticament normal la decisió del Tribunal Constitucional de suspendre cautelarment el pla per al referèndum.

    Últim article del rànquing on es pot comprovar com la distància augmenta en comparació als primers llocs, ja que sols podem observar un *matching* (Congrés).

Articles superposats:

Trobem 2 casos d'articles que se superposen en els articles retornats pel Model Booleà:

Rànquing	**Descripció Resultat**
1	Valor únic: 45305 Summary: El portaveu republicà a la cambra baixa ha seguit la protesta del seu company de files Joan Tardà davant la interlocutòria que obliga a fer classes en castellà a l'escola catalana només que els pares d'un alumne ho demanin. Línia similar: L'expulsió de Joan Tardà de la tribuna del Congrés ha estat només el primer capítol. Valor de la distància entre les frases: 10.533172
2	Valor únic: 77377 Summary: Hu Jintao també ha advertit que "els dirigents del partit no han d'abusar del seu poderï que el partit castigarà "severament"els casos de corrupció independentment del càrrec polític. Línia similar: El president Hu Jintao pronunciant el discurs el primer dia del congrés. Valor de la distància entre les frases: 10.657312

Taula 7.2: Articles superposats Consulta 1º

1. El primer article el podem trobar a la posició **35** del rànquing de resultats, ja que fa *matching* entre $expulsio \approx va\ ser\ expulsat$ i $congres = congres$. Però el conjunt total de la frase és menys similar a les que hi ha en el top 10 segons l'algoritme. Però, al meu judici aquest frase té una similitud major a la cercada.

2. El segon article podem torbar president i congrés en la frase, però en general és menys similar a les anteriors. Es troba en la posició **48**

7.2.1.2. Model Word2Vec

Donat que no hi ha cap article superposat avaluarem individualment els articles resultats de la taula B.19 en l'apèndix.

1. Frase similar:

   López va ser nomenat membre del Tribunal Constitucional el juny del 2013.

   Una valoració contextual de la frase podem dir que segueix el tòpic per es troba molt lluny de la consulta realitzada, encara que la distància diga el contrari.

2. Frase similar:

   El final del govern de José María Aznar va ser dramàtic

   En aquest cas si trobem un aspecte contextual més encertat que l'anterior, però encara així prou allunyat de la cerca realitzada.

3. Frase similar:

   El primer va ser l'octubre del 1985

   Cap mena de relació contextual, un resultat prou deficient. Pot vores beneficiat pel cas de què el summary sí que té similitud en la frase cercada.

4. Frase similar:

   El jutge va exercir de protagonista del debat constituent

   Altre cas d'un resultat incorrecte, ja que no té cap correlació amb la consulta.

5. Frase similar:

   El mutisme del gegant asiàtic va ser total

   El context de la consulta no es manté en aquest resultat.

6. Frase similar:

   Bolton va ser triat líder del UKIP el setembre passa

   En aquest cas podem deduir que entre líder i president hi ha un *matching* de context. Però la resta no trobem relació de similitud.

7. Frase similar:

   El 2015 va dimitir arran del divorci de CiU

   Analitzant la frase podem trobar un context similar entre dimitir i ser expulsat.

8. Frase similar:

   El comiat de Gadea es va fer el 17 novembre del 2009, un cop Parlon va prendre possessió del càrrec

   Com en el cas anterior, la similitud del context és causada per comiat i ser expulsat, però encar així i analitzant el casos de les frases stsb, trobem que és un resultat de baixa similitud.

9. Frase similar:

   El Parlament actual va ser reconstruït després del foc que el va destruir l'any 1834

   Podem trobar un context similar entre les paraules *Parlament* $\approx$ *congrés*.

10. Frase similar:

    El desembre del 2011 va ser nomenat cap de gabinet del president del govern espanyol, Mariano Rajoy

    Trobem un *matching* en la paraula *president*.

7.2.1.3. Conclusions

Analitzant els diferents resultats hem pogut observar com en el model STSB els significats semàntics i contextual de la consulta es recuperaven a cada resultat, variant entre ells un poc. Produint resultats adequats i en un alt grau de satisfacció. Un rànquing que en alguns moments falla, no obstant és prou correcte en termes de similitud.

Per altra banda, els resultats obtinguts pel model Word2Vec no han sigut del tot correctes, observem com el tòpic tractat es perd en alguns resultats, tot causat pel fet que el model Word2Vec no és contextual.

7.2.2. Resultats 2° consulta

La consulta realitzada és: *Només 100 vots a favor de la nova llei*

Una consulta centralitzada en un tòpic legal, que intenta abarcar molt aspectes d'aquest. Paraules claus "vots a favor nova llei"

La següent taula mostra els resultats obtinguts amb les id dels diferents articles extrets.

Model Booleà:	
71182	60863
45781	68460
94287	77168
63929	71691
1760	19452
96019	75636
66497	77025
25127	629
8608	93516
65212	25556
12511	7078
58708	10260
23724	
Nombre d'articles:	25

(a) Model Booleà

Model Word2Vec	Model STSB
17.916	32136
85444	28403
53862	44899
66385	11927
13668	16290
88175	28402
76632	2892
84879	69205
58646	58232
15103	77385

(b) Rànquing model Word2vec i STSB

Taula 7.3: Taules dels id dels articles resultats de la consulta "Només 100 vots a favor de la nova llei"

7.2.2.1. Model STSB

A continuació portarem a terme l'avaluació individual dels articles resultats de la taula B.20 seguint el rànquing.

1. Frase similar:

   Només si aquestes esmenes prosperen acabarà votant a favor de la proposició de llei.

   Una frase amb una similitud prou alta a la consulta, ja que trobem diverses paraules repetides i mes enllà una similitud semàntica molt interessant com *proposici* $\approx$ *nova*. En resum, un resultat molt correcte.

2. Frase similar:

El 'sí' –la proposta de Duran– va guanyar per 95 vots.

Una similitud semàntica trobada en *En sí* $\approx$ *a favor* i d'altres paraules que fan *matching* com vots. També ha relacionat 95 amb 100. En conclusió, una similitud altra i correcta.

3. Frase similar:

   Al final només C's hi va votar a favor.

   En aquest cas trobem una similitud en *votar a favor* i *vots a favor*, i en la paraula només. També, manté el tòpic i significat semàntic la frase.

4. Frase similar:

   Aquesta vegada només seria necessària la majoria simple dels vots: més a favor que en contra

   En aquest cas podem torbar diverses similituds (només, a favor, vots) convertint-la en una frase prou correcta donada per la seua semblança. Pot ser, que haja fet la relació de 100 vots amb majoria simple, però seria molt rara.

5. Frase similar:

   A més, ella mateixa –recorden– va destacar quan va presentar la llei del vot electrònic que era la que havia de servir per resoldre les traves que el Govern posava als catalans que volien votar des de l'estranger.

   Per una part trobem una relació contextual i semàntica entre *va presentar* i *nova*, i per altra algunes paraules iguals en les dues frases (llei, vot, votar).

6. Frase similar:

   El 'sí' –la proposta de Duran– va guanyar per 95 vots.

   Mateixa frase en diferent article [2], pot ser que a l'hora de fer la transformació vectorial alguna característica variara i per això la diferent distància. Però, seguix sent la mateixa frase i hauria de donar el mateix resultat.

7. Frase similar:

   la declaració de la lletrada de justícia... des de la seva fugida pel terrat fins a sentir Carme Forcadell per megafonia quan no va parlar.

   Un resultat poc adequat a la consulta donada. Ja que, no es troba cap similitud entre les dues frases.

8. Frase similar:

   Per aconseguir-ho, cal que superi el 50% dels vots, un llindar que ultrapassa en la mitjana de tots els sondejos.

   Una frase en una distància superior, ja que com es pot observar sols es manté *vot* com a paraula cercada, però estem parlant del mateix tòpic i context, per tant, és correcte.

9. Frase similar:

   no només aquestes, també els canvis que es vulguin operar a la Corporació Catalana de Mitjans Audiovisuals o la llei electoral, que ara els dos partits (que reuneixen els 90 diputats necessaris) podrien impulsar.

   Trobem vàries similituds exactes (*noms*, *llei*) i una similitud contextual com $impulsar \approx nova$, mantenint en tot moment el tòpic i el significat cercat. En resum, un resultat notablement bo.

10. Frase similar:

    Amb el 86% escrutat, el 47,80% dels vots han estat per al líder del PPD, només 0,61 punts i 9.943 vots més que el governador Fortuño.

    Parlem d'una distància superior trobem que el resultat s'allunya del buscat, ja que no parlem de lleis sinó d'una votació per a presidents. Encara així troba similitud amb *vots*. Aquest resultat ho avaluem con deficient, ja que no retorna el context esperat.

7.2.2.2. Model Word2Vec:

Resultats avaluats extrets de la taula B.21.

1. Frase similar:

   Precisament, aquest dimecres la CUP donarà suport a través de dos vots favorables a la llei de pressupostos.

   Trobem un context molt similar al fet en la consulta, on existeixen dues paraules iguals (vots, llei) i una paraula contextual com $favorables \approx afavor$. Per tant, un resultat prou favorable.

2. Frase similar:

   En efecte, a l'article 45 de la LEC ja s'estableix la possibilitat d'integrar centres a la xarxa de la Generalitat per mitjà d'una llei

   Un resultat pitjor que l'anterior, ja que sols coincideix *llei* com a paraula, i pel context tampoc se sembla. Aquest resultat deixa un poc entreveure les debilitats del model.

3. Frase similar:

   El govern espanyol deu a la Generalitat els 750 milions d'euros corresponents a la disposició addicional tercera de l'Estatut.

   Un resultat deficient, ja que no hi ha ni *matching* entre paraules ni pel context trobem una relació directa.

4. Frase similar:

   a setmana passada, Jeremy Corbyn va ordenar als seus diputats de votar a favor de la llei de l'article 50.

   En aquest cas, la frase presenta varies relacion directes (llei, votar a favor) i un context molt semblant a la consulta. Per tant, hauria d'estar mes amunt en el rànquing.

5. Frase similar:

   La CUP no vol donar normalitat democràtica a les eleccions convocades per Mariano Rajoy el 21 de desembre a través de l'article 155 de la Constitució.

   Deduïm que el fet d'aparèixer números semblants als de la consulta ha produït que el resultat obtingut siga poc similar a la consulta, encar així manté el tòpic cercat.

6. Frase similar:

   a normativa ve acompanyada també d'una nova governança per a la Unió Energètica que també s'ha aprovat amb 475 vots a favor i 100 en contra.

   A contrari del resultat anterior, hem obtingut un resultat que conté diverses relacion directe (vots, 100, nova, a favor), opinant que aquest resultat té major similitud que molts dels anteriors.

7. Frase similar:

   Grillo havia mobilitzat la població italiana per a la recollida de firmes per presentar una llei d'iniciativa popular amb l'objectiu d'impedir als condemnats de corrupció l'accés a la política.

   Analitzant, podem observar un context similar a la consulta, obtenint un *matching* com *llei*. A més, podem intuir que el model ha fet la semblança entre *presentar* $\approx$ *nova*

8. Frase similar:

   Enviar una tona de residus a aquesta instal·lació costa uns 41 euros decàn on municipal a l'Ajuntament, l'any que ve la xifra s'encarirà a 47 euros i el 2025 ja valdrà 77 euros la tona.

   Resultat incorrecte, ja que no pertany ni al tòpic cercat.

9. Frase similar

   La petició l'ha fet aquest dijous a Edimburg, hores abans que la cambra autonòmica aprovés per majoria de 68 a 54 vots el tercer i últim estadi de la llei genèrica de referèndums d'Escòcia, a partir de la qual s'hauria de desenvolupar l'específica del segon plebiscit.

   Un resultat prou correcte, on es manté el tòpic cercat amb un context similar, ja que parlem de lleis i vots.

10. Frase similar

    Divendres a mitjanit comença formalment la campanya electoral de l'1-O

    Resultat dolent, on no és mantó cap tòpic, ni conté cap similitud. Es pot extraure que l'article parlara de votacions, però la frase en si cap semblança.

Articles superposats:

Rànquing	**Descripció Resultat**
1	Valor únic: 25556 Summary: L'hora i mitja de la intervenció del president en funcions, transcrita. I en el núvol de 'tags', podeu veure les paraules que més ha utilitzat. Línia similar: En la línia de continuar augmentant l'èxit escolar a Catalunya com ha vingut succeint en aquests darrers anys, es seguirà desplegant la Llei d'Educació de Catalunya, fruit d'un ampli consens assolit en el seu dia en el nostre Parlament. Valor de la distància entre les frases: 10,745329

Taula 7.4: Articles superposats

Aquest article es troba en la posició **23** i va se superposa amb els articles retornats pel model Booleà, com podem comprovar conte la relació directa llei.

7.2.2.3. Conclusions

En aquests nous resultats podem observar la tendència del model STSB a treure resultats que tinguen una relació semàntica i contextual amb la consulta amb un grau de satisfacció alt. I retornant un rànquing prou acorde als objectius.

Per altra banda, el model Word2Vec ha funcionat millor obtenint resultats de similar context, però encar així sol fallar en alguns resultats tornats, ja que o el pes del número és molt gran o el context tornat no és l'adequat.

7.2.3. Resultats 3º Consulta

La consulta és: *L'equip local va guanyar per golejada*

El tòpic seleccionat en aquest cas són els esports, ja que volem comprovar que retornen els diferents models amb una consulta fora dels paràmetres del corpus.

En la següent taula es mostra el valor retornats per els models.

Model Booleà:	
Nombre d'articles:	0

(a) Model Booleà

Model Word2Vec	Model STSB
66562	25026
6881	3357
16533	141
18430	34144
78712	15501
9831	23592
60972	50487
26176	57499
11759	67380
5313	32638

(b) Model Word2vec i STSB

Taula 7.5: Taules dels id dels articles resultats de la consulta L'L'equip local va guanyar per golejada"

Articles superposats: 0

7.2.3.1. Model STSB

Resultats analitzats de la taula B.22.

1. frase similar:

   Vas guanyar per golejada

   Una similitud quasi perfecta, només faltaria fer regència a l'equip local, per tot l'altre manté semàntica i context.

2. frase similar:

   La Trinca el guanya per golejada

   Un resultat amb un grau de similitud molt alt. Ja que trobem diversos elements amb una relació directa (per golejada) i semàntica ($vaguanyar = elguanya$)

3. frase similar:

   Penso que vam guanyar per golejada.

   Si llegim el summary trobem que parlem defenses jurídiques, però en la frase no tenim mes context i per així del resultat. En general, el grau de similitud és gran.

4. frase similar:

   Malauradament, al president del Govern espanyol li falta decisió política, i al de la Generalitat li sobra gosadia i agressivitat.

   Deduïm que ha relacionat *gosadia i agressivitat* amb esport i per això el retorn d'aquest article. Però no se sembla a la consulta feta.

5. frase similar:

   Gairebé ningú en l'esfera política gosa fer-ho

   Resultat prou deficient, si no és pel fet de relacionar esfera en esport.

6. frase similar:

   Pensava que, en plenes negociacions per formar govern, ningú gosaria fer-li ombra

   Les distancia de similitud cada vegada són més grans, és causat pel fet que ja no queden cap resultat esportiu.

7. frase similar:

   Sandalio Gómez

   El model no troba cap resultat esportiu més, i comença a retornar els resultats mes propers, però que no han de veure. És a dir, és un resultat incorrecte.

8. frase similar:

   Ricard Gomà afronta les properes eleccions de maig per primer cop com a cap de llista

   Igual que els exemples de dalt.

9. frase similar:

   I que, de fet, el va ajudar a guanyar les eleccions per golejada.

   Aquest resultat encara que no mantinga el tòpic cercat, hauria d'estar molt mes a dalt, donat que fa relacion directes amb paraules (*guanyar*, *golejada*).

10. frase similar:

    Gomà creu que per "primera vegada" és possible protagonitzar un "triomf social i polític" des de valors de "revolució democràtica"

    En aquest cas trobem un relació de sinonímia amb $triomf \approx guanyar$.

7.2.3.2. Model Word2Vec

Taula dels resultats presentada en B.23. En aquest cas no mostrarem un per un els anàlisis, ja que molts d'ells comparteixen conclusió.

En general, en ser un tòpic poc indexat a l'índex, trobem que els model Word2Vec no ha sigut capaç de trobar resultats favorables a l'hora de fer la consulta. Molts d'ells parlen de política i no tenen similituds de cap mena. En resum, els resultats han sigut nefastos per aquesta consulta, comencem a teoritzar que el corpus usat per SpaCy per a entrenar el model no contenia notícies d'esports o temes menys importants.

7.2.3.3. Conclusions

Per una part, el model STSB ha fet un gran treball retornant les poques notícies que parlàvem sobre esports i amb una similitud molt alta. A més, moltes de les posteriors notícies retornades tenien o alguna similitud directa o algun sinònim. Encara que no mantenien el tòpic de la consulta.

Per altra banda, la conclusió del model Word2Vec és que els resultats han sigut incorrectes.

7.2.4. Resultats 4º Consulta

La consulta realitzada és: **Consulta 4:** Nova llei aprovada al congrés

Model Booleà:	
Nombre d'articles:	0

(a) Model Booleà

Model Word2Vec	Model STSB
34878	37843
34878	28764
9614	32856
15538	1337
61233	27057
17975	28764
51937	37845
60413	47379
7228	21766
10583	38951

(b) Model Word2vec i STSB

Taula 7.6: Taules dels id dels articles resultats de la consulta "Nova llei aprovada al congrés"

7.2.4.1. Model STSB

Resultats mostrats en la taula B.24. En aquest cas trobem un tòpic paregut a la consulta 2, parlant de votacions i de lleis.

1. frase similar:

   El Congrés ha aprovat per una aclaparadora majoria la llei orgànica que contempla l'abdicació del rei Joan Carles

   Es pot observar una clara similitud entre la consulta i la frase, ja que els dos parlen d'aprovar un llei en el congrés.

2. frase similar:

   Necessitem cap de llista al Congrés per Barcelona.

   Baix grau de similitud, sols tenim com a *matching* la paraula Congrés, però el tòpic no es manté

3. frase similar:

   El Congrés va aprovar al novembre una llei orgànica per incloure la norma comunitària en la legislació espanyola.

   Igual que el número 1, mateixes conclusions. Un resultat molt favorable

4. frase similar:

   a Llei de Seguretat Nacional va ser aprovada pel Congrés el 2015

   Es manté el tòpic i les similituds, mateixes conclusions que les anteriors notícies [1]-[3].

5. frase similar:

   Amb tot, la resolució del govern espanyol va ser aprovada amb una majoria més que àmplia al Congrés

   Un resultat molt correcte, ja que manté el significat de la consulta, i conté paraules iguals (*congrs*, *aprovada*).

6. frase similar:

   Necessitem cap de llista al Congrés per BCN

   Igual que el resultat [2]. És la mateixa notícia.

7. frase similar:

   Duran i Lleida avui al Congrés / EF Duran i Lleida avui al Congrés / EF.

   Resultat deficient, poc grau de similitud en la consulta. sols obtenim Congrés com a paraula similar. Encara que i tot el summary de la notícia ens done un grau de similitud alt, però la frase que ha fet *matching* no és correcta.

8. frase similar:

   Una llei elaborada al Congrés s'ha de complir encara que no agradi", ha afegit Torres-Dulce

   Resultat molt favorable, hauria d'estar mes dalt en el rànquing, ja que retorna just la cerca realitzada. Fent uns sinonímia entre *elaborada - aprovada*

9. frase similar:

   Per al segon intentarà sobreviure per aprovar pressupostos i lleis amb C's, CDC, el PNB i els canaris, que completen la majoria de centredreta al Congrés

   Les distàncies comencen a ser mes grans i el grau de similitud menor, en aquest resultat el podem observar, ja que sols tenen relació d'igualtat les paraules com *Congrés, lleis*.

10. frase similar:

    Duran i Lleida, al Congrés / EF Duran i Lleida, al Congrés / EF

    Igual que el resultat [7], no és bon *matching*. I en aquest cas la notícia no té molta similitud en la cerca.

7.2.4.2. Model Word2Vec

1. frase similar:

   Entrevista al cònsol general d'Espanya a Perpinyà publicada al diarid 'Alacant 'Información Entrevista al cònsol general d'Espanya a Perpinyà publicada al diari d'Alacant 'Información

   Mateixa conclusió per als resultats [1]-[2], ja que repetia la mateixa frase dues voltes. Encara i tot és un resultat equívoc, ja que el grau de similitud és molt baix, no conté car mena de relació.

2. frase similar:

   Retreuen així al PP la impugnació de l'Estatut de l'any 2006 al Constitucional

   En aquest cas trobem una relació entre *Constitucional* i *Congrs*, però llevat d'això, altre resultat poc similar.

3. frase similar:

   Nova baixa al grup socialista al Congrés per discrepàncies amb Pedro Sánchez.

   Continuem amb els resultats incorrectes, no trobem cap relació de similitud entra la frase i la consulta

4. frase similar:

   La nova llei es debatrà al febrer a l'Assemblea Nacional

   Un resultat mes adequat a la consulta. Ja que, tracten del mateix tòpic i trobem paraules similars.

5. frase similar:

   La segona llei contra els desnonaments, aprovada per unanimitat al mes de desembre al Parlament, podria acabar al Tribunal Constitucional (TC).

   Un resultat amb un grau de similitud gran, ja que estem parlant de lleis que han estat aprovades i trobem prou similituds.

6. frase similar:

   Alícia Sánchez-Camacho al Parlament de Cataluny Alícia Sánchez-Camacho al Parlament de Cataluny.

   D'aquest resultat sols podem trobar la relació entre $Parlament - Congrs$.

7. frase similar:

   El Parlament egipci aprova una reforma que podria mantenir Al-Sissi al poder fins al 203

   En aquest cas, la frase té prous similituds amb la consulta realitzada com $reforma - llei$ o $Parlament - Congrs$.

8. frase similar:

   Comença el compte enrere per al judici oral contra l'independentisme al Suprem, previst per al gener.

   No es manté el tòpic cercat, ni trobem cap similitud. Un resultat deficient.

9. frase similar:

   L'advocat de Jordi Sànchez, Jordi Pina, ha registrat aquest dimarts almatí al Tribunal Suprem una nova petició al jutge Pablo Llarena perquè el seu client pugui assistir al ple d'investidura de divendres a les 10 h al Parlament.

   Cap mena de similitud, mal resultat.

7.2.4.3. Conclusions

Per una part, el model STSB ens ha proporcionat resultat acordes a la consulta feta, demostrant la seua eficàcia i rendiment, no sols mantenint en tot moment el tòpic i el significat semàntic, sinó que ha proporcionat algun sinònim. Encara que hi ha alguns casos on ha fallat a l'hora de proporcionar-nos un resultat correcte.

Per altra banda, el grau de satisfacció en el model Word2Vec és baix, donat que hi han molts resultats que no es correlacionen amb la consulta. Tot i que tenim alguns exemples on el resultat és adequat i similar, trobant contextos similars.

7.2.5. Resultats 5º Consulta

La consulta feta es la següent: *L'economia creix en aquest darrer període*

On ens centrem en un tòpic econòmic i parlem d'un passat pròxim.

La següent taula mostra els resultats obtinguts dels diferents models.

Model Booleà:	
50955	42732
76475	54054
52993	73096
94078	24142
87587	60485
26204	6077
91655	
83372	
431	
43816	
7526	
63897	
98172	
Nombre d'articles:	19

(a) Model Booleà

Model Word2Vec	Model STSB
73747	6525
8402	75528
83565	51997
2577	18791
82799	87513
64672	25436
18184	69103
22773	51526
69723	19498
74147	36698

(b) Model Word2Vec i STSB

Taula 7.7: Taules dels id dels articles resultats de la consulta L·L'economia creix en aquest darrer període"

7.2.5.1. Model STSB

Com amb anterioritat els resultats estan extrets de la taula B.26.

1. frase similar:

   L'economia està creixent

   Resultat molt similar a la consulta, totes les paraules que componen la frase estan relacionades directament.

2. frase similar:

   Una economia en crisi.

   Podem deduir que el pes de trobar una economia ha guanyat sobre la contrarietat entre crisi i creixent. Per tant, trobem un article que parla del mateix tòpic i assenyala parts paregudes, però amb un antònim clar *creixent* $\neq$ *crisi*.

3. frase similar:

   Respecte a les perspectives de creixement per al 2012, el titular d'Economia reconeix que "aquest any serà difícil.

   Un altre resultat molt acorde a la consulta feta, on podem trobar moltes paraules igual i un context similar.

4. frase similar:

   Un procés que minva el creixement econòmic.

   En aquest cas el resultat té un grau de similitud molt alt, ja que ha reconegut *creixement econòmic* com *l'economia aquesta creixent*.

5. frase similar:

   ¿Aquest és el tipus de persones que «vetllen» per la naturalesa i les que «garanteixen» un equilibri de l'ecosistema?

   Resultat res convenient, fora de lloc, no té cap mena de relació de similitud.

6. frase similar:

   Crec honestament que a Espanya li interessa això per raons econòmiques i democràtiques

   Un resultat de context similar on hi ha paraules similars.

7. frase similar:

   L'alentiment de l'economia té a favor la demografia

   En aquest cas tenim un resultat favorable, on el context és similar i podem fer relacion d'antonímia com $alentimentdel'economia \neq l'economiaestcreixent$.

8. frase similar:

   Creu que l'any que ve l'economia espanyola començarà a créixer

   Observem una clara similitud entre les dues frases (*economia* = *economia*) (*començarà a créixer* $\approx$ *està creixent*).

9. frase similar:

   El grup socialista presentarà una iniciativa al Congrés perquè aquesta norma recuperi la velocitat de creuer ara que l'economia çreix"

   Altre resultat adequat, on trobem prou similituds (*economia creix*). A causa del fet que la frase és llarga el pes causat per aquestes paraules és menor, encara que estiga molt prop a la consulta.

10. frase similar:

    més, la quarta preocupació dels ciutadans, per darrere de l'atur, la corrupció i l'economia, ja són "els polítics, els partits i la política"

    Resultat allunyat de l'objectiu, ja que l'únic que podem trobar com a similar és *economia*.

7.2.5.2. Model Word2Vec

1. Posició 1,2,3,4 trobem articles que no tenen cap mena de relació de similitud amb la consulta feta. Són resultats incorrectes.

2. Posició 5 - frase similar:

   Primer, es redueix el finançament general a tot el sistema universitari en 1.300 milions d'euros en tres anys, aplicant-se la retallada més forta aquest curs 2010 – 2011"

   Trobem un similitud en els contextos, ja que els dos estan parlant d'economia.

3. Posició 6 - frase similar:

   s en aquesta classe mitjana i en el seu creixement exponencial -l'any 2000 només n'eren 5 milions-, en qui confia el govern xinès per transformar l'economia cap a un model apuntalat en el consum intern

   Resultat poc acorde, ja que està parlant de creixement demogràfic i no de creixement econòmic. Però si parla d'algun creixement i menciona l'economia.

4. Posicions 7,9,10 són resultats incorrectes, ja que no mantenen cap similitud contextual.

5. Posició 8 - frase similar:

   El Govern, però, reivindica el model de flux monetari, sobretot "en èpoques de crisi econòmica i taxes d'atur elevades", quan "pren molta més rellevància" l'impacte de "l'activitat de l'administració central en un territori"

Un context molt mes adequat, en el que parlem d'economia i creixement. Hauria d'estar mes a dalt en el rànquing.

7.2.5.3. Conclusions

Per una banda, el model STSB, com ens porta demostrant, ha obtingut resultats acorde a la consulta feta, retornant documents que parlem de creixement econòmic, aportant-nos tant antònims com sinònims.

Per altra banda, sols alguns resultats del model Word2Vec són adequats. Molts d'ells no mantenen el tòpic de la consulta.

7.2.6. Resultats 6ª Consulta

En aquest cas la consulta feta és: *votació d'abril del 2019*

Tornat així a un aspecte polític, però afegint dates. En les següents taules es mostren els valors obtinguts.

Model Booleà:	
66575	4146
92016	1933
3545	92111
90650	1497
1416	8481
7823	3002
7833	92678
2275	
61387	
88443	
60579	
88560	
4480	
Nombre d'articles:	20

(a) Model Booleà

Model Word2Vec	Model STSB
83718	5020
9503	71261
15150	70515
2448	15926
17240	63133
89824	26942
8238	74354
13640	71261
60057	65890
8425	65890

(b) Model Word2Vec i STSB

Taula 7.8: Taules dels id dels articles resultats de la consulta "votació d'abril del 2019"

7.2.6.1. Model STSB

Els resultats són analitzats des de la taula B.28.

Analitzant els resultats individualment es pot observar un clar patró de resultat, en ser 66% de la consulta *abril del 2019* els resultats són tots *abril del ....* Aleshores podem concloure que la consulta devia haver sigut mes especifica.

Exemples: *Abril del 2018, abril del 2014, abril del 2016, etc..*

En altres paraules, la característica principal extreta de la consulta és *abril del 2019*, per tant, els resultats mes propers a la consulta són aquells que tinguen *abril del ...* entre les seues frases.

7.2.6.2. Model Word2Vec

Trobem un agran similitud amb el model STSB, on la majoria dels resultats tornats frases amb dates, en aquest cas menys específic perquè no té en compte la semàntica de la frases, sols el context, aleshores el model no pot distingir entre dates i creu que totes són iguals. Exemples: *gener del 2018, juny del 2018, etc...*

7.2.6.3. Conclusions

Trobem una diferència clara entre model STSB i Word2Vec, com un té en compte les diferències semàntiques i una data és diferent d'altra, ja que els mesos tenen significat, i l'altre model entén totes les dates per iguals donat que comparteixen context.

7.2.7. Resultats 9ª Consulta

Consulta 9: Espanya es prepara per a una crisi

Model Booleà:			
46925	60862	62738	
49982	33073	77669	
13750	12150	69632	
51405	15884	25415	
46833	30071	44775	
44412	3171	99959	
71057	20261	33072	
53729	9501	1932	
36793	20703	62910	
44444	69984	10260	
69924	20700		
34364	75101		
66127	46495		
Nombre d'articles:		36	

(a) Model Booleà

Model Word2Vec	Model STSB
50713	49412
8869	14811
69794	4381
21027	48399
59518	58629
61903	53076
86606	78605
16529	55590
16531	70293
88449	31255

(b) Model Word2Vec i STSB

Taula 7.9: Taules dels id dels articles resultats de la consulta "Espanya es prepara per a una crisi"

7.2.7.1. Model STSB

1. frase similar:

   La crisi econòmica

   Un resultat que torna amb molta precisió el cercat. Encara que no hem buscat quin tipus de crisi, moltes voltes quan ens referim a la crisi fem referència a la crisi econòmica, i aquesta consulta ha sigut capaç de tornar-ho.

2. frase similar:

   diu que "Espanya entra en una crisi d'Estat

   Aquest resultat devia haver sigut el número 1, ja que, el grau de similitud entre les frases és d'1. És a dir, la frase té una semblança enorme en la consulta realitzada.

3. frase similar:

Pel que fa a la crisi interna, Iglesias ho ha atribuït a l'adolescència de la formació"

Aquest resultat sols trobem el fet que estem parlant d'una crisi. Entre altres podríem fer la relació *Iglesias* amb *Espanya* però seria assumir molt.

4. frase similar:

   Hi ha una alternativa social a aquesta crisi.

   Fa una relació amb la paraula *crisi*. I el fet *alternativa* té relació amb *preparaci*, encara que molt llunyana.

5. frase similar:

   La crisi política

   Torna un article que parla de la crisi política, per tant, el tòpic es manté, ja que parlem d'una crisi.

6. frase similar:

   Diplomàcia en temps de crisi

   Molt paregut als resultats anteriors, on trobem un *matching* amb la paraula crisi.

7. frase similar:

   La Croix': Crisi de poder al vaticà i "Espanya desestabilitzada per la crisi del sector bancari

   Podem observar una clara relació amb la paraula *crisi* i la frase extreta, a més trobem com l'article parla de la crisi que sofreix Espanya. Per tant, el resultat és favorable.

8. frase similar:

   El president espanyol ha reclamat un "esforç col·lectiu per sortir de la crisi

   AL meu judici aquesta frase devia estar mes amunt en el rànquing, perquè aquesta parlant d'una preparació davant la crisi. Fent relacion semàntiques entre *reclama...* $\approx$ *preparaci* i les paraules *crisi* i *espanyol* $=$ *Espanya*

9. frase similar:

   Una crisi diplomàtica

   Resultat coix, ja que sols fa *matching* amb *crisi*.

10. frase similar:

    Hem governat en moments de crisi social, política

    Igual que el resultat anterior.

7.2.7.2. Model Word2Vec

La taula dels resultats B.35

1. frase similar:

   Espanya es trenca quan es porta a la fallida les institucions financeres valencianes, quan es condemna a la desocupació milers de famílies"

   Un resultat deficient, sols fa referència a Espanya.

2. frase similar:

   Ara per ara a les ponències que es debatran al congrés no es fa cap menció a la Crida de Puigdemont, i per això no es descarta que a última hora es puguin presentar in situ per discutir-ho

   Un resultat incorrecte, ja que no té cap mena de similitud amb la consulta.

3. frase similar:

   Actualment a Turquia es produeix una situació paradoxal

   Podem fer la relació de *Espanya* $\approx$ *Turquia* i una *situació paradoxal* amb *crisi*. Per tant, considerem un possible resultat correcte.

4. frase similar:

   Ciutadans, que com Podem es presenta per primer cop a unes autonòmiques a Euskadi, es quedaria sense representació a la cambra basca

   Cap mena de similitud amb la consulta.

5. frase similar:

   Activistes es traslladen a Lampedusa per forçar que es doni port a l'Open Arms

   Resultat desfavorable.

6. frase similar:

   tota una bufetada per a un règim que es gasta cada any una milionada per millorar la seva imatge a Occident

   Podem fer la relació que la situació que descriu és una crisi, però, tot i això, segueix sent un resultat desfavorable,

7. frase similar:

   La científica afegeix que cal una formació més humanista per a les persones que es dediquen a desenvolupar tecnologia

   Resultat desfavorable, cap similitud.

8. frase similar:

   Una opció que es torna a valorar ara

   Resultat desfavorable, cap similitud.

9. frase similar:

   Una opció que es torna a valorar ara

   Resultat desfavorable, cap similitud. És una notícia diferent de l'anterior però comparteixen frase.

10. frase similar:

    s tracta de la indemnització més gran que es concedeix a Espanya per una negligència mèdica

    Resultat desfavorable, almenys fa similitud en *Espanya*.

7.2.7.3. Conclusions

Seguint la dinàmica actual, el model STSB ha obtingut uns resultat acordes a la consulta, amb un grau de satisfacció gran, tenint en compt tant el context com la semàntica. Extraent tota classe de crisis que poden ocorre, ja que no vàrem especificar quina en la consulta.

Per altra banda, el model Word2Vec torna a presentar resultats desfavorables, que no tenen un context en comú amb la consulta presentada. Una hipòtesi és el fet que a l'hora d'entrenar aquest model s'ha usat un corpus inadequat o poc extens per fer relacions.

7.2.8. Resultats 10ª Consulta

Consulta 10: El partit d'esquerres guanyarà les futures eleccions

Model Booleà:	
48159	67136
22250	25025
640	16871
38312	12474
54564	25166
27699	625
32707	29408
25286	66649
13182	31060
75047	24481
69566	
62001	
38346	
d'articles:	23

(a) Model Booleà

Model Word2Vec	Model STSB
66202	19618
16373	33270
81053	81
32571	72157
27461	61247
2965	67389
52662	16334
13397	14877
32827	67960
25249	27639

(b) Model Word2Vec i STSB

Taula 7.10: Taules dels id dels articles resultats de la consulta "El partit d'esquerres guanyarà les futures eleccions"

7.2.8.1. Model STSB

1. frase similar:

   Els 'comuns' preparen el manifest fundacional per al que acabarà sent el futur partit de la confluència d'esquerres

   En aquest resultat podem observar dues relacions semàntiques per a una mateixa (*els comuns-confluència d'esquerres –> el partit d'esqueres*) i $futur = futures$.

2. frase similar:

   Lluny de l'acord, encara, sobre el format de les futures eleccions plebiscitàries

   Trobem una clara similitud en la frase *futures eleccions*. El resultat és menys clar que l'anterior.

3. frase similar:

   Les enquestes pronostiquen un gran creixement d'ERC, que guanyaria per primera vegada les eleccions al Parlament des de la Segona República

Un resultat molt interesant on es fa una relació de sinònima clara ($ERC \approx partit\ esquerres$) i paraules igual com *guanyar* i *eleccions*.

4. frase similar:

   El president del Parlament Europeu (PE), Martin Schulz, està convençut que no hi haurà acord sobre una possible quita del deute grec, una de les demandes clau de la coalició d'esquerres Syriza, guanyadora de les eleccions d'ahir

   Es manté el significat i el context, i podem observar paraules repetides ($gunayar$, $eleccions$, $esquerre$) i una sinonímia com $coalici \approx partit$.

5. frase similar:

   En la investidura va parlar de lluitar contra la "submissió ideològica" que encarnen les idees d'esquerres, del progressisme al marxisme

   Resultat desfavorable, ja que parla de la investidura i esquerres però no té molt a vora en eleccions

6. frase similar:

   Tot plegat, el que s'ha fet i com s'ha fet, marcarà història en el panorama dels processos de pau al món, i servirà d'important referència per a futures negociacions

   Resultat deficient, no té el mateix context i sols trobem una paraula igual $futures$, però no té cap classe de semblança amb la consulta.

7. frase similar:

   'L'esquerra en marxa

   Resultat regular, sols parla de l'esquerra, però cal tindre en compte que ha acceptat esquerra com a força política, i no el contrari de dreta. Per tant, ha fet un bon treball.

8. frase similar:

   Economista i politòleg de formació, ja va ocupar un càrrec important en el departament de Vicepresidència durant el tripartit i és el president del Consell Nacional d'Esquerra des del 2011.

   Resultat obviable, ja que no aporta molt, no té quasi semblança en la consulta.

9. frase similar:

   a desacceleració de les economies emergents que demandaven commodities ha impactat en els pressupostos, explica Andrea Costafreda, directora programàtica d'Oxfam Intermón per a l'Amèrica Llatina i el Carib.

   Resultat molt lluny de la consulta i per això, obté aquest lloc en el rànquing.

10. frase similar:

    Si guanya les eleccions, la confluència es compromet a respondre a l'emergència social; fer una auditoria del deute de la Generalitat; revertir la privatització dels serveis públics; garantir la democràcia en les relacions econòmiques, i acabar amb la corrupció, entre altres coses

    Un resultat molt correcte, però a causa de la baixada de pes en el conjunt de la frase que té *Si guanya les eleccions* pertany al 10º lloc del rànquing.

7.2.8.2. Model Word2Vec

1. frase similar:

   El Parlament Europeu investigarà l'eurodiputat polonès per les declaracions misògines

   Resultat desfavorable que no concorda amb el context cercat.

2. frase similar:

   El partit pretén que aquesta conferència nacional marqui les línies bàsiques que defensarà ERC durant el procés constituent

   Encara que no parla de res de les eleccions, trobem diverses paraules que coincideixen, com *partit*.

3. frase similar:

   El ministre sirià d'Exteriors, Walid al-Mualem, ha assegurat avui que el seu país celebrarà eleccions legislatives abans de finals d'any i que les urnes actuaran d'àrbitre de les reformes anunciades pel president Baixar al-Assad

   Cas paregut a l'anterior, té un context que parla de les eleccions però no dels partits d'esquerra.

4. frase similar:

   El programa electoral, passades les eleccions municipals.

   En aquest cas trobem un context mes similar, encara que seguim sense parlar de partits, però apareix *eleccions*.

5. frase similar:

   El caràcter plebiscitari el donen les forces polítiques", subratlla Turull

   Resultat desfavorable, ja que trobem *plotiques*, però el context cercat no apreix en cap moment.

6. frase similar:

   El nou càrrec es crearà dijous i Echenique estrenarà les noves responsabilitats en una primera reunió que tindrà lloc dissabte, quan el partit celebrarà el seu Consell Ciutadà Estatal (CCE)

   Mateix cas que el resultat anterior, trobem *partit* i ha de veure amb la política, però no estem en el context cercat.

7. frase similar:

   El diari assenyala que les sigles P.A.C.

   Resultat inadequat, cap mena de similitud.

8. frase similar:

   El partit no aprovarà finalment dissabte les seves llistes.

   Partit com a única paraula que coincideix en la consulta, però, en general, és un resultat inapropiat.

9. frase similar:

   Les formacions valoren el nou horitzó electoral anunciat per Mas el 27 de setembre d'enguany

   Aquest cas trobem que es fan les relacion de $formacions \approx partits$ i $electoral \approx eleccions$, i un context similar, per tant, podem concloure que és un resultat favorable.

10. frase similar:

    El millor antídot contra el sobiranisme és que Ciutadans guanyi les eleccions

    Una bona relació de similitud trobada en *guanyi les eleccions* i *guanyar les eleccions*, per tant, un bon resultat. Ja que, parla d'un context similar al de la consulta.

7.2.8.3. Conclusions

Per una banda, observem un alta percentatge d'encert per part del model STSB, sempre recuperant notícies similars a la consulta, mantenint un context semblant i recuperant sinònims per al significat semàntic.

Per altra banda, el model Word2Vec treu resultats de baix calibre, és a dir, exceptuant algunes notícies moltes d'elles no eren semblants a la consulta. Recuperaven alguna paraula igual, però el context era diferent.

7.2.9. Altres Resultats

En aquest apartat mostrarem els valors resultats d'algunes consultes. No entrarem en un anàlisis profund, ja que, moltes de les consultes realitzades ja aporten suficient informació en el tòpic a consultar. En altres paraules, aquest consultes som molt similar a algunes fetes o tracten de temes similars.

7.2.9.1. Resultat 7ª Consulta

Consulta 7: El jutge decreta error en la sentència

Model Booleà:	
37242	
18642	
4665	
2693	
315	
63985	
41500	
1192	
10910	
Nombre d'articles:	9

(a) Model Booleà

Model Word2Vec	Model STSB
82802	59458
5555	6731
19116	79380
12402	3759
71688	73223
31232	42560
1252	22698
4665	12958
11594	8306
60736	15526

(b) Model Word2Vec i STSB

Taula 7.11: Taules dels id dels articles resultats de la consulta "El jutge decreta error en la sentència"

Els resultats en mes detall en la taula B.30 i B.31

7.2.9.2. Resultats 8ª Consulta

Consulta 8: Els diputats votaran aquest dijous

Model Booleà:			
52626	3550	15335	4303
79625	80663	38788	9758
74790	70450	68627	12971
53974	444	33434	59519
40599	9613	3915	24817
24681	20241	3424	11299
54577	20424	21954	68444
7632	2303	2272	9744
58323	60929	25625	
7625	39809	59047	
38018	20240	25006	
80298	26408	59035	
21599	10933	7219	
Nombre d'articles:		47	

(a) Model Booleà

Model Word2Vec	Model STSB
53943	27580
21705	66632
2163	40758
77575	3708
9025	13008
89372	34765
37816	15459
71809	3260
9469	14801
89615	7719

(b) Model Word2Vec i STSB

Taula 7.12: Taules dels id dels articles resultats de la consulta "Els diputats votaran aquest dijous"

Els resultats en mes detall en la taula B.32 i B.33

7.2.9.3. Resultats 11ª Consulta

Consulta 11: La mesa electoral va cometre dos errors

Model Booleà:	
1986	13519
Nombre d'articles:	2

(a) Model Booleà

Model Word2Vec	Model STSB
68064	34596
21456	22803
53198	68154
86422	55494
33408	28692
8829	23769
67371	28943
47024	53621
84859	30143
18429	24591

(b) Model Word2Vec i STSB

Taula 7.13: Taules dels id dels articles resultats de la consulta L·La mesa electoral va cometre dos errors"

Els resultats en mes detall en la taula B.38 i B.39

7.2.9.4. Resultats 12ª Consulta

Consulta 12: Els independentistes no accepten el tracte oferit pel Govern

Model Booleà:	
26830	
Nombre d'articles:	1

(a) Model Booleà

Model Word2Vec	Model STSB
52746	63121
21289	3918
21520	16197
12094	2227
25551	5423
20143	22557
5985	3266
63201	7965
24071	23209
42830	32916

(b) Model Word2Vec i STSB

Taula 7.14: Taules dels id dels articles resultats de la consulta "Els independentistes no accepten el tracte oferit pel Govern"

Els resultats en mes detall en la taula B.40 i B.41

7.2.9.5. Resultats 13ª Consulta

Consulta 13: El deute econòmic creix

Model Booleà:	
50150	
49984	
44533	
33461	
76029	
34656	
53485	
77404	
95010	
62474	
Nombre d'articles:	10

(a) Model Booleà

Model Word2Vec	Model STSB
69668	18791
24700	85066
50696	56443
32963	76985
86551	30164
32178	31715
26216	52344
2647	55572
19766	59127
78474	62474

(b) Model Word2Vec i STSB

Taula 7.15: Taules dels id dels articles resultats de la consulta "El deute econòmic creix"

Els resultats en mes detall en la taula B.42 i B.43

Per a la resta de resultats: https://upvedues-my.sharepoint.com/:f:/g/personal/ancase3_upv_edu_es/EmUCtZorwwtFk1pOdhGFC4wBQbSFfjbZ3WrPlGxe-gm25Q?e=1TQy7d.

7.3 Conclusions a partir dels Resultats

Aquest apartat constarà de 3-4 seccions on anirem avaluant els models individualment i, per últim, una comparació entre els models. És a dir, avaluarem el rendiment i encert en

resultats dels diferents models creats, mostrant les característiques principals que extrem de cadascú. Finalment, amb les conclusions extretes farem una comparació de models.

7.3.1. Model Booleà

Comentant amb anterioritat ha sigut el punt de partida per a les avaluacions, és a dir, hem seleccionat els seus resultats com a correctes i hem vist quins articles se superposaven amb els altres models.

Primer, el model Booleà és el model mes simple dels creats i sols retorne aquells articles que contenen les paraules cercades, conseqüentment hem hagut de modificar les consultes per cercar les paraules mes importants perdent part de les característiques de la frase cercada (*Els independentistes no accepten el tracte oferit pel Govern –> independentistes, no, accepten, tracte, Govern*). A més, un dels principals diferencies que podem trobar en el model Booleà és el fet que les paraules poden estar distribuïdes per tot l'article i no ser components d'una mateixa frase, per tant, recuperant articles que no tenen relació directa amb la consulta. Com es pot comprovar en 7.2.9.5, molts dels articles recuperats pel model, no tenen similitud amb la consulta donada. Exemple:

"El carnet número 1 de CDC va ser per a Elisabet Puig (de fet va ser el 3.001, per donar gruix a la militància)" $\neq$ *El president va ser expulsat del congrés*

Per altra banda, ja sabíem de les desavantatges del model Booleà, un model que no té en compte el context ni el significat semàntic, no obstant la seua finalitat que era recuperar notícies que contenien les paraules cercades ha sigut un èxit.

En aquest projecte, hem comprovat com algunes consultes realitzades no tenien resultat a causa que la combinació de paraules cercades no tenien cap resultat possible. és a dir, cap document contenia aquestes paraules. En canvi, si contenien sinònims o construccions similars.

equip local guanyar golejada resultat: 0
2º resultat tronat pel model STSB = *La Trinca el guanya per golejada*

Podem observar, com el model STSB torna un resultat molt acorde a la consulta realitzada. En canvi, el Booleà no ha trobat cap resultat.

En resum, trobem un model molt simple, on la cerca de notícies és o molt extensa o molt curta, ja que segons el conjunt de paraules cercades trobarem un nombre diferent de resultats. Tanmateix, els resultats obtinguts seran el conjunt de totes les notícies que continguen les paraules i depenent del significat de les paraules (algunes poden tindre polisèmia) tindrem resultats concordes a la consulta i altres que no. És a dir, torna resultats que contenen les paraules, però no el cos de l'article no té similitud amb la consulta.

7.3.2. Model Word2Vec

Després d'analitzar diverses consultes realitzades, hem comprovat com en moltes d'elles els resultats tornats pel model no s'adequaven a la consulta, tornava moltes voltes resultats estrany que no mantenien ni el context ni fèiem cap classe de *matching*, tal com hem explicat en el anàlisi de la consulta 3 (7.2.3.2). No sols, en l'exemple, sinó que en moltes altres consultes trobàvem casos pareguts, on alguns resultats sí que coincidien però d'altres no. Per tant, el grau de satisfacció en aquest model és baix.

Analitzant la seua funcionalitat podem observar com en el següent cas:

La nova llei es debatrà al febrer a l'Assemblea Nacional ≈ Nova llei aprovada al congrés

On trobem un context similar a la consulta realitzada. Per tant, en alguns casos sí que ha funcionat.

Hem fet l'hipotesis que el possible error que ha ocorregut per obtindre un model amb tan baixa precisió és el fet d'utilitzar un model *ca-core-news-md* preentrant amb un corpus que no és el nostre. Per tant, les possible relacions que ha fet han sigut diferents de les que necessitem. És a dir, hem utilitzat un model que ha entrenat en un corpus de notícies i articles diferent del nostre, i ha pogut crear relacions contextuals diferents.

Per altra banda, la poca precisió pot ser causada pel fet que els models Word2Vec no són contextuals, són la mitjana de tots el vectors de les paraules que creen les frases, produint un aplanament dels vectors resultants. Per tant, i tenint en compte l'eliminació dels *stopwords*, obtenim un resultat poc precís.

En conclusió, hem pogut analitzar els funcionament del model Word2Vec i treu resultat que han de veure amb el context de la paraula, però hem obtingut una precisió baixa de satisfacció a causa del fet que molts del contextos es pareixien o un mal preentrament del model. És a dir, com no diferencia semànticament moltes de les paraules cercades podien estar en un context o altre, per tant, els resultats obtinguts no eren satisfactoris. Tanmateix, cal tindre en compte que el model Word2Vec és incontextual, no entén el context on està la paraula, per tant, trobem uns resultats poc similars a la consulta, però acordes al que la teoria del model explica.

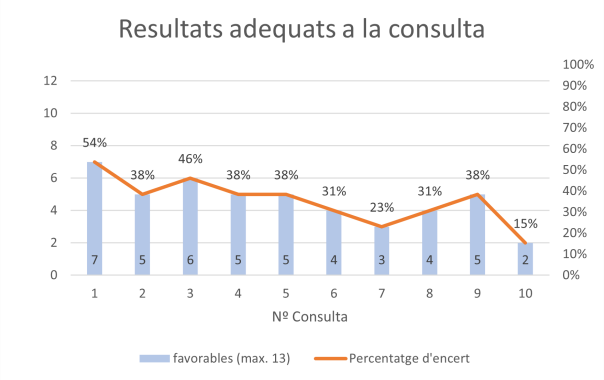


Figura 7.1: Nombre de resultats adequats per posició en el ràquing model Word2Vec

Consulta	favorables (max. 13)	Percentatge d'encert
1ª	7	54%
2ª	5	38%
3ª	6	46%
4ª	5	38%
5ª	5	38%
6ª	4	31%
7ª	3	23%
8ª	4	31%
9ª	5	38%
10ª	2	15%

Taula 7.16: Percentatge de consultes favorables per posició en el ràquing model Word2Vec

7.3.3. Model STSB

El model STSB ha obtingut un alt grau de satisfacció a l'hora d'analitzar els resultats, no sols ens tornava resultats que tenien en compte el context i mantenien el tòpic parlat, sinó que mantenien el significat semàntic de les paraules i frases cercades. Com es poden comprovar en les seccions anteriors, molts dels resultats extrets han sigut satisfactoris. Exemple:

Les enquestes pronostiquen un gran creixement d'ERC, que guanyaria per primera vegada les eleccions al Parlament des de la Segona República ≈ El partit d'esquerres guanyarà les futures eleccions.

No sols ha extret un resultat que manté el context, sinó que moltes de les paraules que apareixen fan un *matching* directe. A més, com es pot observar hi ha clara relació de sinonímia interessant:

$$ERC \approx partit\ d'esquerres$$

És a dir, ha relacionat a Esquerra Republicana de Catalunya (ERC) amb partit d'esquerres. Demostrant així la capacitat d'associar sinònims del model, una habilitat increïble per recuperar articles relacionats directament amb la consulta sense la necessitat de tindre les mateixes paraules.

En conclusió, el model STSB creat ha sigut capaç d'extraure els resultats correctes amb un tassa d'encert alta, amb la capacitat d'extraure resultats que tenen característiques contextuals i semàntiques similars a la cerca. No sols això, sinó que ha sigut capaç d'extraure sinònims, per tant, hem demostrat un rendiment alt i un funcionament correcte.

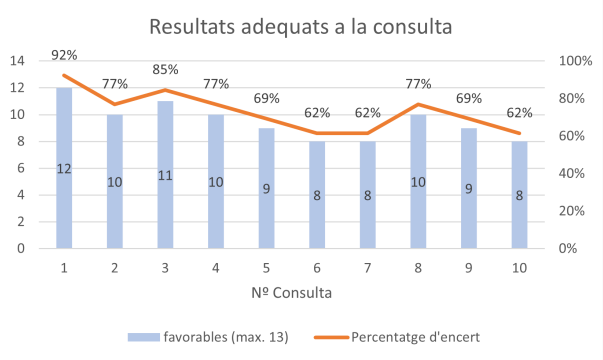


Figura 7.2: Nombre de resultats adequats per posició en el rànquing model STSB

Consulta	favorables (max. 13)	Percentatge d'encert
1ª	12	92%
2ª	10	77%
3ª	11	85%
4ª	10	77%
5ª	9	69%
6ª	8	62%
7ª	8	62%
8ª	10	77%
9ª	9	69%
10ª	8	62%

Figura 7.3: Nombre de consultes favorables per posició en el rànquing model STSB

7.3.4. Comparació dels Models

Analitzant els tres models conjuntament, mes que una comparació per deduir quin és millor o pitjor, ha sigut un anàlisi a l'evolució dels SRI al llarg del temps. És a dir, hem recorregut els diferents models creats en diferents èpoques comprovant com cada model depèn de l'anterior per al seu desenvolupament, per exemple el model STSB és l'evolució natural del Word2Vec, on un model és contextual (entén el context de les paraules tenint en compte la seua semàntica) i l'altre és incontextual (on el vector és la mitjana de les representacions vectorials de les paraules que componen la frase).

En altres paraules, gràcies a la creació dels diferents models SRI hem seguit els passos de la seua evolució, com hem passat d'un model que sols pot comprovar paraules a un que té en compte la semàntica on apareix la paraula o la frase, i per últim a un que a més d'entendre el context entén la semàntica de la frase.

Un altre dels aspectes a avaluar és la unitat d'indexació, ja que hem utilitzat la frase extreta pel tokenitzador de NLTK, que retalla els cossos dels articles en frases, hem comprovat en algunes casos que les frases creades són molt llargues, i perfectament podrien haver retallat per transformar-les en 2 o 3 frases, així augmentant-la precisió de recuperació.

CAPÍTOL 8
Conclusions

Finalment, després d'un llarg procés de desenvolupament, s'han implementat 6 sistemes de recuperació d'informació, 3 en català i 3 en castellà, utilitzant diferents models SRI: un model clàssic (Booleà), que indica si una paraula està o no en un document; un model de representacions vectorials incontextuals (Word2Vec), que transforma les frases en un vector, on és la mitjana de les representacions vectorials de les paraules que componen la frase; un model de representacions contextual (STSB), un model que crea embedding semàntics contextuals, que té la capacitat d'establir distancia entre textos tenint en compte el context i la semàntica, en altres paraules, pot calcular la similitud entre textos.
En segon lloc, hem hagut de fer una sèrie de proves per comprovar el correcte funcionament dels models, i, per últim hem aconseguit obtindre resultats.
Després d'analitzar els resultats obtinguts podem confirmar que hem satisfet l'objectiu principal d'aquest projecte, desenvolupar un sistema de recuperació d'informació basat en representacions denses. Els resultats ens mostren un SRI que utilitza el context i el significat semàntic de les paraules i frases per recuperar la informació buscada. No només, hem aconseguit complir l'objectiu principal, sinó també els secundaris, a més, a mesura que es desenvolupava els projecte nous subobjectius han sortit. Aquest subobjectius es relacionen amb problemes que han sorgit al llarg del desenvolupament, i s'han convertit en unes necessitats bàsiques a tindre en compte. Els subobjectius són els següents:

- Implementació dels 3 models en castellà per utilitzar-los com a referència per comprovar el correcte funcionament dels models en català.
- Una implementació més eficient a l'hora d'utilitzar recursos

Respecte de les coses que he après en aquest TFG, no només ha sigut necessari entendre el funcionament dels SRI, sinó que també he hagut d'aprofundir en l'estudi teòric de les eines i models a utilitzar. L'ús de les diferents llibreries (*SpaCy, NLTK, Whoosh, SentenceTransformer*, etc.) ha sigut fonamental a l'hora de la implementació, ja que un dels principals reptes d'aquest projecte que ha sortit en el projecte ha sigut desenvolupar des de zero un nou model de similitud semàntica de textos per a català com és el STSB. Estem parlant d'uns del models mes avançats en l'àmbit dels SRI, per tant, d'uns del models amb més difícil implementació. En conclusió, amb les noves tecnologies implementades per les llibreries he sigut capaç de desenvolupar els models necessaris d'una forma eficient.

Aquest ha sigut un del projectes més grans que he fet per ara, i no sols ha sigut un repte per mi, sinó que m'ha ajudat a desenvolupar habilitats que necessitaré en un futur. El meu coneixement de Python i llibreries PLN s'han vist incrementats, ja que he hagut de resoldre molts problemes relacionats amb elles. Una de les habilitats que més he millorat

ha sigut el fet de poder resoldre per mi mateix els errors, siga buscant-los en la web o tindre la capacitat i coneixement necessari per resoldre-ho per mi mateixa, la creativitat i el pensament crític han sigut claus en aquest punt. Tanmateix, he descobert que el món del PLN és molt més extens i complicat, descobrint noves funcionalitats i experimentant-les he sigut capaç d'entendre l'enorme àmbit i un munt d'oportunitats noves.

8.1 Relació del treball desenvolupat amb els estudis cursats

L'assignatura de SAR, sistemes d'emmagatzemament i recuperació d'informació, com el seu propi nou indica, és el que més s'apropa als coneixements necessaris per desenvolupar aquest projecte. Més enllà del nom, per al projecte de l'assignatura vam haver de desenvolupar un SRI Booleà des de 0 sense utilitzar cap mena de llibreria especialitzada, marcant així el meu inici en el món dels SRI.

No sols SAR ha sigut útil, sinó que totes aquelles assignatures que m'han ajudat a entendre els algoritmes eficients i els temps d'execució han fet que millore el rendiment d'aquest projecte, deduint i implementat els algoritmes mes eficients i ràpids.

1. Algorítmica
2. Computabilitat i complexitat
3. Tècniques d'optimització
4. Estructures de dades i algorismes

CAPÍTOL 9
Treballs Futurs

En aquest últim capítol detallarem possibles millores que es podrien fer, tot i que el projecte ja és prou extens. No sols parlem de millores d'eficiència o de resultats, sinó d'obrirnos a noves solucions o solucions mes amplies que és podríem implementar.

Millores:

- Utilitzar targetes gràfiques per l'aprenentatge del models i per indexar el corpus mes ràpidament, per mitjà la llibreria CUDA (Compute Unified Device Architecture) per disminuir el temps d'execució.
- La implementació d'altres models per augmentar el rang de la comparació.
- La creació d'una interfície adequada al projecte per poder utilitzar-lo mes còmodament.
- Millorar l'ús de la memòria, ja que algunes limitacions del projecte són derivats de la manca de recursos.
- Augmentar la grandària del corpus.

En conclusió, uns dels problemes més grans al que ens hem hagut d'enfrontar és a la demora en temps d'execució dels processos, ja que, o els algoritmes utilitzats no són eficients o els recursos usats són insuficients. En general, amb més temps podríem haver resolt molts del problemes i millorar el projecte.

Bibliografia

[1] Pérez-Carballo, J. and Strzalkowski, T. Natural language information retrieval: progress report'. *Information Processing and Management 36*, p. 155-178, 2000.

[2] Dominich, S. A unified mathematical definition of classical information retrieval. *Journal of the American Society for Information*, Science, 51 (7), 2000. p. 614-624.

[3] Encarna Segarra, Vicent Ahuir, Lluís-F. Hurtado, José Ángel Gonzalez. DACSA: A large-scale Dataset for Automatic summarization of Catalan and Spanish newspaper Articles *proceding of the Annual Conference of the North American Chapter of the Association for Computational Linguistics* 2022

[4] Croft, W. B. 'Approaches to intelligent information retrieval'. *Computer and Information Science Department, University of Massachusetts, Amherst, MA 01003, USA* 1987

[5] Longley, D. and Shain M. Mac Millan. *Dictionary of IT*. London and Basingstoke: The MacMillan Press, 1989.

[6] Korfhage, R.R. *Information Retrieval and Storage*. New York: Wiley Computer Publisher, 1997.

[7] R. Baeza-Yates, EDS. *Information Retrieval: Data Structures and Algorithm's*. Englewood Cliffs (NJ): Prentice-Hall, 1992, 241–263

[8] Rijsbergen, C.J. Information Retrieval Glasgow, University, 1999 Consulta: 14 de març de 2022 http://www.dcs.gla.ac.uk/Keith/Preface.html.

[9] Baeza-Yates, R. and Ribeiro-Neto, B. Modern information retrieval 1999 Consulta: 2 de febrer de 2022 http://web.cs.ucla.edu/~miodrag/cs259-security/baeza-yates99modern.pdf.

[10] Berry, M. W. y Castellanos, M. (2007). Survey of Text Mining : Clustering , Classification , and Retrieval , Second Edition. Consulta: 21 de febrer de 2022 https://perso.uclouvain.be/vincent.blondel/publications/08-textmining.pdf.

[11] Mikolov, T., Chen, K., Corrado, G. y Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space, pp. 1–12. Consulta: 5 de març de 2022 http://arxiv.org/abs/1301.3781.

[12] Nils Reimers, Iryna Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks (2019) Consulta: 22 de maig de 2022 https://arxiv.org/abs/1908.10084.

[13] STS benchmark dataset and companion dataset Consulta: 2 de maig de 2022 https://huggingface.co/datasets/stsb_multi_mt.

[14] distilbert-base-uncased trained for Semantic Textual Similarity in Spanish Consulta: 2 de maig de 2022 https://huggingface.co/eduardofv/stsb-m-mt-es-distilbert-base-uncased.

[15] es_core_news_md Consulta: 10 de març de 2022 https://spacy.io/models/es.

[16] ca_core_news_md Consulta: 10 de març de 2022 https://spacy.io/models/ca.

[17] Llibreria Whoosh Consulta: 5 març de 2022 https://whoosh.readthedocs.io/en/latest/intro.html

1 OBJECTIUS DE DESENVOLUPAMENT SOSTENIBLE

Grau de relació del treball amb els Objectius de Desenvolupament Sostenible (ODS).

Objectius de Desenvolupament Sostenible	**Alt**	**Mitja**	**Baix**	**No procedeix**
ODS 1. **Fi de la pobresa.**				X
ODS 2. **Fam cero.**				X
ODS 3. **Salut i benestar.**				
ODS 4. **Educació de qualitat.**	X			
ODS 5. **Igualtat de gènere.**				X
ODS 6. **Aigua neta i sanejament.**				X
ODS 7. **Energia assequible i no contaminant.**				
ODS 8. **Treball decent i creixement econòmic.**				X
ODS 9. **Indústria, innovació i infraestructures.**		X		
ODS 10. **Reducció de les desigualtats.**	X			
ODS 11. **Ciutats i comunitats sostenibles.**				X
ODS 12. **Producció i consum responsable.**				X
ODS 13. **Acció pel clima.**				X
ODS 14. **Vida submarina.**				X
ODS 15. **Vida d'ecosistemes terrestres.**				X
ODS 16. **Pau, justícia i institucions sòlides.**				X
ODS 17. **Aliances per aconseguir objectius.**	X			

Reflexió sobre la relació del TFG/TFM amb els ODS i amb el/els ODS mes relacionats.

En aquest projecte hem impulsat em major mesura els ODS 4 i ODS 10. L'objectiu principal d'aquest projecte és crear un SRI amb les últimes tecnologies conegudes per a uns llenguatges minoritaris, en aquest cas el Català. És a dir, que les universitats, individus o institucions siguen capaços d'utilitzar aquest SRI per poder indexar els documents que ells desitgen i després fer consultes satisfactòries. Aconseguint així, una reducció de la desigualtat davant les llengües majoritàries que tenen el privilegi de comptar amb les tecnologies mes avançades en l'àmbit dels SRI, i millorar la qualitat de l'educació actual en Català. Ja que, permetríem l'ús d'aquest SRI per poder indexar milers de llibres o articles i que els estudiants puguen fer ús del buscador per saber quins són els llibres que mes els interessa en el seu aprenentatge, o per satisfer els seus interessos, i d'altres funcions.

Tanmateix, hem pujat el model STSB en Català per poder ser usat en altres aspectes de l'àmbit NLP, aportant noves tecnologies al mercat actual i ajudar a desenvolupar la comunitat de parla Catalana. Produint noves oportunitats d'invocació i ajudant-los en el procés, tenint relació en l'apartat ODS 9.
Per últim, vàrem desenvolupar aquest projecte en l'esperança de ser utilitzat com a eina que ajude a les institucions, universitats, etc. a poder seguir desenvolupant els seus serveis al món, per tant, esperem l'oportunitat de treballar conjuntament per seguir millorant aquest projecte i proporcionar noves ferramentes al món. Tenint relació amb el ODS 17.

APÈNDIX A
Configuració del sistema

A.1 Identificació de dispositius

A.1.1. Procés d'indexació

```
#recorem els directoris i arxius
for root, subdirs, files in os.walk(path):
            #partial count
            pcnt = 0

            #inicialitzaci  de la variable valor_doc com id dels documents
            valor_doc = 0

            #recorrem els arxivaments en el directori
            for filename in files:

                #entren si  s  un arxiu en l'extensi  .jsonl
                if filename.endswith('.jsonl'):

                    #Variable que es guarda el path del arxiu
                    fullname = os.path.join(root, filename)

                    #obrim l'arxiu
                    with open(fullname, "r",encoding="utf-8") as fh:

                        #Reocrrem els diferents articles que hi ha en l'
                                                                arxiu
                        for i, obj in enumerate(fh):

                            #Carreguem l'objecte en mem ria
                            d=json.loads(obj)

                            #En el diccionari guardem el valor_doc en el
                                                                    path
                                                                    obert i
                                                                    en quina
                                                                    lineea es
                                                                     troba l'
                                                                    article
                            self.dic_doc[valor_doc] = (fullname, i)

                            #Separem en un llista les frases que contenen l
                                                                    'article
                            tro_doc = self.tallar(d['article'])

                            #Recorrem les diferents frases
                            for j, frase in enumerate(tro_doc):
```

```
                            #Si estem indexant amb un model Word2Vec, hem d'eliminar les stopwords.
                            if opcio == 0:
                                frase = self.eliminar_stopwords(frase)

                            #Calculem el vector representatiu
                            #de la frase
                            vector = self.calcular_valor(frase, opcio)

                            #transformem el vector en una tuple per
                            #ser indexada
                            a = tuple(vector)

                            """Guardem en el diccionari
                            el vector representatius com a key
                            i valor associat el n mero
                            de frase que s mes
                            el id del document"""
                            self.dic_text.setdefault(a, []).append((valor_doc, j))

                        #Fora del bucle incrementem el valor del id
                        valor_doc += 1

                        #Comprobaci per vore si tot funciona
                        #correctament
                        if valor_doc % 1000 == 0:
                            print("Indexat document: %d" % valor_doc)
                            with open(output, "a") as fh:
                                fh.write("Indexat document: %d" % valor_doc)

                                fh.write("\n")

                        #Para la indexaci als 50000,
                        #ja que si no tindrem problemes
                        #de mem ria
                        if valor_doc == 50000:
                            break

            #Escriu en un arxiu el documents indexats
                pcnt += 1
            if pcnt > 0:
                with open(output, "a") as fh:
                    fh.write("%s: %d" % (root, pcnt))
                    fh.write("\n")
                    fh.write("Numero de documents indexats: %d" %valor_doc)
                    fh.write("\n")
                cnt += pcnt
            elif len(files) > 0:
                with open(output, "a") as fh:
                    fh.write("%s: 0" % root)
                    fh.write("\n")

        #Escriu en un arxiu el temps d'indexaci i el nombre de documents
        ptime = time.time() - time_start
        with open(output, "a") as fh:
            fh.write("=" * 30)
```

```
            fh.write("\n")
            fh.write("Docs: %d, Time: %.1fs." % (cnt, ptime))
            fh.write("\n")
            fh.write("=" * 30)
            fh.write("\n")

        ttime_start = time.time()

        #matrix com una llista del keys del diccionari per poder crear el KDTree
        matrix = list(self.dic_text.keys())

        #Creaci  del KDTree els nodes son els vectores representatius de les frases
        self.tree = scipy.spatial.KDTree(matrix)

        ttime = time.time() - ttime_start

        #Escriu el temps que ha transcorregut mentres es crea el arbre
        with open(output, "a") as fh:
            fh.write("Temps de creaci  del abre " + idioma + ": %.1fs." % (ttime))
            fh.write("\n")
            fh.write('=' * 30)
            fh.write("\n")

        print("Acabat de cargar")
```

A.1.2. Procés de cerca

```
        #El model ha sigut carregat amb anterioritat

       #Carreguem en mem ria el corpus a llegir
        with open(path,"r",encoding="utf-8") as fh:
            content = fh.readlines()

        #Recorrem la llista de consultes
        for query in consultes:

             #Si la consulta es incorrecta no tornem res
            if query is None or len(query) == 0:
                return []

            #Escribim quina ha sigut la consulta feta
            with open(output,'w') as f:
                f.write("La query es: " + query)
                f.write("\n")

            #Si estem cercant amb un model Word2Vec, hem d'eliminar les stopwords.
            if opcio == 0:
                frase = self.eliminar_stopwords(frase)

            #Calculem les representaci  vectorial de la query
            doc = self.calcular_valor(query, opcio)

            #Calculem els n ve ns mes propers
            veins = self.tree.query(doc, k = int(num_doc))

            #El m tode torna en la posici  1 les posicions del vectors
            noticies = veins[1]

```

```
            #I en la posici  0 els valors de les distancies
            valors = veins[0]
            resultat = []


            if len(noticies) == 0:
                print("Cap resultat")
            else:
                #Variable que s'utilitza per contabilitzar la posici
                # en la llista de valors
                i = 0

                print("=" * 30)
                print("\n")
                #Recorrem les noticies recuperades
                for r in noticies:

                    #Fem l'extraci  del vector representatiu
                    value_list = self.tree.data[r]

                    #Convertim el valor en un tupla com feiem en la
                                                            indexaci
                    valor = tuple(value_list)

                    #Del diccionari dic_text extraem els id dels articles a
                                                             recuperar
                    #ja que una mateixa frase pot estar en mes d'una
                                                            noticia
                    # i quin es la posici  de la frase similar
                    valors_ nics = self.dic_text[valor]

                    #Recorrem els valors  nics
                    for valor_ nic , frag in valors_ nics:

                        #Del diccionari dic_doc extraem els path del
                                                               article
                        path_1, line = self.dic_doc[valor_ nic]

                        #Carreguem l'article agafant la posici  on
                        #es troba en l'arxiu complet
                        obj = json.loads(content[line])

                        #Del cos de l'articles el partim en frases
                        tro_doc = self.tallar(obj['article'])

                        #Escribim en un document quin era el valor  nic
                        # el summary de l'article, quina ha sigut la frase
                        #similar a la consulta i per  ltim  la dist ncia
                        # a la consulta
                        with open(output, "a") as f:
                            f.write(str(valor_ nic))
                            f.write("\n")
                            f.write(obj['summary'])
                            f.write("\n")
                            f.write(tro_doc[frag])
                            f.write("\n")
                            f.write("Valor de la dist ncia entre les
                                                                  frases:
                                                                  %f " %
                                                                  valors[i]
                                                                  )
                            f.write("\n")

                        resultat.append(obj[line]['url'])
```

```
88
89                      i = i + 1
90                  with open(output, "a") as f:
91                      f.write("=" * 30)
```

APÈNDIX B
Altres

B.1 Temps d'execució

B.1.1. temps d'execució per consulta

Índex	Número Consulta	169	1000	10000	90000	100000
Indexador Word2Vec castellà	consulta 1:	0,01	0,01	0,1	-	3,63
	consulta 2:	0,01	0,01	0,09	-	1,19
	consulta 3:	0,01	0,01	0,09	-	1,06
	consulta 4:	0,01	0,01	0,016	-	1,05
	consulta 5:	0,01	0,01	0,09	-	1,06
Indexador STSB castellà	consulta 1:	0,07	0,03	0,19	-	2,42
	consulta 2:	0,02	0,02	0,16	-	1,88
	consulta 3:	0,02	0,02	0,16	-	0,7
	consulta 4:	0,02	0,02	0,16	-	0,72
	consulta 5:	0,02	0,03	0,16	-	0,7
Indexador Clàssic castellà	consulta 1:	0,48	0,31	0,11	-	0,42
	consulta 2:	0	0	0,1	-	0,7
	consulta 3:	0	0	0,5	-	0,41
	consulta 4:	0	0	0,6	-	0,42
	consulta 5:	0	0	0,5	-	0,7
Indexador Word2Vec català	consulta 1:	0,01	0,04	0,19	3,45	-
	consulta 2:	0,01	0,02	0,19	0,63	-
	consulta 3:	0,01	0,02	0,21	0,63	-
	consulta 4:	0,01	0,02	0,19	0,64	-
	consulta 5:	0,01	0,02	0,19	0,62	-
Indexador STSB català	consulta 1:	0,05	0,07	0,31	0,94	-
	consulta 2:	0,02	0,06	0,31	0,76	-
	consulta 3:	0,02	0,05	0,31	0,75	-
	consulta 4:	0,02	0,05	0,31	0,75	-
	consulta 5:	0,02	0,06	0,31	0,76	-
Indexador Clàssic català	consulta 1:	0,01	0,4	0,11	0,66	
	consulta 2:	0	0,01	0,1	0,77	-
	consulta 3:	0	0,01	0,5	0,38	-
	consulta 4:	0	0,01	0,6	0,47	-
	consulta 5:	0	0,01	0,5	0,46	-

Taula B.1: Taula del temps d'execució de totes les consultes de prova en tots els índexs

B.2 Resultats proves

B.2.1. Corpus 169 articles

Consulta: Jugar fuera de casa

- **Model STSB castellà:**

Rànquing	Valor únic	Summary	Linia Similar	Distancia
1	157	Todos los datos y el resultado al minuto del partido entre Arina Gabriela Vasilescu - Gina Marie Dittmann del ITF Germany 08A 2019.	La jugadora rumana Arina Gabriela Vasilescu, número 1436 de la WTA, cumplió los pronósticos al ganar por 6-2 y 6-4 a Gina Marie Dittmann, tenista alemana en la ronda previa de calificación del torneo de Darmstadt.	11.572.818
2	160	Todos los datos y el resultado al minuto del partido entre Manon Arcangioli - Vivian Wolff del ITF Germany 08A 2019.	El torneo de Darmstadt (ITF Germany 08A) cuenta con una fase de acceso previa donde las jugadoras con menor ránking tienen que obtener los mayores puntos posibles para lograr clasificarse y participar en el torneo oficial.	12.229.351
3	39	Todos los datos y el resultado al minuto del partido entre Estela Pérez-Somarriba - Katarina Jokic del ITF USA 41A 2019.	Estela Pérez-Somarriba, española venció en dos horas y cuarenta y siete minutos por 7(7)-6(2) y 6-4 a Katarina Jokic, tenista serbia en la ronda previa de calificación del torneo de Macon.	12.371.085
4	119	Todos los datos y el resultado al minuto del partido entre Valentina Losciale - Ines Oliveira del ITF Potugal 14A 2019.	El torneo de Lousada (ITF Portugal 14A) cuenta con una fase de acceso previa donde las jugadoras con menos ránking tienen que obtener la máxima puntuación posible para lograr clasificarse y participar en el torneo oficial.	12.381.634
5	74	Todos los datos y el resultado al minuto del partido entre Fernando Verdasco - Nikoloz Basilashvili del ATP Viena 2019.	El español jugará en los octavos de final de la competición contra el ganador del partido en el que se enfrentarán el austriaco Dominic Thiem y el tenista francés Jo-Wilfried Tsonga.	12.460.785

Taula B.2: Rànquing de la consulta "Jugar fuera de casa"per model STSB castellà

- **Model Word2Vec castellà:**

Rànquing	Valor únic	Summary	Linia Similar	Distancia
1	157	Todos los datos y el resultado al minuto del partido entre Arina Gabriela Vasilescu - Gina Marie Dittmann del ITF Germany 08A 2019.	La jugadora rumana Arina Gabriela Vasilescu, número 1436 de la WTA, cumplió los pronósticos al ganar por 6-2 y 6-4 a Gina Marie Dittmann, tenista alemana en la ronda previa de calificación del torneo de Darmstadt.	11.572.818
2	160	Todos los datos y el resultado al minuto del partido entre Manon Arcangioli - Vivian Wolff del ITF Germany 08A 2019.	El torneo de Darmstadt (ITF Germany 08A) cuenta con una fase de acceso previa donde las jugadoras con menor ránking tienen que obtener los mayores puntos posibles para lograr clasificarse y participar en el torneo oficial.	12.229.351
3	39	Todos los datos y el resultado al minuto del partido entre Estela Pérez-Somarriba - Katarina Jokic del ITF USA 41A 2019.	Estela Pérez-Somarriba, española venció en dos horas y cuarenta y siete minutos por 7(7)-6(2) y 6-4 a Katarina Jokic, tenista serbia en la ronda previa de calificación del torneo de Macon.	12.371.085
4	119	Todos los datos y el resultado al minuto del partido entre Valentina Losciale - Ines Oliveira del ITF Potugal 14A 2019.	El torneo de Lousada (ITF Portugal 14A) cuenta con una fase de acceso previa donde las jugadoras con menos ránking tienen que obtener la máxima puntuación posible para lograr clasificarse y participar en el torneo oficial.	12.381.634
5	74	Todos los datos y el resultado al minuto del partido entre Fernando Verdasco - Nikoloz Basilashvili del ATP Viena 2019.	El español jugará en los octavos de final de la competición contra el ganador del partido en el que se enfrentarán el austriaco Dominic Thiem y el tenista francés Jo-Wilfried Tsonga.	12.460.785

Taula B.3: Rànquing de la consulta "Jugar fuera de casa"per model Word2Vec castellà

- **Model Clàssic castellà:**

Resultats de la consulta = 0

Consulta: El partit d'esquerres va pedre les eleccions

- **Model STSB català:**

Rànquin	Valor úni	Summary	Linia Similar	Distancia
1	81	El partit encara manté obertes les negociacions amb MES, Avancem i Demòcrates i amb independents.	Les enquestes pronostiquen un gran creixement d'ERC, que guanyaria per primera vegada les eleccions al Parlament des de la Segona República.	10,74
2	83	Eren les declaracions que va fer al programa de Mònica Terribas a Catalunya Ràdio.	En aquesta ocasió, el programa que presenta i dirigeix Jenaro Castro ha indignat part de l'audiència de la pública espanyola perquè en el primer reportatge que s'emetia, anomenat 'El periplo catalán', s'ha utilitzat la banda sonora de la pel·lícula 'L'exorcista' per acompanyar unes declaracions del president Carles Puigdemont.	11,05
3	81	El partit encara manté obertes les negociacions amb MES, Avancem i Demòcrates i amb independents.	La candidatura d'Esquerra ha incorporat a la gran majoria dels diputats del partit en la legislatura que es va dissoldre fa dues setmanes per odre del govern espanyol.	11,17
4	37	L'exlíder de Podem insisteix que no anirà a una llista d'ERC ni de la CUP el 21-D.	Tras este resultado, segu"He parlat aquests últims dies amb gent d'ERC, de la CUP, del Procés Constituent, amb gent que més enllà de compartir la independència com a objectiu polític creu que hi ha una manera d'expressar-se des de l'esquerra, que la sobirania recau en el poble de Catalunya.iremos viendo a la pareja ganadora en los cuartos de final del torneo de Szekesfehervar.	11,31
5	59	Els republicans obren les portes de bat a bat a Albano Dante Fachin, amb qui seguiran conversant.	encapçalada per Oriol Junqueras i comptaria amb els presos polítics vinculats al partit, amb els diputats actuals -que van votar la DUI- i amb els alts càrrecs del Govern que han treballat per al referèndum -s'hi van comprometre per escrit a l'abril-.	11,33

Taula B.4: Rànquing de la consulta "El partit d'esquerres va pedre les eleccions"per model STSB català

- **Model Word2Vec català:**

Rànquin	Valor úni	Summary	Linia Similar	Distancia
1	70	Les entitats inicien amb una encartellada una setmana de mobilitzacions per la llibertat dels presos.	de l'ANC, que va demanar superar les 450.000 persones que, el passat dia 21, van exigir l'alliberament de Cuixart i Sànchez.	11,65
2	43	Més de 180 intel·lectuals i polítics defensen un front comú el 21-D.	Hores després de presentar-se, el manifest va rebre el suport de l'alcaldessa Ada Colau, que va desitjar que també s'hi sumi el PSC.	12,10
3	160	El president del Govern creu que hi ha "alternatives", un cop tancat el termini de les coalicions.	El debat sobre la candidatura unitària semblava acabat dimarts quan ERC i el PDECat van decidir presentar-se per separat.	12,75
4	144	Acusa Enric Colet d'induir la directora de serveis a adjudicar a dit un contracte de 45.000 euros a Josep Tous.	El mateix dia Tomás va aprovar el plec de clàusules i va obrir el procediment.	12,84
5	89	Parlem amb els ciutadans que han pujat en autobusos per defensar els membres del Govern empresonats des de les seves pròpies ciutats.	Es va enamorar d'una gurbetana i va venir cap al poble", rememora Anguera, que va viure amb molt d'orgull el dia que el van nomenar conseller.	12,94

Taula B.5: Rànquing de la consulta "El partit d'esquerres va pedre les eleccions"per model STSB català

- **Model Clàssic català:**

Resultats de la consulta = 0

B.2.2. Corpus 1000 articles

Consulta: Jugar en casa

- **Model STSB castellà:**

Rànquin	Valor úni	Summary	Linia Similar	Distancia
1	475	Todos los datos y el resultado al minuto del partido entre Dax Donders - Burak Can Yilmaz del ITF Turquía F26 2019.	El jugador neerlandés Dax Donders, número 1708 de la ATP, venció por 6-1 y 6-1 al jugador turco Burak Can Yilmaz en la ronda previa de calificación del torneo de Antalya.	10,88
2	992	Todos los datos y el resultado al minuto del partido entre Hugo Cazaban - Robert Ziganshin del ITF France F15 2019.	El torneo de Bagneres-De-Bigorre (ITF France F15) cuenta con una fase de acceso previa en la que los jugadores con menor ránking tienen que obtener los mayores puntos posibles para lograr clasificarse y participar en el torneo oficial.	11,00
3	742	Todos los datos y el resultado al minuto del partido entre Alen Avidzba - Karl Friberg del ITF France F16 2019.	El jugador ruso jugará en los cuartos de final de la competición contra el tenista frances Dan Added, núCon esta victoria, el jugador logra sumar nuevos puntos a su ranking para participar en el torneo de Sajur.mero 449.	11,04
4	952	Todos los datos y el resultado al minuto del partido entre Eduardo Cohen - Mattan Kermish del ITF Israel F6 2019.	Con esta victoria, el jugador logra sumar nuevos puntos a su ranking para participar en el torneo de Sajur.	11,07
5	493	Todos los datos y el resultado al minuto del partido entre Mats Hermans - Michalis Sakellaridis del ITF Turquía F26 2019.	En esta fase en concreto participan 48 jugadores.	11,08

Taula B.6: Rànquing de la consulta "Jugar en casa"per model STSB castellà

- **Model Word2Vec castellà:**

Rànquin	Valor úni	Summary	Linia Similar	Distancia
1	201	Todos los datos y el resultado al minuto del partido entre Karolina Pliskova - Margarita Gasparyan del WTA Eastbourne 2019.	Gasparyan consiguió romper el saque en una ocasión, mientras que la jugadora checa lo hizo en 4 ocasiones.	17,66
2	272	Todos los datos y el resultado al minuto del partido entre David Klier y Bruno Sant'anna - Fabrizio Ornago y Elmar Ejupovic del ITF France F13 2019.	La pareja derrotada consiguió romper el saque en una ocasión, mientras que los vencedores lo lograron en 4 ocasiones.	17,69
3	182	Todos los datos y el resultado al minuto del partido entre Jason Kubler - Henri Laaksonen del Wimbledon Men Singles 2019.	Durante esta fase en concreto, participan 128 jugadores.	18,12
4	247	Todos los datos y el resultado al minuto del partido entre Brayden Schnur - Maxime Janvier del Wimbledon Men Singles 2019.	Durante esta fase en concreto, participan 128 jugadores.	18,12
5	388	entre Kristian Lozan - Mirko Cutuli del ITF Morocco F2 2019.	Cutuli logró romper el saque a su rival en una ocasión, mientras que Lozan lo consiguió en 4 ocasiones.	18,24

Taula B.7: Rànquing de la consulta "Jugar en casa"per model Word2Vec castellà

- **Model Clàssic castellà:**

Resultats de la consulta = 0

Consulta: El partit d'esquerres va pedre les eleccions

- **Model STSB català:**

Rànquin	Valor úni	Summary	Linia Similar	Distancia
1	875	Sánchez lamenta que el líder de Podem ho vegi com una "derrota" quan és un "triomf de la democràcia".	El president del govern espanyol en funcions, Pedro Sánchez, s'ha mostrat sorprès per les crítiques d'Iglesias i s'ha preguntat "com és possible" que partits d'esquerres critiquin que el seu executiu compleixi la promesa de treure Franco del Valle de los Caídos.	10,5051
2	613	Entre l'octubre del 34 i el febrer del 36 les institucions catalanes havien estat intervingudes. No va caldre convocar eleccions catalanes per restituir-les.	La complicitat amb l'esquerra espanyola feia preveure que en les següents eleccions generals exigirien el seu alliberament.	10,5524
3	391	Torra insta el PSOE i Podem a crear una taula de diàleg on es pugui parlar d'autodeterminació.	Amb el resultat a la mà, els independentistes -com a mínim Esquerra- són imprescindibles perquè la legislatura a Espanya pugui començar.	10,6105
4	541	El candidat de JxCat al Senat fa un circuit propi durant la campanya electoral.	En les passades eleccions municipals, en aquest barri, va guanyar Esquerra.	10,7348
5	630	El vicepresident del Govern acusa Sánchez de dir "mentides" per intentar dividir l'independentisme.	penjant les trucades que li arribin des del Palau de la Generalitat.	10,786700

Taula B.8: Rànquing de la consulta "El partit d'esquerres va pedre les eleccions"per model STSB català

- **Model Word2Vec català:**

Rànquin	Valor úni	Summary	Linia Similar	Distancia
1	919	El SEM calcula que hi ha 600 ferits, 10 dels quals hospitalitzats, entre ells un policia molt greu.	El Sistema d'Emergències Mèdiques va atendre aquest dilluns tres persones.	10,8768
2	443	Albert Rivera va cometre el pitjor error en un polític: no saber aprofitar el seu moment.	El projecte polític que significa Ciutadans va quedar aquest diumenge seriosament tocat.	11,0039
3	502	El PSOE va guanyar aquelles eleccions per majoria absoluta i Felipe González va començar el seu periple de catorze anys com a president.	El PSOE va guanyar aquelles eleccions per majoria absoluta i Felipe González va començar el seu periple de catorze anys com a president, precisament el que volien evitar els colpistes.	11,1322
4	488	Dolents, normals o bons? Així es llegiran els resultats d'aquest diumenge a les seus dels partits.	El 28 d'abril va guanyar unes generals per primera vegada.	11,2750
5	76	La defensa de Puigdemont estudia accions legals per les declaracions de l'alt representant de la UE.	declaracions de l'exministre espanyol va l'equip legal de Carles Puigdemont.	11,2826

Taula B.9: Rànquing de la consulta "El partit d'esquerres va pedre les eleccions"per model STSB català

- **Model Clàssic català:**

Resultats de la consulta = 0

B.2.3. Corpus 10000 articles

Consulta: La economía crece en este último período

- **Model STSB castellà:**

Rànquin	Valor úni	Summary	Linia Similar	Distancia
1	6032	El presidente de la Xunta, Alberto Núñez Feijóo, y representantes de los trabajadores de Alcoa San Cibrao han alertado este lunes de que las promesas que el Gobierno central "incumplió" abocan a la compañía al cierre si no se toman medidas urgentes.	En este momento el Gobierno de España no tiene ni política industrial ni política energética, y lo que es más grave, la política energética e industrial lo que están es destruyendo el empleo, agrega.	9,7799
2	3992	El secretario de Estado de Medio Ambiente en funciones, Hugo Morán, ha reiterado este jueves el compromiso de España para que la Unión Europea (UE) alcance la neutralidad climática en 2050, según informa el Ministerio para la Transición Ecológica (MITECO).	De este modo, se contribuye a paliar efectos de escasez, a facilitar la adaptación al cambio climático y se avanza en la economía circular.	9,9961
3	2261	Mallorca y presidente del Instituto Mallorquín de Asuntos Sociales (IMAS), Javier de Juan, y la directora insular del área de gente mayor del IMAS, Sofia Alonso, han presidido el 'VI Encuentro del voluntariado', que ha organizado la institución	La temática de este encuentro ha sido 'Voluntariado y demencias: protección, atención y formación'.	10,2288
4	2802	La Consejería de Transición Ecológica, Lucha contra el Cambio Climático y Planificación Territorial del Gobierno de Canarias contará por primera vez con una agenda propia en la 'Cumbre del Clima', que actualmente se está desarrollando en Madrid.	Por este motivo, el Ejecutivo regional ha programado una serie de proyecciones en la Delegación del Gobierno de Canarias relacionadas con los efectos y las previsiones del cambio climático en todo el planeta.	10,2549
5	5896	El Servicio de Protección de la Naturaleza (Seprona) en Gran Canaria ha investigado el 20 de diciembre a un hombre que responde a las iniciales de C.A.F., de 18 años y vecino de Santa Lucía de Tirajana, como presunto autor de un delito de maltrato animal al supuestamente ocasionar la muerte a golpes a un gato en la localidad de Sardina del Sur.	De este modo, se pudo localizar e investigar a C.A.F.	10,3514

Taula B.10: Rànquing de la consulta L·La economía crece en este último período"per model STSB castellà

- **Model Word2Vec castellà:**

Rànquin	Valor úni	Summary	Linia Similar	Distancia
1	3957	Albacete, con 15 euros, y Ciudad Real, con 20 euros, son las ciudades con la tarjeta mensual de transporte más barata, según un estudio anual realizado por Facua-Consumidores en Acción.	La diferencia en este caso alcanza el 218%.	10,7962
2	4347	EMECA, la asociación europea que agrupa a los principales propietarios de recintos feriales de Europa, ha aceptado a Bilbao Exhibition Centre (BEC) como uno de sus miembros, lo que le permitirá compartir información con los 24 principales espacios del 'Viejo Continente', según ha informado el recinto ferial vizcaíno.	La Asociación EMECA se fundó en 1992 para centrar la atención en el impacto económico que la industria ferial aporta en Europa.	11,5819
3	4033	El Partido Riojano pide a Renfe que "rebaje" los precios de los billetes de tren en La Rioja "igual que los nuevos billetes 'low cost' entre Madrid y Barcelona que ha anunciado recientemente el ministro de Fomento, José Luis Ábalos".	Renfe con esta decisión avala que La Rioja continúe en el furgón de cola en materia ferroviaria".	11,6136
4	5411	(CSIF) alerta sobre "la saturación del servicio de urgencias del hospital San Pedro de Logroño, donde cada uno de los boxes ha tenido que acoger a dos pacientes, mientras las 18 camas de Preingresos se hallaban totalmente ocupadas".	En La Rioja, la gripe alcanza ya el nivel de epidemia.	12,0073
5	9766	Enero fue en La Rioja un mes normal tanto en temperaturas como en precipitaciones, según los primeros datos del Avance Climatológico de la comunidad, que ha dado a conocer la Agencia Estatal de Meteorología (Aemet).	De acuerdo con este parte, en términos generales, el mes de enero en La Rioja tuvo un carácter normal en general en toda la comunidad respecto a las temperaturas, con tendencia a cálido en la sierra de La Rioja Alta y a frío en el valle de La Rioja Baja.	12,1593

Taula B.11: Rànquing de la consulta L·La economía crece en este último período"per model Word2Vec castellà

- **Model Clàssic castellà:**

Resultats de la consulta = 0

Consulta: L'economia creix en aquest darrer període.

- **Model STSB català:**

Rànquin	Valor úni	Summary	Linia Similar	Distancia
1	6525	Entrevista de la directora de l'ARA a la consellera de la Presidència.	L'economia està creixent.	7,6417
2	9588	L'analista polític sosté que Europa ha d'afrontar els seus reptes amb unió.	L'economia d'Espanya està creixent", ha dit en la conferència La tornada de la geopolítica a la Reunió del Cercle d'Economia que se celebra a Sitges.	9,4926
3	9488	Sánchez tria un executiu extens, amb més dones que homes i amb independents.	El líder d'ERC empresonat creu que Europa es troba en una cruïlla entre l'auge de l'extrema dreta, que vol minvar la força de les institucions comunitàries, i la d'un federalisme europeu d'esquerres, "socialment ambiciós, internacionalista i que defensa una economia eficient,	9,6346
4	5784	En un article a 'Euractiv' el líder d'ERC defensa una Catalunya lliure dins d'una Europa federal.	Sobre l'escorcoll a Economia, l'agent que va recopilar les imatges per a l'atestat ha qualificat la situació de "setge".	9,7452
5	4573	Els agents de la Guàrdia Civil intenten apuntalar el relat de la violència.	És una aposta estratègica que requereix d'una "convergència" molt gran no només dels partits polítics, sinó també de la societat civil, de sectors com les lletres, l'economia, la cultura i la judicatura, entre d'altres.	9,7707

Taula B.12: Rànquing de la consulta L·L'economia creix en aquest darrer període."per model STSB català

- **Model Word2Vec català:**

Rànquin	Valor úni	Summary	Linia Similar	Distancia
1	8402	L'ANC destaca el "bon ritme" en la venda de samarretes i espera almenys un milió d'assistents.	Romeu subratlla que les últimes setmanes d'agost i sobretot l'inici de setembre serà el període en què el degoteig d'inscripcions anirà en augment.	8,1442
2	2577	Els cupaires, reticents en el passat a formar part del cartipàs, són en dotze executius més que el 2015.	"Aquest cop hem sigut menys dogmàtics per l'emergència social en què viu Figueres, en el top 3 de fracàs escolar i amb un 60% de població en risc d'exclusió social",	8,2221
3	9432	La militància dels joves es va disparar de manera generalitzada des d'un mes abans del referèndum.	En el cas d'Arran, l'organització juvenil de l'esquerra independentista, també hi ha hagut un notable creixement de militància en els últims mesos, si bé en l'actual context "repressiu" prefereix no fer públiques les dades detallades de cada mes.	8,6640
4	3759	Un 55% dels ciutadans creuen que la independència dels jutges és pobra, i ho atribueixen a pressions econòmiques o polítiques.	Justament l'any en què viu un dels processos judicials més importants de la seva història, Espanya empitjora respecte a l'any anterior: si l'any passat un 49% creien que la independència judicial era molt dolenta o força dolenta, aquest 2019 se situa en un 55%.	8,7474
5	3881	El PP i Cs la volen "ordenar" i la resta de partits parlen de drets.	Rivera presumeix d'emmirallar-se en el Canadà en matèria d'immigració, ja que la formació taronja promourà un sistema de "visat per punts" -reproduint el del país nord-americà- basat en "barems de puntuació objectius".	8,9412

Taula B.13: Rànquing de la consulta L·L'economia creix en aquest darrer període."per model STSB català

- **Model Clàssic català:**

Resultats de la consulta = 0

B.2.4. Corpus 90000/100000 articles

Consulta: El juez decreto fallo en la sentencia

- **Model STSB castellà:**

Ràngui	Valor úni	Summary	Linia Similar	Distancia
1	74147	La acusación de que el primer ministro podría haber mentido a la Reina sobre las razones reales de la prórroga fue alimentada por el juicio del tribunal de sesión escocés que argumenta que «el gobierno no dijo la verdad».	En su resumen de la sentencia , el tribunal dijo.	8,8898
2	91868	El Fiscal pide 1 año de prisión por lesionar a un directivo.	El juicio quedó visto para sentencia.	9,2658
3	21236	La Fiscalía ha pedido este jueves retirar la patria potestad al padre acusado de abusar sexualmente de su hija de cuatro años en Palma, hechos por los que también solicita una condena de 12 años de	El juicio ha quedado visto para sentencia.	9,2968
4	13913	Dos empresarios han aceptado este jueves un año de cárcel cada uno por un delito de estafa, al reconocer que abonaron el coste de unas obras con pagarés de empresas que, en realidad, eran insolventes, por lo que sabían que nunca podrían	Finalmente, el juicio quedó visto para sentencia.	9,3178
5	42731	El pleno de las Cortes de Aragón ha aprobado por unanimidad una proposición no de ley (PNL) del Partido Popular -enmendada por PSOE, Podemos-Equo, Chunta Aragonesista y Partido Aragonés- para apoyar las reclamaciones de devolución de los bienes de las parroquias aragonesas que se	SENTENCIAS JUDICIALE	9,7215

Taula B.14: Rànquing de la consulta "El juez decreto fallo en la sentencia"per model STSB castellà

- **Model Word2Vec castellà:**

Ràngui	Valor úni	Summary	Linia Similar	Distancia
1	83102	La competencia por los puestos de trabajo y la inmigración tensan en la actualidad una relación salpicada de conflictos.	El Tribunal Supremo tumbaría la medida en 2016.	12,5000
2	11031	Agresión sexual y violación no son lo mismo ni se castigan con la misma pena. El caso Diana Quer y el 'caso Lanza' cuestionan los jurados populares: pros y contras. Juan Carlos Quer: "Mi hija ha estado 500 días en un pozo diciendo no es no".	El juez Ángel Pantín dictará la sentencia.	12,6554
3	71247	El antiguo fiscal del caso AMIA presentó contra la entronces presidenta argentina y varios miembros de su Gobierno por el Memorándum de Entendimiento suscrito en 2013 por Argentina e Irán para destrabar las investigaciones.	El también fiscal Germán Moldes apeló el fallo de Rafecas ante la Cámara Federal, que ratificó la decisión.	12,8116
4	7271	El joven granadino de 24 años entró voluntariamente sobre las 19.20 h en el centro penitenciario de Albolote (Granada).Su madre, visiblemente afectada, ha considerado una "injusticia" el ingreso en la cárcel de su hijo: "Me lo van a matar ahí dentro".Su abogado presentó el pasado viernes un recurso ante la Audiencia Nacional, que finalmente no ha prosperado.A partir de ahora, "la lucha se centrará en que Alejandro permanezca en la cárcel lo menos posible", sostiene el letrado. El PSOE granadino ha defendido que si el final de la pena es la reinserción, él "ha cumplido plenamente este objetivo".Alejandro reconoció este domingo que se sentía "acojonado" y "hundido".	El fallo fue recurrido ante el Tribunal Supremo, que confirmó la pena.	13,3237
5	44543	El Tribunal Supremo ha anulado la sentencia del Tribunal Superior de Justicia de Cantabria (TSJC) que absolvió al guardia civil de Tráfico condenado por la Audiencia provincial por beneficiar a la empresa de transportes de su esposa, en marzo de 2018, después de que un tribunal de jurado le declarara culpable de un delito de actividades prohibidas.	El agente recurrió la sentencia en apelación ante la Sala Civil y Penal del TSJC, que en junio de ese mismo año le absolvió.	13,3819

Taula B.15: Rànquing de la consulta "El juez decreto fallo en la sentencia"per model Word2Vec castellà

- **Model Clàssic castellà:** Paraules cercades: Juez decretar sentencia

Documents: 97912
Resultats de la consulta = 1

Consulta: Només 100 vots a favor de la nova llei

- **Model STSB català:**

Rànquin	Valor úni	Summary	Linia Similar	Distancia
1	32136	ERC compara els dos partits amb Cánovas i Sagasta, mentre que IU titlla l'acord de "populisme punitiu" per considerar que "tots som terroristes".	Només si aquestes esmenes prosperen acabarà votant a favor de la proposició de llei.	10,1606
2	28403	El líder del sector renovador dels democristians admet a RAC1 que ara el debat serà per qui es queda les sigles.	El 'sí' –la proposta de Duran– va guanyar per 95 vots.	10,5740
3	44899	El partit de Rivera defensa retirar la sanitat als immigrants, les escoles d'elit i els lobis.	Al final només C's hi va votar a favor.	10,5757
4	11927	El president del Parlament es reuneix aquest dimecres amb Puigdemont i els diputats a Brussel·les.	Aquesta vegada només seria necessària la majoria simple dels vots: més a favor que en contra.	10,5827
5	16290	Governació lamenta que al registre impulsat per Exteriors només hi ha 5.000 persones apuntades.	A més, ella mateixa –recorden– va destacar quan va presentar la llei del vot electrònic que era la que havia de servir per resoldre les traves que el Govern posava als catalans que volien votar des de l'estranger.	10,6466

Taula B.16: Rànquing de la consulta "Només 100 vots a favor de la nova llei."per model STSB català

- **Model Word2Vec català:**

Rànquin	Valor úni	Summary	Linia Similar	Distancia
1	17916	La presidenta del Parlament ha explicat que, de moment, la CUP no formarà part de la mesa del Parlament, tot i que ha agraït la seva voluntat d'entrar-hi.	Precisament, aquest dimecres la CUP donarà suport a través de dos vots favorables a la llei de pressupostos.	9,5483
2	85444	Els professors es podran quedar a treballar al centre i els alumnes també conserven la plaça.	En efecte, a l'article 45 de la LEC ja s'estableix la possibilitat d'integrar centres a la xarxa de la Generalitat per mitjà d'una llei.	9,9692
3	53862	A banda dels 1.750 milions del fons de competitivitat, el Govern reclama a Madrid la transferència corresponent a la disposició addicional tercera de l'Estatut.	El govern espanyol deu a la Generalitat els 750 milions d'euros corresponents a la disposició addicional tercera de l'Estatut.	10,0213
4	66385	L'ex primer ministre britànic demana la creació d'un moviment "políticament transversal" per combatre el resultat d'un referèndum basat en el "desconeixement" de les conseqüències de la	La setmana passada, Jeremy Corbyn va ordenar als seus diputats de votar a favor de la llei de l'article 50.	10,1689
5	13668	Boya no descarta "cap escenari" i demana al Govern que prengui acords per impulsar la República.	La CUP no vol donar normalitat democràtica a les eleccions convocades per Mariano Rajoy el 21 de desembre a través de l'article 155 de la Constitució.	10,4632

Taula B.17: Rànquing de la consulta "Només 100 vots a favor de la nova llei."per model STSB català

- **Model Clàssic català:** Paraules cercades: 100 vots favor llei

Documents: 71182 45781; 94287; 63929; 1760; 96019; 66497; 25127; 8608; 65212; 12511; 58708; 23724; 60863; 68460; 77168; 71691; 19452; 75636; 77025; 629; 93516; 25556; 7078; 10260;
Resultats de la consulta = 25

B.3 Resultats Consultes finals

B.3.1. Consulta 1: El president va ser expulsat del congrés

Rànquing	Descripció Resultat
1	Valor únic: 12625 Summary: Pastor torna a fer un discurs en clau d'humor als premis de l'APP. Línia similar: ", relata la presidenta del Congrés. Valor de la distància entre les frases: 8.941197
2	Valor únic: 21708 Summary: Santi Vila és conseller de Cultura i aspira a presidir el consell nacional del Partit Demòcrata català (PDC) fent un discurs crític amb el congrés fundacional de la formació que substitueix CDC. Línia similar: El congrés del PDC va ser un èxit? Valor de la distància entre les frases: 9.241630
3	Valor únic: 17940 Summary: El testimoni més emotiu va ser el del fill del president, Josep Tarradellas i Macià. "El meu pare sempre va estar convençut que tornaria a Catalunya com a president, cada hora, cada minut, cada segon". Línia similar: El testimoni més emotiu va ser el del fill del president, Josep Tarradellas i Macià. Valor de la distància entre les frases: 9.704752
4	Valor unic: 21707 Summary: Homs optarà a president en un gest que la resta atribueixen a un pacte per escollir Pastor i tenir grup. Linia similar: El president del Congrés es tria nominalment i en secret. Valor de la distancia entre las frases: 9.741816
5	Valor únic: 19838 Summary: Un infart acaba amb la vida de l'exalcaldessa de València, que s'enfrontava a acusacions de corrupció. Línia similar: El Congrés es va congelar. Valor de la distància entre les frases: 9.848175
6	Valor únic: 8261 Summary: Endureix el discurs contra Maduro però evita parlar de dictadura com el PP. Línia similar: El va acompanyar, en representació de l'Estat, la presidenta del Congrés, l'exministra del PP Ana Pastor. Valor de la distància entre les frases: 9.896319
7	Valor únic: 19838 Summary: Un infart acaba amb la vida de l'exalcaldessa de València, que s'enfrontava a acusacions de corrupció. Línia similar: Un minut de silenci ordenat per la presidenta del Congrés, Ana Pastor, va provocar la polèmica. Valor de la distància entre les frases: 9.902490
8	Valor únic: 37228 Summary: PP, PSOE i UPyD es neguen a rebre el ministre principal de la colònia britànica. Línia similar: El desembre del 2013 va ser el portaveu dels republicans al Congrés, Alfred Bosch, qui es va entrevistar amb Picardo durant la visita a la zona. Valor de la distància entre les frases: 9.911806
9	Valor únic: 11181 Summary: La decisió de Torrent sobre la investidura és d'obligat compliment per al jutge. Línia similar: El president del Congrés, o del Parlament, és l'única autoritat que ordena com es desenvolupa l'acte. Valor de la distància entre les frases: 9.957330
10	Valor únic: 19565 Summary: El conseller d'Interior, Jordi Jané, es mostra convençut que la policia catalana estarà "al costat del poble català, de les seves institucions i de la legalitat vigent a Catalunya". Línia similar: El portaveu d' ERC al Congrés, Gabriel Rufián, troba "dramàticament normalll'la decisió del Tribunal Constitucional de suspendre cautelarment el pla per al referèndum. Valor de la distància entre les frases: 9.975480

Taula B.18: Resultats consulta 1 model STSB.

Rànquing	**Descripció Resultat**
1	Valor únic: 2010 Summary: Enrique López va dimitir com a magistrat del TC perquè va donar positiu en un control d'alcoholèmia. Línia similar: López va ser nomenat membre del Tribunal Constitucional el juny del 2013. Valor de la distància entre les frases: 8.797681
2	Valor únic: 48768 Summary: El ministre confessa que vol que l'escola catalana fomenti l'amor a Espanya en els seus alumnes. Línia similar: El final del govern de José María Aznar va ser dramàtic. Valor de la distància entre les frases: 9.191519
3	Valor unic: 28429 Summary: Els quatre grups lamenten els recents casos de corrupció i les retallades i critiquen el Govern per no complir les resolucions del Parlament. Linia similar: El primer va ser l'octubre del 1985. Valor de la distancia entre las frases: 9.230586
4	Valor únic: 32358 Summary: El magistrat desafia la Fiscalia i obre el text a l'esmena ciutadana. Línia similar: El jutge va exercir de protagonista del debat constituent. Valor de la distància entre les frases: 9.660270
5	Valor únic: 64848 Summary: Trump aposta per adoptar mesures unilaterals contra el règim de Kim Jong-un. Línia similar: El mutisme del gegant asiàtic va ser total. Valor de la distància entre les frases: 9.792452
6	Valor únic: 64000 Summary: Dimiteixen perquè el líder, Henry Bolton, es nega a deixar el càrrec. Línia similar: Bolton va ser triat líder del UKIP el setembre passat. Valor de la distància entre les frases: 9.813232
7	Valor únic: 17720 Summary: És diplomàtic i va ser delegat del 2013 al 2015, quan va dimitir després del divorci de CiU. Línia similar: El 2015 va dimitir arran del divorci de CiU. Valor de la distància entre les frases: 9.844411
8	Valor únic: 40632 Summary: L'excap de planificació i l'exinterventora múnicipal van denunciar l'alcaldessa i el gerent per assetjament laboral i injúries quan els van rellevar dels càrrecs. Línia similar: El comiat de Gadea es va fer el 17 novembre del 2009, un cop Parlon va prendre possessió del càrrec. Valor de la distància entre les frases: 9.889872
9	Valor únic: 63922 Summary: Els diputats aproven traslladar-se a una altra seu durant les obres de restauració, que duraran almenys 6 anys. Línia similar: El Parlament actual va ser reconstruït després del foc que el va destruir l'any 1834. Valor de la distància entre les frases: 9.973517
10	Valor únic: 9511 Summary: L'ambaixador d'Espanya davant d'aquest organisme ocuparà el nou càrrec al setembre. Línia similar: El desembre del 2011 va ser nomenat cap de gabinet del president del govern espanyol, Mariano Rajoy. Valor de la distància entre les frases: 10.000707

Taula B.19: Resultats consulta 1 model Word2Vec.

B.3.2. Consulta 2: Només 100 vots a favor de la nova llei

Rànquing	Descripció Resultat
1	Valor únic: 32136 Summary: ERC compara els dos partits amb Cánovas i Sagasta, mentre que IU titlla l'acord de "populisme punitiu"per considerar que "tots som terroristes". Línia similar: Només si aquestes esmenes prosperen acabarà votant a favor de la proposició de llei. Valor de la distància entre les frases: 10.160643
2	Valor únic: 28403 Summary: El líder del sector renovador dels democristians admet a RAC1 que ara el debat serà per qui es queda les sigles. Línia similar: El 'sí' –la proposta de Duran– va guanyar per 95 vots. Valor de la distància entre les frases: 10.573980
3	Valor únic: 44899 Summary: El partit de Rivera defensa retirar la sanitat als immigrants, les escoles d'elit i els lobis. Línia similar: Al final només C's hi va votar a favor. Valor de la distància entre les frases: 10.575716
4	Valor únic: 11927 Summary: El president del Parlament es reuneix aquest dimecres amb Puigdemont i els diputats a Brussel·les. Línia similar: Aquesta vegada només seria necessària la majoria simple dels vots: més a favor que en contra. Valor de la distància entre les frases: 10.582693
5	Valor únic: 16290 Summary: Governació lamenta que al registre impulsat per Exteriors només hi ha 5.000 persones apuntades. Línia similar: A més, ella mateixa –recorden– va destacar quan va presentar la llei del vot electrònic que era la que havia de servir per resoldre les traves que el Govern posava als catalans que volien votar des de l'estranger. Valor de la distància entre les frases: 10.646573
6	Valor únic: 28402 Summary: El secretari general d'Unió, Ramon Espadaler, defensa a Catalunya Ràdio que el partit democristià "no aposta pel trencament de la federacióÇiU, i atribueix la sortida dels consellers d'Unió del Govern a l'ültimàtum"de Convergència. El sector sobiranista plantarà cara si la direcció aposta per defensar la tercera via. Línia similar: El sí –la proposta de Duran– va guanyar per 95 vots. Valor de la distància entre les frases: 10.665938
7	Valor únic: 2892 Summary: Van den Eynde, també advocat de Romeva, carrega contra l'"exageració"de la Fiscalia. Línia similar: I la declaració de la lletrada de justícia... des de la seva fugida pel terrat fins a sentir Carme Forcadell per megafonia quan no va parlar. Valor de la distància entre les frases: 10.703554
8	Valor únic: 69205 Summary: L'únic candidat de dretes és el favorit a les presidencials, però ja avança que no tombarà Costa. Línia similar: Per aconseguir-ho, cal que superi el 50% dels vots, un llindar que ultrapassa en la mitjana de tots els sondejos. Valor de la distància entre les frases: 10.704723
9	Valor únic: 58232 Summary: Els socialistes tindran "presència en els mecanismes de negociació bilateral Estat-Generalitat". Les reformes del marc estatutari o constitucional es pactaran entre els dos partits. Línia similar: I no només aquestes, també els canvis que es vulguin operar a la Corporació Catalana de Mitjans Audiovisuals o la llei electoral, que ara els dos partits (que reuneixen els 90 diputats necessaris) podrien impulsar. Valor de la distància entre les frases: 10.712012
10	Valor únic: 77385 Summary: En un referèndum no vinculant, els illencs s'han pronunciat a favor de ser l'estat 51 dels Estats Units, i l'opció independentista ha tingut un suport de més del 5% de l'electorat. Línia similar: Amb el 86% escrutat, el 47,80% dels vots han estat per al líder del PPD, només 0,61 punts i 9.943 vots més que el governador Fortuño. Valor de la distància entre les frases: 10.775621

Taula B.20: Resultats consulta 2 model STSB.

Rànquing	Descripció Resultat
3	Valor únic: 17916 Summary: La presidenta del Parlament ha explicat que, de moment, la CUP no formarà part de la mesa del Parlament, tot i que ha agraït la seva voluntat d'entrar-hi. Línia similar: Precisament, aquest dimecres la CUP donarà suport a través de dos vots favorables a la llei de pressupostos. Valor de la distància entre les frases: 9.548260
2	Valor únic: 85444 Summary: Els professors es podran quedar a treballar al centre i els alumnes també conserven la plaça. Línia similar: En efecte, a l'article 45 de la LEC ja s'estableix la possibilitat d'integrar centres a la xarxa de la Generalitat per mitjà d'una llei. Valor de la distància entre les frases: 9.969210
3	Valor únic: 53862 Summary: A banda dels 1.750 milions del fons de competitivitat, el Govern reclama a Madrid la transferència corresponent a la disposició addicional tercera de l'Estatut. Línia similar: El govern espanyol deu a la Generalitat els 750 milions d'euros corresponents a la disposició addicional tercera de l'Estatut. Valor de la distància entre les frases: 10.021252
4	Valor únic: 66385 Summary: L'ex primer ministre britànic demana la creació d'un moviment "políticament transversal"per combatre el resultat d'un referèndum basat en el "desconeixement"de les conseqüències de la decisió. Línia similar: La setmana passada, Jeremy Corbyn va ordenar als seus diputats de votar a favor de la llei de l'article 50. Valor de la distància entre les frases: 10.168932
5	Valor únic: 13668 Summary: Boya no descarta çap escenariï demana al Govern que prengui acords per impulsar la República. Línia similar: La CUP no vol donar normalitat democràtica a les eleccions convocades per Mariano Rajoy el 21 de desembre a través de l'article 155 de la Constitució. Valor de la distància entre les frases: 10.463237
6	Valor únic: 88175 Summary: Aprova que el 2030 el 32% de l'energia consumida a Europa hagi de ser renovable. Línia similar: La normativa ve acompanyada també d'una nova governança per a la Unió Energètica que també s'ha aprovat amb 475 vots a favor i 100 en contra. Valor de la distància entre les frases: 10.487649
7	Valor únic: 76632 Summary: Aquest actor i còmic italià ha sabut agrupar un conjunt de la població italiana decebuda amb els polítics i els constants casos de corrupció. Línia similar: Grillo havia mobilitzat la població italiana per a la recollida de firmes per presentar una llei d'iniciativa popular amb l'objectiu d'impedir als condemnats de corrupció l'accés a la política. Valor de la distància entre les frases: 10.488225
8	Valor únic: 84879 Summary: La recollida selectiva arriba al 41,8% mentre creix la generació de brossa: 1,43 quilos per habitant i dia. Línia similar: Enviar una tona de residus a aquesta instal·lació costa uns 41 euros de cànon municipal a l'Ajuntament, l'any que ve la xifra s'encarirà a 47 euros i el 2025 ja valdrà 77 euros la tona. Valor de la distància entre les frases: 10.511790
9	Valor únic: 58646 Summary: El Parlament d'Edimburg aprova la llei genèrica que el faculta per poder convocar plebiscits. Línia similar: La petició l'ha fet aquest dijous a Edimburg, hores abans que la cambra autonòmica aprovés per majoria de 68 a 54 vots el tercer i últim estadi de la llei genèrica de referèndums d'Escòcia, a partir de la qual s'hauria de desenvolupar l'específica del segon plebiscit. Valor de la distància entre les frases: 10.530191
9	Valor únic: 15103 Summary: Tarragona, Santa Coloma i Barcelona: els tres actes unitaris de la campanya del 'sí'. Linia similar: Divendres a mitjanit comença formalment la campanya electoral de l'1-O. Valor de la distancia entre las frases: 10.550881

Taula B.21: Resultats consulta 2 model Word2Vec.

B.3.3. Consulta 3: L'equip local va guanyar per golejada

Rànquing	Descripció Resultat
1	Valor únic: 25026 Summary: El president de la Generalitat en funcions, en el míting final de Democràcia i Llibertat a les comarques de Tarragona, condemna els fets de Pontevedra tot i no tenir çap amistat"amb el president espanyol. Línia similar: "Vas guanyar per golejada. Valor de la distància entre les frases: 7.474182
2	Valor únic: 3357 Summary: Res millor que l'ocurrència de Messi i les ocurrències de Núria de Gispert a Twitter per revifar la popularitat de les Creus de Sant Jordi. Línia similar: La Trinca el guanya per golejada. Valor de la distància entre les frases: 7.956905
3	Valor únic: 141 Summary: L'advocat presenta el llibre on recull la seva vivència com a defensor de presos polítics. Línia similar: "Penso que vam guanyar per golejada. Valor de la distància entre les frases: 8.231167
4	Valor únic: 34144 Summary: 'El País' assegura que el president espanyol i Mas es poden convertir en "obstacles"per a qualsevol solució en les relacions Catalunya-Espanya. 'La Vanguardia' i 'El Periódico' no valoren les paraules del líder del PP sobre el 9-N. Línia similar: Malauradament, al president del Govern espanyol li falta decisió política, i al de la Generalitat li sobra gosadia i agressivitat. Valor de la distància entre les frases: 9.305518
5	Valor únic: 15501 Summary: Els pensadors de l'Estat que han explicitat el suport al referèndum es poden comptar amb els dits d'una mà. Línia similar: Gairebé ningú en l'esfera política gosa fer-ho. Valor de la distància entre les frases: 9.783986
6	Valor únic: 23592 Summary: La cita estava prevista per al 21 i 22 de maig i les candidatures a les primàries s'havien de presentar ja els dies 11 i 14 d'abril. Línia similar: Pensava que, en plenes negociacions per formar govern, ningú gosaria fer-li ombra. Valor de la distància entre les frases: 9.874027
7	Valor únic: 50487 Summary: Alguns empresaris i experts creuen que l'èxit de la selecció estatal de futbol ha generat el millor moment per rellançar la marca Espanya a l'exterior. Línia similar: Sandalio Góme. Valor de la distància entre les frases: 9.979721
8	Valor únic: 57499 Summary: Ricard Gomà, però, fuig del continuisme, i avisa que exigirà noves condicions als seus socis per obrir una nova etapa. Línia similar: Ricard Gomà afronta les properes eleccions de maig per primer cop com a cap de llista. Valor de la distància entre les frases: 10.036296
9	Valor únic: 67380 Summary: La imatge transmesa en l'enfrontament televisat ha sigut sovint decisiva. Línia similar: I que, de fet, el va ajudar a guanyar les eleccions per golejada. Valor de la distància entre les frases: 10.044290
10	Valor únic: 32638 Summary: El líder múnicipal d'ICV-EUiA a Barcelona defensa el pacte de confluència per ser d'una "rellevància inèdita sense precedents". Línia similar: Gomà creu que per "primera vegada"és possible protagonitzar un "triomf social i polític"des de valors de "revolució democràtica". Valor de la distància entre les frases: 10.164925

Taula B.22: Resultats consulta 3 model STSB.

Rànquing	Descripció Resultat
3	Valor únic: 66562 Summary: El nou president nord-americà rectifica la fredor d'Obama amb una conversa "molt amable" amb Netanyahu. Línia similar: L'israelià va ser ahir el primer líder estranger que va poder conversar amb Trump per telèfon. Valor de la distància entre les frases: 9.946684
2	Valor únic: 6881 Summary: Marc Solsona, també diputat al Parlament, ho ha anunciat en un tuit. Línia similar: Marc Solsona va ser el primer alcalde i alhora diputat que va anar a declarar per l'1-O. Valor de la distància entre les frases: 9.976754
3	Valor únic: 16533 Summary: La formació ha canviat dues vegades de portaveu al Parlament valencià en tan sols dos anys. Línia similar: Qui primer va abandonar aquesta responsabilitat va ser Carolina Punset, que va canviar de Parlament i es va traslladar a l'hemicicle europeu per substituir Joan Carles Girauta. Valor de la distància entre les frases: 10.023018
4	Valor únic: 18430 Summary: El partit lila, reforçat després de l'assemblea de Madrid, reclama tenir-hi més representació. Línia similar: Podem Catalunya va néixer l'any 2014 i va estar dirigit durant mesos per una executiva provisional. Valor de la distància entre les frases: 10.218748
5	Valor únic: 78712 Summary: Quan falta una setmana perquè Eurovisió arribi a l'Azerbaidjan, molts dels seus ciutadans viuen allunyats del glamur kitsch del festival. El govern ha empresonat dissidents i ha expulsat famílies de casa seva. Línia similar: Ismailova va entendre l'amenaça però va decidir seguir endavant amb una investigació que va publicar la setmana passada. Valor de la distància entre les frases: 10.428913
6	Valor únic: 9831 Summary: També ha estat imputat l'expresident de les Corts i exdirector de la policia Juan Cotino. Línia similar: Qui també va reaccionar va ser Telefónica, que va suspendre "amb caràcter immediat" la relació laboral amb l'expresident valencià. Valor de la distància entre les frases: 10.506383
7	Valor únic: 60972 Summary: Un estudi recent posa de manifest que va causar un refredament del planeta. Línia similar: L'efecte d'aquesta reforestació va ser considerable: va eliminar tal quantitat de CO que va refredar el planeta. Valor de la distància entre les frases: 10.525112
8	Valor únic: 26176 Summary: Projecten l'escut i la bandera 'rojigualda' en una pantalla gegant a Santa Coloma. Línia similar: També va reivindicar que Santa Coloma "va ser la primera ciutat que va posar en marxa un projecte d'immersió lingüística". Valor de la distància entre les frases: 10.587693
9	Valor únic: 11759 Summary: Els republicans esperen una última proposta que la llista del president afirma que ja havien acceptat. Línia similar: Dimarts Torrent va ajornar el ple d'investidura i ahir ERC va començar l'ofensiva per deixar clara la seva posició. Valor de la distància entre les frases: 10.648343
10	Valor únic: 5313 Summary: 45.000 persones es manifesten amb el PP, Cs i Vox contra la política del diàleg i demanen eleccions. Línia similar: L'únic que va intentar sortir d'aquest marc va ser el candidat a l'alcaldia de Barcelona, Manuel Valls, que va defensar una inexistent transversalitat. Valor de la distància entre les frases: 10.696311

Taula B.23: Resultats consulta 3 model Word2Vec.

B.3.4. Consulta 4: Nova llei aprovada al congrés

Rànquing	Descripció Resultat
1	Valor únic: 37843 Summary: La norma passa amb 299 vots a favor, 19 en contra i 23 abstencions. Dos diputats socialistes han trencat la disciplina de vot i Amaiur ha marxat de l'hemicicle. Línia similar: El Congrés ha aprovat per una aclaparadora majoria la llei orgànica que contempla l'abdicació del rei Joan Carles. Valor de la distància entre les frases: 9.335949
2	Valor únic: 28764 Summary: El líder dels socialistes catalans vol que l'exministra sigui la cap de cartell del partit, tal com va fer el 2011. Línia similar: Necessitem cap de llista al Congrés per Barcelona. Valor de la distància entre les frases: 9.824880
3	Valor únic: 32856 Summary: Aquest criteri obligarà a revisar les excarceracions dictades per l'Audiència Nacional. Línia similar: El Congrés va aprovar al novembre una llei orgànica per incloure la norma comunitària en la legislació espanyola. Valor de la distància entre les frases: 9.952205
4	Valor únic: 1337 Summary: El conseller diu que no existeix çap dada objectiva"que justifiqui una possible aplicació de la llei de seguretat nacional. Línia similar: La Llei de Seguretat Nacional va ser aprovada pel Congrés el 2015. Valor de la distància entre les frases: 10.163414
5	Valor únic: 27057 Summary: De Guindos converteix el ple per aprovar el tercer rescat en un atac als "miratges dels populismes". Línia similar: Amb tot, la resolució del govern espanyol va ser aprovada amb una majoria més que àmplia al Congrés. Valor de la distància entre les frases: 10.295118
6	Valor únic: 28764 Summary: El líder dels socialistes catalans vol que l'exministra sigui la cap de cartell del partit, tal com va fer el 2011. Línia similar: Necessitem cap de llista al Congrés per BCN. Valor de la distància entre les frases: 10.367120
7	Valor únic: 37845 Summary: El portaveu de CiU al Congrés defensa l'abstenció de CiU en la llei d'abdicació i respon a Rajoy que són ells els que fan "política petita". Línia similar: Duran i Lleida avui al Congrés / EF Duran i Lleida avui al Congrés / EF. Valor de la distància entre les frases: 10.539861
8	Valor únic: 47379 Summary: Torres-Dulce adverteix a la Ser que les consultes s'han de fer sempre dins el marc de la Constitució. Línia similar: Üna llei elaborada al Congrés s'ha de complir encara que no agradi", ha afegit Torres-Dulce. Valor de la distància entre les frases: 10.544841
9	Valor únic: 21766 Summary: El líder del PP crida a la "reflexió", també personal, i el del PSOE li aconsella ser investit pel PDC. Línia similar: Per al segon intentarà sobreviure per aprovar pressupostos i lleis amb C's, CDC, el PNB i els canaris, que completen la majoria de centredreta al Congrés. Valor de la distància entre les frases: 10.555518
10	Valor únic: 38951 Summary: El líder d'Unió afirma que, si eventualment calgués, el text l'hauria de redactar "el Parlament i no persones privades". Línia similar: Duran i Lleida, al Congrés / EF Duran i Lleida, al Congrés / EF. Valor de la distància entre les frases: 10.584393

Taula B.24: Resultats consulta 4 model STSB.

Rànquing	Descripció Resultat
1	Valor únic: 34878 Summary: Gauden Villas ha dit en una entrevista que a Perpinyà "el sobiranisme català és residual". Línia similar: Entrevista al cònsol general d'Espanya a Perpinyà publicada al diari d'Alacant 'Información Entrevista al cònsol general d'Espanya a Perpinyà publicada al diari d'Alacant 'Información. Valor de la distància entre les frases: 12.045829
2	Valor únic: 34878 Summary: Gauden Villas ha dit en una entrevista que a Perpinyà "el sobiranisme català és residual". Línia similar: Entrevista al cònsol general d'Espanya a Perpinyà publicada al diari d'Alacant 'Información Entrevista al cònsol general d'Espanya a Perpinyà publicada al diari d'Alacant 'Información. Valor de la distància entre les frases: 12.045829
3	Valor únic: 9614 Summary: El president s'obre a parlar-ho tot però afirma que ha de ser el poble de Catalunya qui decideixi. Línia similar: Retreuen així al PP la impugnació de l'Estatut de l'any 2006 al Constitucional. Valor de la distància entre les frases: 12.133776
4	Valor únic: 15538 Summary: Trevín censura la nova estratègia de Sánchez, que el va rellevar com a portaveu d'interior i seguretat nacional pel seu suport a la campanya de Díaz. Línia similar: Nova baixa al grup socialista al Congrés per discrepàncies amb Pedro Sánchez. Valor de la distància entre les frases: 12.578547
5	Valor únic: 61233 Summary: Crearan una llista negra de manifestants i multaran els encaputxats i les protestes no autoritzades. Línia similar: La nova llei es debatrà al febrer a l'Assemblea Nacional. Valor de la distància entre les frases: 12.662225
6	Valor únic: 17975 Summary: En declaracions a l'ACN, la consellera Borràs afirma que el Govern "està tranquil"en aquest tema i veu en l'actitud de l'Estat ün atac a les competències de Catalunya". Línia similar: La segona llei contra els desnonaments, aprovada per unanimitat al mes de desembre al Parlament, podria acabar al Tribunal Constitucional (TC). Valor de la distància entre les frases: 12.680275
7	Valor únic: 51937 Summary: La líder del PP català rebat així les paraules d'Esperanza Aguirre, que ahir apostava en una trobada amb Rajoy perquè les autonomies retornessin les competències a l'administració central amb l'argument que els suposaria un estalvi de 48.000 milions d'euros. Línia similar: Alícia Sánchez-Camacho al Parlament de Cataluny Alícia Sánchez-Camacho al Parlament de Cataluny. Valor de la distància entre les frases: 12.72542
8	Valor únic: 60413 Summary: El president vol blindar el seu lideratge després de les revoltes d'Algèria i el Sudan. Línia similar: El Parlament egipci aprova una reforma que podria mantenir Al-Sissi al poder fins al 203. Valor de la distància entre les frases: 12.763490
9	Valor únic: 7228 Summary: Dona cinc dies a la Fiscalia i l'Advocacia de l'Estat i a Vox perquè presentin els escrits d'acusació. Línia similar: Comença el compte enrere per al judici oral contra l'independentisme al Suprem, previst per al gener. Valor de la distància entre les frases: 12.786393
10	Valor únic: 10583 Summary: Esgrimeix la resolució de l'ONU en què insta Espanya a garantir els seus drets polítics. Línia similar: L'advocat de Jordi Sànchez, Jordi Pina, ha registrat aquest dimarts al matí al Tribunal Suprem una nova petició al jutge Pablo Llarena perquè el seu client pugui assistir al ple d'investidura de divendres a les 10 h al Parlament. Valor de la distància entre les frases: 12.871646

Taula B.25: Resultats consulta 4 model Word2Vec.

B.3.5. Consulta 5: L'economia creix en aquest darrer període

Rànquing	Descripció Resultat
1	Valor únic: 6525 Summary: Entrevista de la directora de l'ARA a la consellera de la Presidència. Línia similar: L'economia està creixent. Valor de la distància entre les frases: 7.641679
2	Valor únic: 75528 Summary: Si es tanca l'aixeta del petroli, Cuba entrarà en crisi. Línia similar: Una economia en cris. Valor de la distància entre les frases: 9.015269
3	Valor únic: 51997 Summary: En una entrevista al diari alemany 'Frankfurter Allgemeine', el ministre d'Economia es mostra convençut que Espanya sortirà de la crisi "per si mateixa". Admet, però, que el 2012 "serà un any dur". Línia similar: Respecte a les perspectives de creixement per al 2012, el titular d'Economia reconeix que "aquest any serà difícil". Valor de la distància entre les frases: 9.183023
4	Valor únic: 18791 Summary: El delegat del govern espanyol assegura que la Moncloa segueix "de prop"els plans sobiranistes i que actuarà amb "fermesaï "proporcionalitat". Crida al "diàleg"entre administracions i carrega contra la consulta per separar entre "guanyadors i vençuts". Línia similar: Un procés que minva el çreixement"econòmi. Valor de la distància entre les frases: 9.191608
5	Valor únic: 87513 Summary: El PACMA ha difós el vídeo i ha demanat la col·laboració ciutadana per identificar l'agressor. Línia similar: "¿Aquest és el tipus de persones que «vetllen» per la naturalesa i les que «garanteixen» un equilibri de l'ecosistema? Valor de la distància entre les frases: 9.363764
6	Valor únic: 25436 Summary: El candidat de Democràcia i Llibertat a les eleccions espanyoles pronostica que hi haurà acord entre Junts pel Sí i la CUP per a la investidura del president de la Generalitat. Línia similar: Crec honestament que a Espanya li interessa això per raons econòmiques i democràtiques. Valor de la distància entre les frases: 9.373934
7	Valor únic: 69103 Summary: Pequín vol que l'any del mico serveixi per adaptar-se a la "nova normalitat" fruit de l'alentiment econòmic. Línia similar: L'alentiment de l'economia té a favor la demografia. Valor de la distància entre les frases: 9.406595
8	Valor únic: 51526 Summary: Assegura que les retallades no suposen la "demolició de l'estat del benestar". Línia similar: Creu que l'any que ve l'economia espanyola començarà a créixer. Valor de la distància entre les frases: 9.412511
9	Valor únic: 19498 Summary: L'expresident espanyol retreu a Puigdemont que no assisteixi a la conferència de presidents i reclama a Rajoy un "esforç"en finançament per a la llei de Dependència, que avui compleix 10 anys. Línia similar: El grup socialista presentarà una iniciativa al Congrés perquè aquesta norma recuperi l·la velocitat de creuer"ara que l'economia çreix". Valor de la distància entre les frases: 9.434928
10	Valor únic: 36698 Summary: Irromp al baròmetre del CIS com a tercera força, deixant enrere IU i trepitjant ja els talons al PSOE. Línia similar: A més, la quarta preocupació dels ciutadans, per darrere de l'atur, la corrupció i l'economia, ja són "els polítics, els partits i la política". Valor de la distància entre les frases: 9.487665

Taula B.26: Resultats consulta 5 model STSB.

Rànquing	Descripció Resultat
1	Valor únic: 73747 Summary: Algunes informacions apunten que hi poden viure uns cinc mil presoners. Línia similar: L'anàlisi, que cobreix l'activitat observada en el camp en els últims 12 mesos, s'ha basat en imatges recollides per Airbus aquest març. Valor de la distància entre les frases: 8.075607
2	Valor únic: 8402 Summary: L'ANC destaca el "bon ritme" en la venda de samarretes i espera almenys un milió d'assistents. Línia similar: Romeu subratlla que les últimes setmanes d'agost i sobretot l'inici de setembre serà el període en què el degoteig d'inscripcions anirà en augment. Valor de la distància entre les frases: 8.144165
3	Valor únic: 83565 Summary: L'ACCO considera que les clàusules són restrictives de la competència. Línia similar: En aquestes licitacions s'exigeix un nivell d'experiència prèvia en nombre de quilòmetres recorreguts anuals en transport urbà molt elevat que limita injustificadament el nombre d'operadors que poden participar en les licitacions, indica Competència. Valor de la distància entre les frases: 8.151097
4	Valor únic: 2577 Summary: Els cupaires, reticents en el passat a formar part del cartipàs, són en dotze executius més que el 2015. Línia similar: "Aquest cop hem sigut menys dogmàtics per l'emergència social en què viu Figueres, en el top 3 de fracàs escolar i amb un 60% de població en risc d'exclusió social", argumenta. Valor de la distància entre les frases: 8.222145
5	Valor únic: 82799 Summary: Andrea Fumagalli, professor d'economia política de la Universitat de Pavia, una veu molt crítica amb el govern, ens explica els punts bàsics i més controvertits de la reforma universitària italiana. Una llei que està previst que s'aprovi entre avui i demà i que canviarà de dalt a baix l'estructura organitzativa i econòmica de la institució. Línia similar: " Primer, es redueix el finançament general a tot el sistema universitari en 1.300 milions d'euros en tres anys, aplicant-se la retallada més forta aquest curs 2010 – 2011". Valor de la distància entre les frases: 8.271723
6	Valor únic: 64672 Summary: La classe mitjana ha pujat de 5 a 300 milions de persones a la Xina en 17 anys i el 2020 superarà Europa i els EUA. Línia similar: És en aquesta classe mitjana i en el seu creixement exponencial -l'any 2000 només n'eren 5 milions-, en qui confia el govern xinès per transformar l'economia cap a un model apuntalat en el consum intern. Valor de la distància entre les frases: 8.360283
7	Valor únic: 18184 Summary: El mateix estudi, centrat en el mercat immobiliari europeu, afirma que el procés podria desviar inversions immobiliàries de Barcelona a Lisboa. Línia similar: També Madrid sembla que perd atractiu entre les ciutats europees per rebre inversió del sector immobiliari, ja que en aquesta edició ocupa el lloc número 9, quan l'any passat estava en setena posició. Valor de la distància entre les frases: 8.389676
8	Valor únic: 22773 Summary: La Generalitat calcula que Catalunya va aportar 14.623 milions a l'Estat el 2012 que no van tornar. Línia similar: El Govern, però, reivindica el model de flux monetari, sobretot "en èpoques de crisi econòmica i taxes d'atur elevades", quan "pren molta més rellevància" l'impacte de "l'activitat de l'administració central en un territori". Valor de la distància entre les frases: 8.412851
9	Valor únic: 69723 Summary: La potent pertorbació tropical 'Patricia' entrarà pel Pacífic amb vents sostinguts de més de 300 km/h i un mínim de pressió atmosfèrica inèdit. Línia similar: S'espera que l'huracà segueixi una trajectòria en direcció nord-oest, un cop hagi tocat terra anirà perdent força, però les últimes previsions del National Hurricane Center alerten que mantindrà categoria màxima encara durant 12 hores. Valor de la distància entre les frases: 8.438246
10	Valor únic: 74147 Summary: El gas, el petroli i els interessos comercials frenen les sancions econòmiques d'Europa contra Moscou. Línia similar: L'expert en qüestions energètiques del Centre d'Estudis Europeus (CEPS) Arno Behrens creu que en el cas de sancions econòmiques en l'àmbit energètic el gran perdedor és Rússia, sobretot si Europa aprova un full de ruta per reduir la dependència energètica del país en els pròxims anys. Valor de la distància entre les frases: 8.463324

Taula B.27: Resultats consulta 5 model Word2Vec.

B.3.6. Consulta 6: votació d'abril del 2019

Rànquing	Descripció Resultat
1	Valor únic: 5020 Summary: El fotògraf Isidre García Puntí ha retratat alguns protagonistes de la repressió. Línia similar: Abril del 2018. Valor de la distància entre les frases: 9.425442
2	Valor únic: 71261 Summary: Sismes a la Xina i Haití han estat els pitjors per nombre de morts. Línia similar: - 20 d'abril de 2014. Valor de la distància entre les frases: 10.268601
3	Valor únic: 70515 Summary: Els BRICS es conjuren per prescindir del dòlar en el seu comerç. Línia similar: Abril del 2016. Valor de la distància entre les frases: 11.035303
4	Valor únic: 15926 Summary: Diosdado Toledano, portaveu de l'entitat promotora de la ILP, assegura que és un "dia històric"per als ciutadans més vulnerables de Catalunya. Línia similar: Es tracta d'un import que augmentarà progressivament i que l'1 d'abril del 2020 arribarà als 664 euros. Valor de la distància entre les frases: 11.037055
5	Valor únic: 63133 Summary: Les dones musulmanes engeguen una campanya per salvar la noia, forçada a casar-se als 16 anys. Línia similar: Era l'abril del 2017. Valor de la distància entre les frases: 11.151681
6	Valor únic: 26942 Summary: El periodista Albert Cuesta recull al seu blog una selecció de textos en castellà sobre el dret a decidir i el 27-S. Línia similar: Departament de la Presidència (20 abril 2015). Valor de la distància entre les frases: 11.297127
7	Valor únic: 74354 Summary: Renzi serà el tercer primer ministre que no és escollit a les urnes. Línia similar: Abril 2013. Valor de la distància entre les frases: 11.677885
8	Valor únic: 71261 Summary: Sismes a la Xina i Haití han estat els pitjors per nombre de morts. Línia similar: - 6 d'abril de 2009. Valor de la distància entre les frases: 11.739299
9	Valor únic: 65890 Summary: Recull de portades internacionals sobre el resultat de la primera volta de les eleccions franceses. Línia similar: Portada de 'Libération' de 24 d'abril de 2017. Valor de la distància entre les frases: 11.786069
10	Valor únic: 65890 Summary: Recull de portades internacionals sobre el resultat de la primera volta de les eleccions franceses. Línia similar: / The Guardia Portada de 'The Guardian' de 24 d'abril de 2017. Valor de la distància entre les frases: 11.864983

Taula B.28: Resultats consulta 6 model STSB.

Rànquing	Descripció Resultat
1	Valor únic: 83718 Summary: La redacció del projecte s'endarrereix i els treballadors es queixen del silenci que envolta la reforma. Línia similar: L'últim trajecte del Tramvia Blau, el 28 de gener del 2018 L'últim trajecte del Tramvia Blau, el 28 de gener del 2018. Valor de la distància entre les frases: 9.329926
2	Valor únic: 9503 Summary: El president afirma que el seu Govern ha vingut per "fer fora el 155". Línia similar: L'hemicicle del Parlament, durant el ple del 6 de juny del 2018 / CÈLIA ATSE L'hemicicle del Parlament, durant el ple del 6 de juny del 2018 / CÈLIA ATSE. Valor de la distància entre les frases: 9.571959
3	Valor únic: 15150 Summary: La societat civil s'ha fet un lloc a la taula principal de negociació des del 2012. Línia similar: L'11 de setembre del 2012, el primer punt d'inflexió polític del Procés. Valor de la distància entre les frases: 9.604166
4	Valor únic: 2448 Summary: El major dels Mossos, acusat de rebel·lió, serà jutjat a l'Audiència Nacional amb Soler, Puig i Laplana. Línia similar: El 20 de gener del 2020 arrencarà a l'Audiència Nacional el judici contra el major dels Mossos, Josep Lluís Trapero, acusat d'un delicte de rebel·lió pels fets del 20 de setembre i l'1 d'octubre del 2017. Valor de la distància entre les frases: 9.760231
5	Valor únic: 17240 Summary: L'expresident del PP de Castelló ja ha complert tres quartes parts de la condemna per un delicte de frau fiscal. Línia similar: L'excap del PP provincial gaudia del tercer grau penitenciari des de l'abril del 2016, concedit 1 any i 3 mesos després del seu ingrés a presó. Valor de la distància entre les frases: 9.767204
6	Valor únic: 89824 Summary: Alguns veïns han acabat marxant de casa seva després d'onze anys denunciant la situació. Línia similar: El juny del 2008 l'inspector del districte Joan Llucià ja ordenava el cessament de l'activitat de 10 pisos turístics. Valor de la distància entre les frases: 9.819455
7	Valor únic: 8238 Summary: L'expresident publicarà uns dies abans el llibre 'La crisi catalana, una oportunitat per a Europa'. Línia similar: Aquesta conferència també tindrà lloc l'endemà del primer aniversari del referèndum d'autodeterminació de l'1-O. Valor de la distància entre les frases: 9.851211
8	Valor únic: 13640 Summary: L'alt tribunal adverteix Forcadell que no pot fer res que "atorgui valor jurídic"a la proclamació. Línia similar: El ple ordinari del Tribunal Constitucional ha decidit suspendre cautelarment la declaració d'independència del Parlament del 27 d'octubre. Valor de la distància entre les frases: 10.007367
9	Valor únic: 60057 Summary: L'executiu de Nicola Sturgeon vol celebrar la nova consulta el segon semestre de 2020. Línia similar: Nicola Sturgeon vol fer el segon referèndum d'independència d'Escòcia abans del maig del 202. Valor de la distància entre les frases: 10.039828
10	Valor únic: 8425 Summary: El Governconsidera que és una mostra del "respecte a la bilateralitat". Línia similar: Els pressupostos del 2019 seran el primer examen important del Govern abans d'acabar l'any. Valor de la distància entre les frases: 10.142507

Taula B.29: Resultats consulta 6 model Word2Vec.

B.3.7. Consulta 7: El jutge decreta error en la sentència

Rànquing	Descripció Resultat
1	Valor únic: 59458 Summary: La fiscalia demanava 40 anys de presó per a la noia, de 21 anys, que havia sigut violada. Línia similar: El jutge José Virgilio Jurado Martínez, del Tribunal de Sentència de la ciutat de Cojutepeque, ha presidit el nou judici contra la jove, que es va celebrar entre dijous i divendres passat. Valor de la distància entre les frases: 8.391141
2	Valor únic: 6731 Summary: Sentencia que és ïmpossible"arribar a cap acord ni negociació sobre un referèndum acordat. Línia similar: El president de la Generalitat, Quim Torra, considera que seria ïrresponsableçonvocar eleccions després de la sentència dels judicis pel Procés. Valor de la distància entre les frases: 8.550101
3	Valor únic: 79380 Summary: El tribunal militar al·lega que no hi ha proves suficients per condemnar Ahmed Adel pels abusos denunciats per l'activista Samira Ibrahim. El jutge ha detectat çontradiccions"en un dels testimonis de la defensa. Línia similar: En la seva sentència, el jutge ha detectar çontradiccions"en les declaracions de les testimonis, principalment en les d'una jove que havia estat convocada a petició de la defensa. Valor de la distància entre les frases: 8.870489
4	Valor únic: 3759 Summary: Un 55% dels ciutadans creuen que la independència dels jutges és pobra, i ho atribueixen a pressions econòmiques o polítiques. Línia similar: En aquest sentit, doncs, el nombre de jutges no és una garantia de percepció de bon funcionament perquè Dinamarca és el sistema judicial amb millor percepció d'independència però està a la cua en nombre de jutges. Valor de la distància entre les frases: 8.991712
5	73223 Summary: Un tribunal canvia la pena de mort que se li havia imposat i que va ser recorreguda per la Fiscalia. Línia similar: Aquest tribunal havia enviat els expedients dels implicats fins a dues ocasions al muftí de la República i màxima autoritat religiosa del país, Shauqi Alam, perquè es pronunciés sobre la pena capital a què van ser sentenciats de forma provisional. Valor de la distància entre les frases: 9.211520
6	Valor únic: 42560 Summary: Carles Mateu ha conegut aquest divendres la sentència de l'Audiència de Castelló, que també el condemna a un any i un dia de retirada del permís de conduir. Línia similar: Aquesta sentència revoca la decisió del jutge de primera instància, que va resoldre a favor de Carles Mateu Blay i que suposava la seva absolució dels delictes penals que se li imputaven. Valor de la distància entre les frases: 9.218497
7	Valor únic: 22698 Summary: El magistrat situa el partit com a beneficiari de 204.198 euros de l'entramat. Línia similar: La causa a què fa referència el jutge versa sobre les adjudicacions suposadament irregulars que va obtenir el cap de la trama, Francisco Correa, per a les seves empreses, amb l'ajut dels altres acusats, entre els quals l'ex-alcalde de Boadilla, Arturo González Panero, l'exregidor José Galeote i l'exdiputat del PP, Alfonso Bosch. Valor de la distància entre les frases: 9.314591
8	Valor únic: 12958 Summary: L'exdirectora general de Turisme ha declarat aquest dimecres com a investigada en el cas Cursach. Línia similar: L'exdirectora general també ha explicat que ha respost a totes les preguntes del jutge, el fiscal anticorrupció Miguel Ángel Subirán i el seu advocat, Enrique Ordóñez. Valor de la distància entre les frases: 9.367248
9	Valor únic: 8306 Summary: El Consell Fiscal decidirà si demana al ministeri de Justícia la documentació sobre la demanda. Línia similar: Reclamen que el Consell Fiscal tingui çoneixement complet"de les actuacions dutes a terme en defensa de la sobirania jurisdiccional espanyola, la independència dels seus tribunals i la del jutge Llarena. Valor de la distància entre les frases: 9.381698
10	Valor únic: 15526 Summary: Dona per primer cop als jutges ordinaris la potestat de no aplicar normes que contradiguin les estatals. Línia similar: Els requisits perquè els jutges ordinaris puguin decidir inaplicar una llei autonòmica són, segons marca la sentència, que la llei estatal posterior sigui bàsica. Valor de la distància entre les frases: 9.466796

Taula B.30: Resultats consulta 7 model STSB.

Rànquing	Descripció Resultat
1	Valor únic: 82802 Summary: La Fiscalia demana cadena perpètua i el militar justifica la seva actuació per fer front a l'l'agressió terrorista contra els béns materials i les persones". Línia similar: El tribunal que jutja el dictador argentí Jorge Rafael Videla farà pública avui la sentència en el procés per delictes de repressió. Valor de la distància entre les frases: 8.967111
2	Valor únic: 5555 Summary: A una setmana del judici la sala respon a les defenses que no s'està çriminalitzant una ideologia". Línia similar: El Tribunal Constitucional ajorna la decisió sobre el recurs de llibertat d'Oriol Junquera. Valor de la distància entre les frases: 9.196900
3	Valor únic: 19116 Summary: Endureix la sentència de l'Audiència de Madrid perquè considera que els ultres van actuar la Diada del 2013 per motius de "discriminació ideològica". Línia similar: El tribunal estima el recurs presentat per la Fiscalia, en contra de la decisió de l'Audiència Provincial de Madrid el febrer passat. Valor de la distància entre les frases: 9.294696
4	Valor únic: 12402 Summary: El tribunal resoldrà sobre Junqueras el 4 de gener i la fiscalia demana imputar Jové i Trapero. Línia similar: El Suprem també ha indicat el 4 de gener la data en què resoldrà sobre el recurs d'apel·lació d'Oriol Junqueras a la decisió de Pablo Llarena de mantenir-lo en presó provisional. Valor de la distància entre les frases: 9.525572
5	Valor únic: 71688 Summary: Conclou així el procés obert des de el 15 de febrer de 2011 contra l'antic 'Cavaliere', acusat de presumpte abús de poder i incitació a la prostitució de menors. Línia similar: El Suprem ha confirmat així la sentència d'absolució dictada en segona instància pel Tribunal d'Apel·lació de Milà el juliol de 2014 i apel·lada per la fiscalia d'aquesta ciutat el passat novembre. Valor de la distància entre les frases: 9.538102
6	Valor únic: 31232 Summary: Fernández Díaz afirma que Espanya és ün dels països més segurs del món". Línia similar: El jutge també descriu la importància d'Internet en aquest nou terrorisme. Valor de la distància entre les frases: 9.613154
7	Valor únic: 1252 Summary: El tribunal argumenta que la qüestió prejudicial només afecta la situació personal del líder d'ERC. Línia similar: El Tribunal Suprem ha rebutjat aquest dimarts deixar en suspens la sentència del judici del Procés fins que el Tribunal de Justícia de la Unió Europea (TJUE) resolgui sobre la immunitat d'Oriol Junqueras, tal com havia demanat la seva defensa el 17 de setembre. Valor de la distància entre les frases: 9.635337
8	Valor únic: 4666 Summary: Assegura que va advertir el Govern que no comptés amb el cos per fer la independència. Línia similar: El president del tribunal ha introduït una novetat en l'interrogatori atorgant-se la prerrogativa prevista a la Llei d'Enjudiciament Criminal d'interpel·lar el testimoni. Valor de la distància entre les frases: 9.701994
9	Valor únic: 11594 Summary: Un recurs de Zapatero tomba part de la llei de l'Aran aprovada pel Parlament el 2010. Línia similar: En la sentència, l'alt tribunal estima parcialment el recurs presentat pel govern socialista de José Luis Rodríguez Zapatero emparant-se en la sentència de l'Estatut. Valor de la distància entre les frases: 9.727953
10	Valor únic: 60736 Summary: Üna cosa és el control migratori i l'altra el dret de les persones que transiten fronteres", diu l'activista. Línia similar: El jutjat d'instrucció ja havia tancat provisionalment el cas el 12 de desembre, però la fiscalia havia recorregut la decisió. Valor de la distància entre les frases: 9.728551

Taula B.31: Resultats consulta 7 model Word2Vec.

B.3.8. Consulta 8: Els diputats votaran aquest dijous

Rànquing	Descripció Resultat
1	Valor únic: 27580 Summary: També ha aprovat traslladar a la Fiscalia les incompareixences en la comissió sobre el frau fiscal. Línia similar: Els diputats populars han votat en contra d'aquest punt. Valor de la distància entre les frases: 9.292365
2	Valor únic: 66632 Summary: El cap de l'Oficina d'Ètica Governamental diu que el magnat serà vulnerable a "sospites de corrupció". Línia similar: "Cada decisió que prendrà com a president serà perseguida pel fantasma del dubte", va afirmar a Bloomberg News Noah Bookbinder, director de Ciutadans per la Responsabilitat i l'Ètica. Valor de la distància entre les frases: 9.941839
3	Valor únic: 40758 Summary: L'alcalde de Lleida assegura que prendran una decisió aquest dimecres, però que no té per què ser unànime. Línia similar: En aquest sentit, ha assegurat que "tenia més profunditat política la votació en què cinc diputats vam trencar la disciplina que la de dijous". Valor de la distància entre les frases: 9.648776
4	Valor únic: 3708 Summary: El PSC pren als comuns el primer lloc a la demarcació de Barcelona. Línia similar: Si a les eleccions generals del 2016 ERC ja va ser l'opció més votada a la circumscripció de Girona, els republicans van confirmar el seu ascens: per primer cop van aconseguir tres diputats. Valor de la distància entre les frases: 9.702732
5	Valor únic: 13008 Summary: Més de 200 personalitats de diferents àmbits impulsen una plataforma de suport a Miquel Iceta. Línia similar: A votar per posar punt final a aquesta bogeria i deixar de perdre temps, oportunitats i amistats", va reblar la candidata de Cs, que va destacar que "ara sí que votarem per tenir un govern de tots, un govern que respecti tots els catalans". Valor de la distància entre les frases: 9.888430
6	Valor únic: 34765 Summary: El Barcelonès, amb 9.473, és la comarca amb més voluntaris. Hi ha nou comarques que n'han registrat més de mil. Línia similar: La xifra total sorgeix una vegada depurades, per evitar duplicitats, les dades de les 38.706 inscripcions recollides fins dilluns per internet més la suma del personal de centres docents (4.170) i els que es van apuntar a les delegacions territorials de la Generalitat (684). Valor de la distància entre les frases: 9.923623
7	Valor únic: 15459 Summary: La Generalitat i l'Ajuntament de Barcelona han convocat un minut de silenci per demà a les 12 del migdia a plaça Catalunya. Línia similar: Puigdemont ha comparegut acompanyat del vicepresident del Govern, Oriol Junqueras, i de l'alcaldessa de Barcelona, Ada Colau, per demostrar que estan ünits contra la barbàrie". Valor de la distància entre les frases: 9.930522
8	Valor únic: 3260 Summary: "Hem de parlar", li diu Junqueras a Sánchez en una sessió marcada pel boicot de la triple dreta als acatament dels independentistes. Línia similar: Van ofegar les paraules d'uns diputats electes -que parlaven sense micròfon- perquè, segons cridaven alguns diputats, era "una vergonya" que les poguessin pronunciar. Valor de la distància entre les frases: 9.958739
9	Valor únic: 14801 Summary: La llista de punts de votació inclou equipaments múnicipals de Barcelona i diversos CAP. Línia similar: En aquesta web trobaràs el lloc on et correspon: La web, que és idèntica a la del referèndum que ha sigut anul·lada reiteradament per la justícia espanyola, inclou aquest cop un apartat a la part inferior que diu "On votar". Valor de la distància entre les frases: 9.971309
10	Valor únic: 7719 Summary: Milers de persones van plantar cara a l'intent de l'Estat d'aturar el referèndum per la força. Línia similar: El vicepresident del Govern, Oriol Junqueras, també diposita el seu vot al centre de la Guàrdia, a Sant Vicenç dels Horts: "Estic segur que ens en sortirem com a país". Valor de la distància entre les frases: 9.980681

Taula B.32: Resultats consulta 8 model STSB.

Rànquing	Descripció Resultat
1	Valor únic: 53943 Summary: Obren sense incidents destacables els 2.696 col·legis electorals catalans, on estan cridades a votar gairebé 5,4 milions de persones. A les urnes es resoldrà la lluita frec a frec entre CiU, PSC i PP que auguren les enquestes. Línia similar: Els cinc candidats catalans tenen previst votar tots aquest matí. Valor de la distància entre les frases: 10.657254
2	Valor únic: 21705 Summary: Un portaveu de les víctimes de l'Alvia protesta en contra de l'elecció de Pastor com a presidenta del Congrés. Línia similar: Els diputats canaris d'Units Podem han ocupat els vuit llocs que durant l'anterior legislatura tenien els nacionalistes catalans. Valor de la distància entre les frases: 11.491500
3	Valor únic: 2163 Summary: "Són acusacions inèdites que pretenen desmobilitzar, atemorir i silenciar la lluita", denuncien. Línia similar: Els Mossos d'Esquadra detenen set activistes independentistes acusats d'organització crimina. Valor de la distància entre les frases: 11.556428
4	Valor únic: 77575 Summary: El govern de Portugal ha confirmat el que els ciutadans temien des de fa setmanes: un augment exponencial dels impostos. El ministre de Finances va presentar ahir uns pressupostos beneïts per la troica. Línia similar: Els pressupostos presentats ahir per l'executiu lusità ataquen sobretot els ciutadans. Valor de la distància entre les frases: 11.890424
5	Valor únic: 9025 Summary: Cospedal creu que és un dia important "per Espanyaï Santamaría que són el partit de la llibertat. Línia similar: Els inscrits poden votar fins les 20:30 d'aquest dijous. Valor de la distància entre les frases: 11.954458
6	Valor únic: 89372 Summary: Els carrers guarnits treuen pit 12 mesos després que els atemptats de la Rambla monopolitzessin l'agost. Línia similar: Els homes l'estan liant!"–. Valor de la distància entre les frases: 11.984314
7	Valor únic: 37816 Summary: A l"Abc': L·Lliçó democràtica de Rajoy i Rubalcaba a l'esquerra radical i el nacionalisme". Línia similar: Els diaris d'aquest dijous titulen. Valor de la distància entre les frases: 12.202709
8	Valor únic: 71809 Summary: Alemanya fa costat a Rajoy en la seva polèmica amb Tsipras: "Això a l'Eurogrup no es fa". Línia similar: Els diputats sí que podran votar l'acord definitiu. Valor de la distància entre les frases: 12.219653
9	Valor únic: 9469 Summary: El ministre d'Interior, assenyalat per no investigar tortures i qüestionar la llibertat d'expressió. Línia similar: Els 17 ministres de Pedro Sánchez ja han pres possessió després que aquest dimecres s'acabessin de tancar tots els noms. Valor de la distància entre les frases: 12.266848
10	Valor únic: 89615 Summary: El personal de seguretat podrà aprendre els indicis que poden ajudar a detectar radicals violents. Línia similar: Els Mossos d'Esquadra formaran els vigilants privats sobre terrorisme jihadista. Valor de la distància entre les frases: 12.467051

Taula B.33: Resultats consulta 8 model Word2Vec.

B.3.9. Consulta 9: Espanya es prepara per a una crisi

Rànquing	Descripció Resultat
1	Valor únic: 49412 Summary: Çatalunya necessita l'instrument d'un estat, qualsevol nació el pot arribar a tenir; aquest és el clam de la gent", diu el president de la Generalitat a la capital espanyola. "Havíem pensat durant molt de temps que aquest estat podia ser l'espanyol i hem treballat molt perquè això fos així", afegeix. Mas reclama que no es minimitzi la manifestació de la Diada i una consulta per preguntar si Catalunya és una nació. Línia similar: La crisi econòmic. Valor de la distància entre les frases: 8.776558
2	Valor únic: 14811 Summary: "Madrid ha de recórrer a arguments i no a la força per evitar un trencament", opina el 'Times'. Línia similar: I diu que "Espanya entra en una crisi d'Estat". Valor de la distància entre les frases: 9.020654
3	Valor únic: 4381 Summary: "Si només hi ha dues respostes es condiciona la votació", opina Pablo Iglesias. Línia similar: Pel que fa a la crisi interna, Iglesias ho ha atribuït a l'"adolescència de la formació". Valor de la distància entre les frases: 9.238693
4	Valor únic: 48399 Summary: "Per culpa de les polítiques d'Artur Mas un nen català no té les mateixes oportunitats que un altre depenent de la família on neixi", assegura el líder socialista durant l'aprovació del programa electoral. Línia similar: Hi ha una alternativa social a aquesta crisi. Valor de la distància entre les frases: 9.247504
5	Valor únic: 58629 Summary: Fill de Daphne Caruana Galizia, la reportera maltesa assassinada l'octubre del 2017. Línia similar: La crisi polític. Valor de la distància entre les frases: 9.355510
6	Valor únic: 53076 Summary: El president té previst fer la seva primera visita oficial aquesta primavera al Marroc. El secretari d'exteriors del Govern assegura que no s'obriran ni es tancaran més delegacions a l'estranger. Línia similar: Diplomàcia en temps de crisi. Valor de la distància entre les frases: 9.366208
7	Valor únic: 78605 Summary: Les matances a Síria, les intrigues al Vaticà i la l'evolució de les borses centren l'interès de la premsa internacional. Línia similar: 'La Croix': Çrisi de poder al vaticàï "Espanya desestabilitzada per la crisi del sector bancari. Valor de la distància entre les frases: 9.403340
8	Valor únic: 55590 Summary: El president espanyol, amb to de comiat, ha reclamat ün esforç col·lectiu"per sortir de la crisi i ha confiat que el PIB creixi per sobre de l'1,5% el quart trimestre d'enguany. Anuncia mesures de protecció per als que no poden pagar la hipoteca. Demana ün respecte sincer, no retòric"pel moviment del 15M. Línia similar: El president espanyol ha reclamat un "esforç col·lectiu"per sortir de la crisi. Valor de la distància entre les frases: 9.451238
9	Valor únic: 70293 Summary: La pressió a la Mànega tensa les relacions entre Londres i París. Línia similar: Una crisi diplomàtica. Valor de la distància entre les frases: 9.491224
10	Valor únic: 31255 Summary: L'alcalde i candidat a la reelecció acusa Colau de fer "demagògiaçontra l'activitat econòmica de Barcelona. Línia similar: "Hem governat en moments de crisi social, política. Valor de la distància entre les frases: 9.531265

Taula B.34: Resultats consulta 9 model STSB.

Rànquing	Descripció Resultat
1	Valor únic: 50713 Summary: El portaveu socialista, Antonio Torres, considera "vergonyós"que el president de les Corts Valencianes s'atreveixi a qüestionar la sentència del Tribunal Constitucional. Línia similar: Espanya es trenca quan es porta a la fallida les institucions financeres valencianes, quan es condemna a la desocupació milers de famílies". Valor de la distància entre les frases: 11.248956
2	Valor únic: 8869 Summary: Preveu fer un congrés fundacional perquè els adherits decideixin la fórmula jurídica del moviment. Línia similar: Ara per ara a les ponències que es debatran al congrés no es fa cap menció a la Crida de Puigdemont, i per això no es descarta que a última hora es puguin presentar in situ per discutir-ho. Valor de la distància entre les frases: 11.277195
3	Valor únic: 69794 Summary: Brussel·les demana a Istanbul que talli el flux migratori cap a Europa i Erdogan exigeix contrapartides. Línia similar: Actualment a Turquia es produeix una situació paradoxal. Valor de la distància entre les frases: 11.295153
4	Valor únic: 21027 Summary: EH Bildu seria segona força amb 17 diputats (quatre menys que ara), i Podem superaria els socialistes i entraria al parlament amb 15 escons. Línia similar: Ciutadans, que com Podem es presenta per primer cop a unes autonòmiques a Euskadi, es quedaria sense representació a la cambra basca. Valor de la distància entre les frases: 11.535843
5	Valor únic: 59518 Summary: Malta s'ofereix a acollir-les però rebutja l'entrada de les altres 121, que porten nou dies a bord. Línia similar: Activistes es traslladen a Lampedusa per forçar que es doni port a l''Open Arms. Valor de la distància entre les frases: 11.553274
6	Valor únic: 61903 Summary: El règim carrega la responsabilitat sobre el sotscap dels serveis d'intel·ligència, Ahmad Al-Assiri. Línia similar: Tota una bufetada per a un règim que es gasta cada any una milionada per millorar la seva imatge a Occident. Valor de la distància entre les frases: 11.650898
7	Valor únic: 86606 Summary: Eines com Alexa o Siri, amb veus femenines, perpetuen la idea que les que reben ordres són elles. Línia similar: La científica afegeix que cal una formació més humanista per a les persones que es dediquen a desenvolupar tecnologia. Valor de la distància entre les frases: 11.728108
8	Valor únic: 16529 Summary: Entre l'anunci de la pregunta i la jornada de votació el Govern té encara un camí d'obstacles a recórrer. Línia similar: Una opció que es torna a valorar ara. Valor de la distància entre les frases: 11.859256
9	Valor únic: 16531 Summary: El Govern situa el referèndum al calendari i es prepara per afrontar els entrebancs de l'Estat. Línia similar: Una opció que es torna a valorar ara. Valor de la distància entre les frases: 11.859256
10	Valor únic: 88449 Summary: Ha de pagar 3,3 milions a la família d'un nen que va néixer amb paràlisi cerebral. Línia similar: Es tracta de la indemnització més gran que es concedeix a Espanya per una negligència mèdica. Valor de la distància entre les frases: 11.929145

Taula B.35: Resultats consulta 9 model Word2Vec.

B.3.10. Consulta 10: El partit d'esquerres guanyarà les futures eleccions

Rànquing	Descripció Resultat
1	Valor únic: 19618 Summary: Els republicans confien que la cimera del referèndum del proper 23 de desembre servirà per comprovar çom de fort i sa"està el pacte pel dret a decidir. Línia similar: Els 'comuns' preparen el manifest fundacional per al que acabarà sent el futur partit de la confluència d'esquerres. Valor de la distància entre les frases: 10.097030
2	Valor únic: 33270 Summary: El líder d'ERC demana al president que respongui als gestos republicans amb noves eleccions. Línia similar: Lluny de l'acord, encara, sobre el format de les futures eleccions plebiscitàries. Valor de la distància entre les frases: 10.202158
3	Valor únic: 81 Summary: El partit encara manté obertes les negociacions amb MES, Avancem i Demòcrates i amb independents. Línia similar: Les enquestes pronostiquen un gran creixement d'ERC, que guanyaria per primera vegada les eleccions al Parlament des de la Segona República. Valor de la distància entre les frases: 10.344327
4	Valor únic: 72157 Summary: El president del Parlament Europeu veu com a alternativa la lluita contra l'evasió fiscal. Línia similar: El president del Parlament Europeu (PE), Martin Schulz, està convençut que no hi haurà acord sobre una possible quita del deute grec, una de les demandes clau de la coalició d'esquerres Syriza, guanyadora de les eleccions d'ahir. Valor de la distància entre les frases: 10.555622
5	Valor únic: 61247 Summary: El nou president ha nomenat un gabinet de militars, antifeministes i una religiosa per "redreçar" el país. Línia similar: En la investidura va parlar de lluitar contra la "submissió ideològica" que encarnen les idees d'esquerres, del progressisme al marxisme. Valor de la distància entre les frases: 10.568587
6	Valor únic: 67389 Summary: L·La metodologia de les negociacions ha sigut un encert: rondes quinzenals, amb quinze dies de pausa i reflexió". Línia similar: Tot plegat, el que s'ha fet i com s'ha fet, marcarà història en el panorama dels processos de pau al món, i servirà d'important referència per a futures negociacions. Valor de la distància entre les frases: 10.664132
7	Valor únic: 16334 Summary: El president valencià ha escollit el lema 'L'esquerra en marxa' en una clara al·lusió al títol del 39è congrés del PSOE, 'Som l'esquerra'. Línia similar: 'L'esquerra en marxa. Valor de la distància entre les frases: 10.673818
8	Valor únic: 14877 Summary: És la mà dreta de Junqueras al Govern, el coordinador tècnic del referèndum i un dels 12 detinguts. Línia similar: Economista i politòleg de formació, ja va ocupar un càrrec important en el departament de Vicepresidència durant el tripartit i és el president del Consell Nacional d'Esquerra des del 2011. Valor de la distància entre les frases: 10.715225
9	Valor únic: 67960 Summary: La punxada econòmica a l'Amèrica Llatina posa en risc els avenços socials de l'última dècada. Línia similar: La desacceleració de les economies emergents que demandaven commodities ha impactat en els pressupostos, explica Andrea Costafreda, directora programàtica d'Oxfam Intermón per a l'Amèrica Llatina i el Carib. Valor de la distància entre les frases: 10.726598
10	Valor únic: 27639 Summary: La confluència d'esquerres busca erigir-se com a alternativa a les polítiques de Mas. Línia similar: Si guanya les eleccions, la confluència es compromet a respondre a l'emergència social; fer una auditoria del deute de la Generalitat; revertir la privatització dels serveis públics; garantir la democràcia en les relacions econòmiques, i acabar amb la corrupció, entre altres coses. Valor de la distància entre les frases: 10.770010

Taula B.36: Resultats consulta 10 model STSB.

Rànquing	Descripció Resultat
1	Valor únic: 66202 Summary: També se li prohibeix representar el Parlament Europeu en reunions interparlamentàries, conferències o qualsevol tipus de fòrum durant un any. L'ultra Janusz Korwin-Mikke va defensar que les dones han de cobrar menys "perquè són més dèbils i menys intel·ligents". Línia similar: El Parlament Europeu investigarà l'eurodiputat polonès per les declaracions misògine. Valor de la distància entre les frases: 9.281645
2	Valor únic: 16373 Summary: El president d'ERC carrega contra l·l'irresponsable"govern espanyol, que actua "sistemàticament en contra"dels interessos de la ciutadania. Línia similar: El partit pretén que aquesta conferència nacional marqui les línies bàsiques que defensarà ERC durant el procés constituent. Valor de la distància entre les frases: 9.286292
3	Valor únic: 81053 Summary: El règim anuncia que seran "netes i justes i portaran a una Assemblea Popular que representi el poble sirià a través del pluralisme polític", segons l'agència de SANA. Línia similar: El ministre sirià d'Exteriors, Walid al-Mualem, ha assegurat avui que el seu país celebrarà eleccions legislatives abans de finals d'any i que les urnes actuaran "d'àrbitre"de les reformes anunciades pel president Baixar al-Assad. Valor de la distància entre les frases: 9.405838
4	Valor únic: 32571 Summary: Assegura que no té por d'anar sense CDC als comicis del setembre. Línia similar: El programa electoral, passades les eleccions múnicipals. Valor de la distància entre les frases: 9.447585
5	Valor únic: 27461 Summary: CDC revela que el decret de convocatòria serà ordinari per evitar un recurs de Rajoy. Línia similar: El caràcter plebiscitari el donen les forces polítiques", subratlla Turull. Valor de la distància entre les frases: 9.534200
6	Valor únic: 2965 Summary: Fonts de la direcció justifiquen el cessament perquè tindrà un rol en la negociació amb el PSOE. Línia similar: El nou càrrec es crearà dijous i Echenique estrenarà les noves responsabilitats en una primera reunió que tindrà lloc dissabte, quan el partit celebrarà el seu Consell Ciutadà Estatal (CCE). Valor de la distància entre les frases: 9.539958
7	Valor únic: 52662 Summary: El diari assegura que les van cobrar a través de dos contractes d'un total de 12 milions d'euros per organitzar les campanyes electorals del PP per a les autonòmiques del 2003 i les generals del 2004. Línia similar: El diari assenyala que les sigles P.A.C. Valor de la distància entre les frases: 9.564263
8	Valor únic: 13397 Summary: Puigdemont ha reclamat un últim intent per buscar una proposta àmplia amb la societat civil. Línia similar: El partit no aprovarà finalment dissabte les seves llistes. Valor de la distància entre les frases: 9.584255
9	Valor únic: 32827 Summary: Després de la recuperació del çlima d'acord", les formacions polítiques es posicionen en la nova fase del procés sobiranista. Iceta titlla de "nou fracàs"el pacte Mas-Junqueras. Línia similar: Les formacions valoren el nou horitzó electoral anunciat per Mas el 27 de setembre d'enguany. Valor de la distància entre les frases: 9.707771
10	Valor únic: 25249 Summary: Rivera coincideix amb el PP a voler desinflar el procés amb pressió sobre els funcionaris i asfíxia econòmica. Línia similar: "El millor antídot contra el sobiranisme és que Ciutadans guanyi les eleccions". Valor de la distància entre les frases: 9.764372

Taula B.37: Resultats consulta 10 model Word2Vec.

B.3.11. Consulta 11: La mesa electoral va cometre dos errors

Rànquing	Descripció Resultat
1	Valor únic: 34596 Summary: El líder d'ERC serà president de taula d'una mesa electoral de l'Institut Frederic Montpou de Sant Vicenç dels Horts. Línia similar: President de mesa electora. Valor de la distància entre les frases: 8.692840
2	Valor únic: 22803 Summary: El secretari general de Sortu ha vingut a Barcelona a aprendre del procés català, que ell voldria importar a Euskadi. Línia similar: Vam cometre diversos errors. Valor de la distància entre les frases: 9.513745
3	Valor únic: 68154 Summary: D'aquí uns mesos es commemoraran vint-i-cinc anys de la caiguda definitiva del Teló d'Acer, que es va produir amb l'extinció de la Unió Soviètica, però algun dels seus hereus continua mantenint intacte aquest teló de cara al món. Línia similar: Gran error. Valor de la distància entre les frases: 9.649062
4	Valor únic: 55494 Summary: El dirigent republicà assegura en el seu bloc que no presentarà una candidatura pròpia per "evitar la imatge de lluita interna dins del partit". Línia similar: D'aquesta manera, el dirigent, diputat a Madrid, deixa en mans de la nova executiva la seva reelecció com a candidat a les generals. Valor de la distància entre les frases: 10.006902
5	Valor únic: 28692 Summary: La coalició ha volgut fer una çrida"a la calma i ha assegurat que si depèn d'ells hi haurà un govern de canvi com reclama la ciutadania. Línia similar: Acord de la mesa de les Cort. Valor de la distància entre les frases: 10.014405
6	Valor únic: 23769 Summary: 'El País' l'acusa d'haver enganyat els militants distraient-los amb el dret a decidir, mentre que 'El Mundo' creu que el seu primer discurs va ser çaspós i anacrònic". Línia similar: "Àlvarez cometrà un greu error si introdueix factors de confrontació territorial aliens a la gestió sindical", afirma l'editorial del diari de Prisa. Valor de la distància entre les frases: 10.079089
7	Valor únic: 28943 Summary: L'esquerra valenciana assumeix cinc eixos programàtics bàsics però difereix sobre qui ha de tenir la presidència. Línia similar: Acord per a la mesa de les Cort. Valor de la distància entre les frases: 10.081998
8	Valor únic: 53621 Summary: El líder abertzale reconeix en una entrevista a la Cadena SER que es va sentir alleujat quan va conèixer el cessament de la violència, fa l'ullet al PP i creu que ETA i l'esquerra abertzale han de "reconèixer i reparar les víctimes del terrorisme". Línia similar: Crec sincerament que vam cometre un error". Valor de la distància entre les frases: 10.082339
9	Valor únic: 30143 Summary: L'ex número 3 de Podem vol "empènyer amb més força" des de fora. Línia similar: El Pablo com a secretari general a la mesa executiva. Valor de la distància entre les frases: 10.114677
10	Valor únic: 24591 Summary: El diputat de Junts pel Sí assegura que Catalunya està inventant una nova manera d'arribar a la independència. Línia similar: Però la CUP també va cometre un error. Valor de la distància entre les frases: 10.183270

Taula B.38: Resultats consulta 11 model STSB.

Rànquing	Descripció Resultat
1	Valor únic: 68064 Summary: Els conservadors acceleren el procés d'elecció de nou 'premier' i el laborisme ja viu una guerra civil oberta. Línia similar: La sessió parlamentària d'ahir va tenir dos clars protagonistes. Valor de la distància entre les frases: 10.714993
2	Valor únic: 21456 Summary: El PP pensa en una investidura a l'octubre facilitada pel desgast i el PSOE vol cremar etapes. Línia similar: La mateixa data va triar Alberto Núñez Feijóo, que va descol·locar la fragmentada oposició gallega. Valor de la distància entre les frases: 10.838203
3	Valor únic: 53198 Summary: El president de la Generalitat avisa des de Madrid que "el pacte fiscal és una de les últimes oportunitats per refer el divorci creixent entre la societat catalana i les institucions espanyoles". Línia similar: La Moncloa va mostrar-se ahir satisfeta d'aquesta fotografia conjunta amb CiU. Valor de la distància entre les frases: 11.043422
4	Valor únic: 86422 Summary: La fiscalia i l'acusació particular li rebaixen la pena perquè estava borratxo. Línia similar: La víctima va interposar denúncia dos dies després. Valor de la distància entre les frases: 11.121744
5	Valor únic: 33408 Summary: Prepara una via legal per captar els tècnics tributaris a qui l'Estat posa traves per marxar. Línia similar: Segons va publicar ahir La Vanguardia, la comissió delegada d'assumptes econòmics es va reunir dijous per abordar la situació. Valor de la distància entre les frases: 11.450391
6	Valor únic: 8829 Summary: Els republicans acusen la llista del president de "mentir" i afirmen que la confiança queda tocada. Línia similar: La formació va votar dividida perquè l'altre membre de la mesa, Eusebi Campdepadrós, va abstenir-se. Valor de la distància entre les frases: 11.707012
7	Valor únic: 67371 Summary: La candidata domina un aspre cara a cara presidencial i col·loca Donald Trump a la defensiva. Línia similar: La resposta de Trump va ser ràpida: "Publicaré la declaració quan la secretària Clinton publiqui els 33.000 correus que va esborrar". Valor de la distància entre les frases: 11.724539
8	Valor únic: 47024 Summary: El jutge ha condemnat dos manifestants acusats d'agredir els agents de l'autoritat i ha absolt tres policies involucrats en els incidents ocorreguts al barri valencià l'any 2010. Línia similar: La diputada Mònica Oltra va tractar d'aturar els enderrocs. Valor de la distància entre les frases: 11.747499
9	Valor únic: 84859 Summary: Nieto explota en la comissió d'investigació dels atemptats del 17-A al Parlament. Línia similar: La d'ahir va ser la quinzena sessió d'aquesta comissió. Valor de la distància entre les frases: 11.750795
10	Valor únic: 18429 Summary: La del cas Vidal i la de l'operació Catalunya se sumen a la vintena que s'han creat des del 1980. Línia similar: La comissió d'investigació va acordar reprovar l'expresiden. Valor de la distància entre les frases: 11.804725

Taula B.39: Resultats consulta 11 model Word2Vec.

B.3.12. Consulta 12: Els independentistes no accepten el tracte oferit pel Governs

Rànquing	Descripció Resultat
1	Valor únic: Valor únic: 63121 Summary: El govern, l'únic importador, en dificulta l'accés als mitjans crítics. Línia similar: El govern acostuma a optar pel silenci administratiu: ni accepta ni denega els permisos, es limita a no donar-hi resposta. Valor de la distància entre les frases: 9.759110
2	Valor únic: 3918 Summary: El president diu "Ho tornarem a fer", sense concretar si vol fer un altre referèndum unilateral. Línia similar: S'ha ofert per ser soci "estable" del PSOE si accepta el referèndum com una solució al diàleg i així poder-lo concretar en els "propers anys". Valor de la distància entre les frases: 9.823105
3	Valor únic: 16197 Summary: La gran diferència continua sent la resposta al Procés, però el líder de Podem "apreciala defensa de la plurinacionalitat del PSOE. Línia similar: "Demano que torni la sensatesa al PSOE", ha afegit, i ha recordat que Sánchez "no té prou vots per poder-ho fer [governar] sense incloure cessions als independentistes i a Bildu". Valor de la distància entre les frases: 9.887424
4	Valor únic: 2227 Summary: Alguns experts consultats apunten que el PP es podria abstenir per facilitar la investidura. Línia similar: Si no s'arrisca a nous comicis, el PSOE pot intentar per segona vegada un pacte similar o un vot a favor d'Unides Podem sense entrar a l'executiu. Valor de la distància entre les frases: 9.912768
5	Valor únic: 5423 Summary: Assegura que Sánchez buscarà el sí d'ERC i el PDECat als pressupostos "fins a l'últim moment". Línia similar: "Però no només depèn d'ell, i l'independentisme s'ha de preguntar si li interessa afeblir aquest govern", ha exposat. Valor de la distància entre les frases: 10.016506
6	Valor únic: 22557 Summary: La reunió entre el grup parlamentari i el secretariat acaba sense consens i només si tres assemblees territorials ho demanen la formació revisarà l'esmena a la totalitat. Línia similar: Fins i tot si això passés al llarg d'aquest cap de setmana, no està gens clar que les bases acabin votant a favor de retirar l'esmena a la totalitat i acceptar la proposta del Govern. Valor de la distància entre les frases: 10.024039
7	Valor únic: 3266 Summary: Meritxell Batet és la nova presidenta d'un òrgan que té membres del PSOE, el PP, Cs i Podem. Línia similar: Aquest dimarts hauria estat escollit vicepresident primer si els independentistes no haguessin votat a favor de Gloria Elizo, d'Unides Podem. Valor de la distància entre les frases: 10.077696
8	Valor únic: 7965 Summary: La direcció republicana aposta per la desobediència com a arma de negociació amb l'Estat pel referèndum. Línia similar: "Només serem independents si som prous per forçar un referèndum acceptat per l'Estat", afegeix una altra. Valor de la distància entre les frases: 10.114261
9	Valor únic: 23209 Summary: L'entitat es concentra a fer pedagogia del procés i trasllada a les institucions aplicar el mandat del 27-S. Línia similar: "Tenim socis independentistes, colauistes i partidaris del dret a decidir", apunta aquesta font, deixant clar que "sense retirar-se de l'independentisme" volen tornar a fer la funció de "tauler". Valor de la distància entre les frases: 10.274199
10	Valor únic: 32916 Summary: El PP i el PSOE s'obliden, per ara, de Catalunya i descarten eleccions en breu. Línia similar: "L'independentisme és, sobretot, insolidaritat i egoisme", va afegir, fugint del discurs més moderat amb el procés a Catalunya, que pel PSOE passa per una reforma de la Constitució. Valor de la distància entre les frases: 10.286675

Taula B.40: Resultats consulta 12 model STSB.

Rànquing	Descripció Resultat
1	Valor únic: 66202 Summary: També se li prohibeix representar el Parlament Europeu en reunions interparlamentàries, conferències o qualsevol tipus de fòrum durant un any. L'ultra Janusz Korwin-Mikke va defensar que les dones han de cobrar menys "perquè són més dèbils i menys intel·ligents". Línia similar: El Parlament Europeu investigarà l'eurodiputat polonès per les declaracions misògine. Valor de la distància entre les frases: 9.281645
2	Valor únic: 16373 Summary: El president d'ERC carrega contra l·l'irresponsable"govern espanyol, que actua "sistemàticament en contra"dels interessos de la ciutadania. Línia similar: El partit pretén que aquesta conferència nacional marqui les línies bàsiques que defensarà ERC durant el procés constituent. Valor de la distància entre les frases: 9.286292
3	Valor únic: 81053 Summary: El règim anuncia que seran "netes i justes i portaran a una Assemblea Popular que representi el poble sirià a través del pluralisme polític", segons l'agència de SANA. Línia similar: El ministre sirià d'Exteriors, Walid al-Mualem, ha assegurat avui que el seu país celebrarà eleccions legislatives abans de finals d'any i que les urnes actuaran "d'àrbitre"de les reformes anunciades pel president Baixar al-Assad. Valor de la distància entre les frases: 9.405838
4	Valor únic: 32571 Summary: Assegura que no té por d'anar sense CDC als comicis del setembre. Línia similar: El programa electoral, passades les eleccions múnicipals. Valor de la distància entre les frases: 9.447585
5	Valor únic: 27461 Summary: CDC revela que el decret de convocatòria serà ordinari per evitar un recurs de Rajoy. Línia similar: El caràcter plebiscitari el donen les forces polítiques", subratlla Turull. Valor de la distància entre les frases: 9.534200
6	Valor únic: 2965 Summary: Fonts de la direcció justifiquen el cessament perquè tindrà un rol en la negociació amb el PSOE. Línia similar: El nou càrrec es crearà dijous i Echenique estrenarà les noves responsabilitats en una primera reunió que tindrà lloc dissabte, quan el partit celebrarà el seu Consell Ciutadà Estatal (CCE). Valor de la distància entre les frases: 9.539958
7	Valor únic: 52662 Summary: El diari assegura que les van cobrar a través de dos contractes d'un total de 12 milions d'euros per organitzar les campanyes electorals del PP per a les autonòmiques del 2003 i les generals del 2004. Línia similar: El diari assenyala que les sigles P.A.C. Valor de la distància entre les frases: 9.564263
8	Valor únic: 13397 Summary: Puigdemont ha reclamat un últim intent per buscar una proposta àmplia amb la societat civil. Línia similar: El partit no aprovarà finalment dissabte les seves llistes. Valor de la distància entre les frases: 9.584255
9	Valor únic: 32827 Summary: Després de la recuperació del çlima d'acord", les formacions polítiques es posicionen en la nova fase del procés sobiranista. Iceta titlla de "nou fracàs"el pacte Mas-Junqueras. Línia similar: Les formacions valoren el nou horitzó electoral anunciat per Mas el 27 de setembre d'enguany. Valor de la distància entre les frases: 9.707771
10	Valor únic: 25249 Summary: Rivera coincideix amb el PP a voler desinflar el procés amb pressió sobre els funcionaris i asfíxia econòmica. Línia similar: "El millor antídot contra el sobiranisme és que Ciutadans guanyi les eleccions". Valor de la distància entre les frases: 9.764372

Taula B.41: Resultats consulta 12 model Word2Vec.

B.3.13. Consulta 13: El deute economic creix

Rànquing	Descripció Resultat
1	Valor únic: 18791 Summary: El delegat del govern espanyol assegura que la Moncloa segueix "de prop"els plans sobiranistes i que actuarà amb "fermesaï "proporcionalitat". Crida al "diàleg"entre administracions i carrega contra la consulta per separar entre "guanyadors i vençuts". Línia similar: Un procés que minva el çreixement"econòmic. Valor de la distància entre les frases: 9.509012
2	Valor únic: 85066 Summary: Després d'explorar l'orgasme femení a Mèxic, l'artista visual s'endinsa en el pensament amorós a Sèrbia. Línia similar: El capitalisme l'explota. Valor de la distància entre les frases: 9.607563
3	Valor únic: 56443 Summary: L'expresident del Partit Popular de Catalunya denuncia que "es pretén que només el País Basc i Navarra tinguin un tracte preferencial"pel que fa al pacte fiscal. Línia similar: Els problemes que frenen el creixement econòmic són, segons Piqué, les subvencions, "que acaben ofegant la iniciativa", i l'atur juvenil, "que és una pèrdua potencial per a l'economia". Valor de la distància entre les frases: 9.830784
4	Valor únic: 76985 Summary: Molts apostaven per un divorci, però el primer ministre britànic i el vice primer ministre han hagut d'adoptar decisions difícils "necessàries i adequades"–segons el 'premier'– per redreçar l'economia del Regne Unit. Línia similar: Un context econòmic difícil. Valor de la distància entre les frases: 10.038761
5	Valor únic: 30164 Summary: A 'El País': "Alemanya va ajudar els EUA en l'espionatge als líders europeus". Línia similar: ARA: "¿Es pot crear ocupació sense creixement econòmic? Valor de la distància entre les frases: 10.059344
6	Valor únic: 31715 Summary: L'alcaldable del PSC creu que no es compliran els resultats tan negatius que mostren les enquestes. Línia similar: Les dificultats del PSC ¿les atribueix a l'esgotament del partit, a la crisi del socialisme, a l'econòmica. Valor de la distància entre les frases: 10.076995
7	Valor únic: 52344 Summary: La renda inferior se suma a un finançament mal resolt a València, mentre que les Illes registren el greuge fiscal relatiu més gran de tot l'Estat. Línia similar: José Ignacio Aguiló, vicepresident econòmic del govern Balear, no està content amb el model anterior, pactat el 2009 pels socialistes, perquè fa que l'executiu espanyol "tingui la paella pel mànec". Valor de la distància entre les frases: 10.147810
8	Valor únic: 55572 Summary: L'exconseller d'Economia trenca un silenci de mig any per defensar la seva gestió i es queixa de "falsedats"que han difós alguns representants de CiU. Lamenta que Madrid fomenti visions pròpies de qui voldria que l·l'estat de les autonomies fos un parèntesi en la història d'Espanya". Línia similar: Davant les mancances pròpies d'un país de la Unió Europea –no control de la política monetària–, Castells ha apostat per un "ajustament del dèficit", però "gradual", i que l·la UE porti a terme polítiques per fer créixer l'economia". Valor de la distància entre les frases: 10.157731
9	Valor únic: 59127 Summary: Els comptes del país surten de la zona de perill però la millora no arriba a la butxaca de la població. Línia similar: Però Cerejeira alerta de la precarietat a l'alça, perquè la nova ocupació "ofereix contractes temporals i el sou mitjà és molt més baix que a Europa", diu l'economista. Valor de la distància entre les frases: 10.165856
10	Valor únic: 62474 Summary: El sociòleg portuguès defensa que l'acadèmia ha d'incorporar els coneixements creats en les lluites populars. Línia similar: Perquè van seguir el mateix model econòmic neoliberal i extractivista, amb la diferència que feien una mica més de distribució de la riquesa. Valor de la distància entre les frases: 10.600260

Taula B.42: Resultats consulta 13 model STSB.

Rànquing	Descripció Resultat
1	Valor únic: 69668 Summary: El triomf d'un 'outsider' evidencia el vot insatisfet pels polítics convencionals, a un any de les eleccions presidencials dels Estats Units. Línia similar: El multimilionari novaiorquès n'obté el 22%. Valor de la distància entre les frases: 15.430121
2	Valor únic: 24700 Summary: El PSOE proposa Patxi López perquè la presideixi. Tindria el suport de C's, mentre que Podem manté com a condició els grups de les aliances territorials. Línia similar: El sistema d'elecció el beneficia. Valor de la distància entre les frases: 15.559031
3	Valor únic: 50696 Summary: En una entrevista al canal internacional des de Nova York, el president de la Generalitat reivindica que hi ha territoris del sud d'Europa, com Catalunya, que estan fent bé la feina per superar la situació crítica. Línia similar: El problema és el deute de l'Estat. Valor de la distància entre les frases: 15.559454
4	Valor únic: 32963 Summary: Rajoy segella a Andorra un acord amb condicions que no es posarà en marxa fins al 2018. Línia similar: "El problema català el solucionarà Espanya. Valor de la distància entre les frases: 15.756096
5	Valor únic: 86551 Summary: Un trasplantament de cèl·lules mare fa remetre el virus del VIH en un pacient i marca el camí de la recerca. Línia similar: El consorci SciStem arrenca el 2014. Valor de la distància entre les frases: 15.836631
6	Valor únic: 32178 Summary: Ramon Espadaler valora l'exercici de "transparència"del president de la Generalitat i insta Alícia Sánchez-Camacho a fer el mateix. Línia similar: El pacte d'Estat contra el jihadism. Valor de la distància entre les frases: 15.874846
7	Valor únic: 26216 Summary: Obama complau Rajoy i s'afegeix a Merkel i Cameron per presentar com a inviable i sense aliats una Catalunya independent. Línia similar: El primer diari estonià justifica el procés. Valor de la distància entre les frases: 15.898243
8	Valor únic: 2647 Summary: Els comuns acusen el Govern de "fer-li un favor a La Caixa", propietari de l'espai. Línia similar: El Govern sosté que el cost l'assumirà l'empresa. Valor de la distància entre les frases: 15.981304
9	Valor únic: 19766 Summary: Incrementa en 1.170 milions d'euros la despesa social, que és del 74,7% del total, 2,9 punts més que el 2015 i Junqueras diu que són els comptes de l·la fi de les retallades". Línia similar: El Govern presenta un pressupost de 28.310 milions, el 3,6% mes. Valor de la distància entre les frases: 16.067628
10	Valor únic: 78474 Summary: La cadena britànica ha obert el debat en la seva pàgina web després de l'escàndol que va provocar Rajoy en comparar els dos països. Les xifres i opinions recollides posen en dubte l'opinió del president espanyol. Línia similar: El creixement del PIB d'Espanya és negatiu, el d'Uganda creix un 5,2%". Valor de la distància entre les frases: 16.208808

Taula B.43: Resultats consulta 13 model Word2Vec.