UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Informática

Analizador de legibilidad de textos

Trabajo Fin de Grado

Grado en Ingeniería Informática

AUTOR/A: Santolaria Leiva, Guillermo

Tutor/a: Rebollo Pedruelo, Miguel

CURSO ACADÉMICO: 2021/2022

Resumen

La legibilidad es un elemento clave para la comprensión de un texto y puede utilizarse para crear textos más claros y fáciles de entender. A pesar de la gran importancia de este tema, existen pocas herramientas que permitan realizar un análisis complejo para textos en español. Por eso, en este trabajo he desarrollado una herramienta que calcula e interpreta ciertos valores relacionados con la legibilidad utilizando tanto los métodos tradicionales como el análisis de redes complejas. Además, incluye otros tipos de análisis, que permiten hacerse una idea de forma rápida del contenido temático del texto. Finalmente se ha verificado el funcionamiento de la aplicación y se usado para realizar unos experimentos relacionados con la legibilidad de los textos de comunicación.

Palabras clave: legibilidad, análisis de redes, algoritmos tradicionales.

Abstract

Readability is a key element for the comprehension of a text and can be used to create clearer and easier to understand texts. Despite the great importance of this topic, there are few tools that allow a complex analysis for Spanish texts. Therefore, in this work I have developed a tool that calculates and interprets certain values related to readability using both traditional methods and complex network analysis. In addition, it includes other types of analysis, which allow to get a quick idea of the thematic content of the text. Finally, the application has been verified and used to perform some experiments related to the readability of communication texts.

Keywords : readability, network analysis, traditional algorithms.

Tabla de contenidos

Lista de ilustraciones

Lista de tablas

Lista de ecuaciones

1. Introducción

La legibilidad se define en la RAE como la cualidad de lo que es legible y es, además, un elemento clave para la comprensión de un texto. (Ferrando Belart, 2004). Sin embargo, la lectura es un proceso complejo en el cual intervienen muchos factores. Por tanto, la legibilidad es un término que puede ser definido de distintas formas. Los dos tipos de legibilidad más estudiados han sido la legibilidad tipográfica y la lingüística. (Barrio Cantalejo & Simón Lorda, 2003). La legibilidad tipográfica, es aquella que se centra en aspectos como el estilo, color o tamaño de los caracteres que forman el texto, mientras que la legibilidad lingüística, es aquella que se focaliza en las construcciones gramaticales del texto

En el caso de esta tesis me centraré en la segunda definición descrita de la legibilidad.

La legibilidad de un texto es un problema muy importante cuyo interés no ha hecho más que aumentar en los últimos años, ya que, como enuncia la Constitución Española, cualquier individuo puede y debe ser capaz de "comunicar y recibir información veraz por cualquier medio de difusión". Este lenguaje debe ser comprensible para todo ciudadano. (Peñaranda Cortés, 2015). Esto es más cierto todavía en la época actual en la que, gracias a internet, tenemos acceso muy sencillo a gran cantidad de texto.

Uno de los principales campos en los que se ha usado este análisis ha sido en los textos médicos, en concreto a los textos dirigidos a los pacientes. La información debería ser fácil de entender pues cuánto más amplia se hace la brecha comunicativa que se establece entre esta información y el profesional sanitario, más crece la brecha entre el informe médico y su correcta asimilación. (Porras-Garzón & Estopà, 2020).

Aunque últimamente ha cobrado mucha importancia también en el campo del SEO (*Search Engine Optimization*), que consiste en un conjunto de acciones orientadas a mejorar el posicionamiento de un sitio web en la lista de resultados, ya que un texto claro y fácil de leer, invita a los usuarios a permanecer un mayor tiempo en tu página *web*.

Además de esto, el análisis de legibilidad se podría utilizar para permitir que los textos pudieran llegar a más gente con un nivel educativo menor, haciendo así accesible el contenido de estos para un número más amplio de personas y favoreciendo de esa forma la difusión del conocimiento y dando oportunidad a gente con menos recursos a poder optar a una educación de calidad.

En el presente trabajo crearé una herramienta que utilice tanto los algoritmos clásicos para medir la legibilidad en español como las técnicas más nuevas de procesamiento de lenguaje natural, en concreto el análisis de redes, para dar una mayor fiabilidad a las medidas tomadas y realizaré una verificación de la herramienta con unos conjuntos de textos seleccionados.

Finalmente, realizaré un experimento con distintos textos de los medios de comunicación para ver si existen diferencias en la legibilidad de distintas secciones de un mismo periódico digitales y también se comprobará entre la misma sección de distintos periódicos.

1.1. Motivación

La disponibilidad de una cantidad cada vez mayor de datos que ha traído consigo la era de la información ha tenido un fuerte impacto en la computación, dando lugar a novedosas perspectivas del análisis de datos y, más concretamente, del análisis del lenguaje natural, que permiten una observación más avanzada de los temas importantes en un texto determinado o estudiar las relaciones entre las palabras o los temas de un documento o un conjunto de documentos.

Considero que esto abre un gran número de posibilidades para estudiar en profundidad cómo la forma en que un texto está escrita influye en los potenciales lectores perciben el contenido, pudiendo hacer que conceptos más complicados se hagan más fáciles de entender o pudiendo atraer más fácilmente la atención de posibles lectores, siendo esto de especial importancia en la época actual en la que el acceso a gran cantidad de información y texto es tan inmediato.

Todos estos nuevos métodos pueden unirse a los métodos tradicionales de cálculo de legibilidad basados en la estadística y el conteo de palabras, frases o sílabas para crear una herramienta que proporcione un análisis más completo de los textos en español.

1.2. Objetivos

La presente tesis tiene como objetivo aplicar los métodos del procesamiento del lenguaje natural a la creación de una herramienta, y estos objetivos se dividen en:

- Estudiar de los métodos a utilizar en el análisis de los textos: algoritmos de legibilidad, procesos estadísticos, creación de redes…
- Desarrollar e implementar una aplicación de escritorio con una interfaz gráfica que permita introducir texto y mostrar un análisis de este devolviendo resultados y la interpretación de estos.
- Interpretar los resultados obtenidos con los algoritmos clásicos e interpretar la red generada a partir de los textos para obtener un análisis más completo.
- Realizar el experimento basado en la comparación de distintas secciones de un periódico explicado anteriormente e interpretar los resultados.

Por tanto, la estructura del trabajo estará dividida entorno a estos cuatro grandes objetivos.

1.3. Objetivos de Desarrollo Sostenible

En cuanto a los objetivos de desarrollo sostenible propuestos por la UPV, esta herramienta podría ayudar con varios de ellos.

En primer lugar, y como ya he comentado anteriormente, esta se podría utilizar para facilitar la comprensión de los textos científicos, haciendo que el contenido de estos fuera accesible a una mayor cantidad de público con distintos niveles de educación.

En segundo lugar, esta se podría utilizar para evitar las *fake news*, ya que proporciona una visión global del texto además del propio análisis de legibilidad.

Por último, podría utilizarse para simplificar otro tipo de textos que tienen como objetivo un público general, para hacer más accesible para todo el mundo textos como el de las prescripciones médicas.

Objetivos de Desarrollo Sostenibles	**Alto**	**Medio**	**Bajo**	**No Procede**
ODS 1. **Fin de la pobreza.**				**x**
ODS 2. **Hambre cero.**				**x**
ODS 3. **Salud y bienestar.**		**x**		
ODS 4. **Educación de calidad.**	**x**			
ODS 5. **Igualdad de género.**		**x**		
ODS 6. **Agua limpia y saneamiento.**				**x**
ODS 7. **Energía asequible y no contaminante.**		**x**		
ODS 8. **Trabajo decente y crecimiento económico.**				**x**
ODS 9. **Industria, innovación e infraestructuras.**				**x**
ODS 10. **Reducción de las desigualdades**.			**x**	
ODS 11. **Ciudades y comunidades sostenibles**.				**x**
ODS 12. **Producción y consumo responsables.**				**x**
ODS 13. **Acción por el clima.**		**x**		
ODS 14. **Vida submarina.**			**x**	
ODS 15. **Vida de ecosistemas terrestres.**			**x**	
ODS 16. **Paz, justicia e instituciones sólidas.**	**x**			
ODS 17. **Alianzas para lograr objetivos.**				**x**

- ODS 4. GARANTIZAR UNA EDUCACIÓN INCLUSIVA, EQUITATIVA Y DE CALIDAD Y PROMOVER OPORTUNIDADES DE APRENDIZAJE DURANTE TODA LA VIDA PARA TODOS:

Unos textos con mayor legibilidad y por tanto más comprensibles y sencillos de entender podrían aumentar la tasa de rendimiento discente de alumnos/as y, potencialmente, reducir la tasa de abandono de estudios de alumnos/as.

Además, la aplicación ayudaría a cumplir la meta 4: aumentar considerablemente el número de jóvenes y adultos que tienen las competencias necesarias, en particular técnicas y profesionales, para acceder al empleo, el trabajo decente y el emprendimiento de aquí a 2030, ya que unos textos más sencillos facilitarían la adquisición de esos conocimientos.

- ODS 16. PROMOVER SOCIEDADES JUSTAS, PACÍFICAS E INCLUSIVAS.

La meta 3 de este ods que busca promover el estado de derecho en los planos nacional e internacional y garantizar la igualdad de acceso a la justicia para todos y la meta b que busca promover y aplicar leyes y políticas no discriminatorias en favor del desarrollo sostenible, se beneficiarían se la herramienta, ya que unos textos legales más sencillos y accesibles para más gente facilitarían poder alcanzar ese objetivo.

- El enfoque de evitar las *fake news* de la aplicación podría usarse para ayudar en la consecución se distintos ods que sufren mucho de este tipo de problemas que extienden la desinformación. Esto es algo a tener en cuenta en el campo de la salud y más con lo que hemos vivido estos años con el Covid y toda la desinformación que se extendió con este, por tanto, se favorecería a la consecución de los objetivos del ods 3: Garantizar una vida sana y promover el bienestar para todos en todas las edades.

  Otros campos donde esto también es importante son el de la igualdad de género, que ha cobrado especial importancia este mes de junio, el de las energías que una vez más es muy sonado en la actualidad, debido a las subidas en el precio de la luz y la gasolina que estamos viviendo y por último en el campo de la acción por el clima y en menor medida también en el mantenimiento y cuidado de la vida submarina y los ecosistemas terrestres.

  Por tanto, también influiría en los ods 5, 7, 13 y en menos medida en los 14 y 15.

Dada la importancia de todos los temas mencionados, la realización y prueba de esta herramienta se enfocará en que pueda ser capaz de ayudar en la consecución de estos objetivos.

2. Estado del Arte

Los orígenes de la legibilidad se remontan al año 900 a.C., donde los estudiosos del Talmud contaban las palabras y conceptos de la Torá para determinar cuántas veces y con qué rango de frecuencia aparecía cualquier palabra inusual. (Blanco, 2004).

En 1917, Thorndike expuso las conclusiones de diversas pruebas comprobadas sobre grupos de alumnos. Estas pruebas probaron que el conocimiento semántico de las palabras durante la lectura no implicaba la comprensibilidad de los párrafos o frases formadas por esas palabras. Para Thorndike, la comprensibilidad en la lectura es bastante equivalente al razonamiento en las matemáticas. Por ello, el orden de las palabras, su correspondencia y la dinámica de la frase determinan cómo de fácil es de comprender el texto. Así, niños que al parecer saben leer no comprenden hasta que aprenden a razonar simultáneamente con las palabras que leen en los textos de estudio (Blanco, 2004).

Así, la predicción de la legibilidad de los textos comenzó a ser un campo de estudio de creciente importancia, ya que todo aquel medio de comunicación que utilice textos escritos (como la prensa, revistas, folletos, programas…) puede utilizar estos estudios. Incluso se han efectuado acercamientos al mundo del lenguaje oral a partir de estas técnicas. La radio y la televisión también se han visto favorecidas, a veces, con estas predicciones (Ríos, 2009).

En consecuencia, se desarrollaron fórmulas primitivas de legibilidad para determinar lo que parecía legible. Estas son más rápidas de calcular que otras medidas más precisas de complejidad sintáctica y semántica ya que están basadas en el conteo de palabras, oraciones y sílabas. En muchos casos, se utilizan para estimar el nivel de grado correspondiente al texto.

En lo relativo a la lengua inglesa, Rudolf Flesch fue un pionero de las fórmulas de legibilidad y es uno de los autores más influyentes en este tema. *Reading Ease Score* que fue diseñada por él, es considerada como una de las fórmulas más conocidas del ámbito de la legibilidad. (Szigriszt Pazos, 1993). Además, participó en la creación del nivel de lectura Flesch-Kincaid (F-K) que fue desarrollado en 1975 por J. Peter Kincaid y su equipo, bajo contrato con la Marina de Estados Unidos (Wikipedia W. , 2022).

La fórmula *Reading Ease Score,* se calcula realizando operaciones con el total de palabras, el total de frases y el total de sílabas con una serie de constante calculadas previamente. En la ecuación 1 se puede ver la fórmula.

$$206.835 - 1.015\left[\frac{total\ de\ palabras}{total\ de\ frases}\right] - 84.6\left[\frac{total\ de\ sílabas}{total\ de\ palabras}\right]$$

Ecuación 1: Reading Ease Score

La tabla 1 muestra el resultado de aplicar la anterior fórmula, a partir de la cual se puede extraer la supuesta audiencia a la cual va dirigida el texto.

Puntuaje	Nivel	Tipos de Revista	Supuesta Audiencia
Hasta 1	Muy Fácil	Cómicas	4to grado
1 a 2	Fácil	Revista de quioscos	5to grado
2 a 3	Ligeramente fácil	Policíaca, aventuras	6to grado
3 a 4	Estándar	Resúmenes	7mo/8vo grado
4 a 5	Algo árido	Selecto	Institutos
5 a 6	Duro	Académico	Universidad
6 o más	Muy duro	Científico	Titulados

Tabla 1: Interpretación de la Reading Ease Score traducida al español.

En cuanto a la lengua española, la primera fórmula para calcular la legibilidad de un texto fue presentada en 1956 por Seth Spauldin. (Spaulding, 1956)

Pero no fue hasta 1992 cuando Francis Szigriszt Pazos, presentó su tesis donde realizó una verdadera validación de la fórmula RES al castellano, modificando las constantes de la fórmula y nombrándola Fórmula de Perspicuidad. Además de adaptar la escala de interpretación de Flesch. (Szigriszt Pazos, 1993)

La fórmula es la que se puede ver en la ecuación 2 donde S, el total de sílabas; P, la cantidad de palabras; F, el número de frases.

$$206.835 - \frac{62.3S}{P} - \frac{P}{F}$$

Ecuación 2: Fórmula de Perspicuidad

Igual que en el caso anterior, la tabla 2 sirve para interpretar el resultado obtenido con la fórmula anterior.

Puntos	Estilo	Palabras por frase	Número de sílabas cada 100 palabras	Estudios
Hasta 1	Muy Difícil	29	261	Universitarios
1 a 2	Difícil	23	230	Secundarios
2 a 3	Bastante Difícil	21	210	Secundarios incompletos
3 a 4	Medio	18	199	7mo/8vo grado
4 a 5	Bastante fácil	14	189	6to grado
5 a 6	Fácil	11	178	5to grado
6 o más	Muy fácil	8	166	4to grado

Tabla 2: Interpretación de la Perspicuidad

Existen una gran variedad de algoritmos de este tipo que miden distintas características del texto, como el *gunning fog,* que tiene en cuenta el número de

palabras largas por palabra o la Legibilidad μ que tiene en cuenta la media del número de letras por palabra o su varianza.

En inglés existen varias páginas web como *web.fx* o *readable.com* que utilizan algoritmos para ofrecer un análisis de la legibilidad. En cuanto a la lengua en español existe la aplicación INFLESZ que está enfocada principalmente a textos del ámbito de la medicina o la página *legible.es*.

Existen además ciertos modelos estadísticos que permiten extraer información del texto, como por ejemplo la ley de *Zipf*, según la cual, en una determinada lengua, la frecuencia de aparición de distintas palabras sigue una distribución que puede aproximarse por

$$Pn \sim 1/n^a$$

Ecuación 3: Ley de Zipf

donde Pn representa la frecuencia de la n-ésima palabra más frecuente y el exponente a es un número real positivo, en general ligeramente superior a 1.1 Esto significa que el segundo elemento se repetirá aproximadamente con una frecuencia de 1⁄2 de la del primero, el tercer elemento con una frecuencia de 1⁄3 del primero y así sucesivamente. Esta ley se cumple para la mayoría de las lenguas. (Ley de Zipf, Wikipedia)

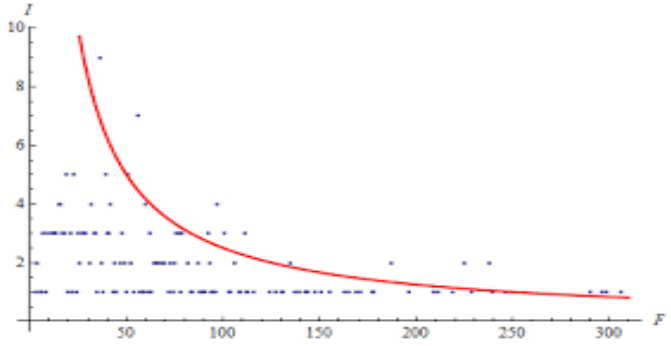


Ilustración 1: Aplicación de las Leyes de Zipf en los textos españoles - Wikipedia

Al representarse se obtiene un gráfico similar al de la ilustración 1.

Se han ideado métodos más sofisticados para análisis más complejos, como cuantificar la relevancia de las palabras en un documento (Hotho, 2005) o usar modelos para rastrear los temas de interés actuales e identificar las tendencias emergentes en las bibliotecas digitales en línea y la literatura científica (Alsumait, 2008).

Sin embargo, ninguno de estos parecía predecir la legibilidad del texto tan bien como las fórmulas de legibilidad mucho más sencillas mencionadas anteriormente (Todirascu, 2016).

Uno de los primeros enfoques de clasificación de la legibilidad basados en un clasificador SVM fue propuesto por Schwarm y Ostendorf (Schwarm & Ostendorf, 2005). Se entrenó en un corpus de *WeeklyReader*, que contiene artículos agrupados en cuatro clases según la edad del público objetivo. En el modelo se utilizan características tradicionales, sintácticas y de modelo lingüístico. Este enfoque fue ampliado y mejorado por Petersen y Ostendorf (Petersen & Ostendorf, 2009).

En un enfoque propuesto por Vajjala y Lučić, se utilizaron 155 características tradicionales, de cohesión del discurso, léxico-semánticas (Vajjala & Lučić., 2018), probado en un corpus *OneStopEnglish* y el clasificador de optimización mínima secuencial (SMO) con el núcleo lineal logró una precisión de clasificación del 78,13% para tres clases de legibilidad (nivel de lectura elemental, intermedio y avanzado).

Más tarde, Madrazo Azpiazu y Soledad Pera propusieron una red neuronal con el mecanismo de atención (Madrazo Azpiazu & Soledad Pera, 2019). Su sistema, llamado Vec2Read, es una RNN multiatención capaz de aprovechar las estructuras jerárquicas del texto con la ayuda de mecanismos de atención a nivel de palabra y de frase, y un mecanismo de agregación construido a medida. Utilizaron la red en un entorno multilingüe, en corpus que contenían textos en euskera, catalán, holandés, inglés, francés, italiano y español. Aunque la complejidad de este experimento supera la necesidad de explicarlo, su conclusión fue que, aunque el número de instancias utilizadas para el entrenamiento tiene un fuerte efecto en el rendimiento general del sistema, no surgieron patrones específicos de cada idioma que indicaran que la predicción de la legibilidad en algunos idiomas es más difícil que en otros. (Martinc, Pollak, & Robnik-Šikonja, 2021)

Todos los métodos descritos hasta el momento tienen en común que analizan los textos desde un punto de vista léxico y morfológico, sin tener en cuenta lo que se conoce como la escala mesoscópica del texto. En física y química, la escala mesoscópica se refiere a la escala de longitud en la que se puede discutir razonablemente las propiedades de un material o fenómeno, sin tener que discutir el comportamiento de los átomos individuales (Wikipedia, 2020). Podemos aplicar ese concepto al análisis de textos y decir que la estructura mesoscópica de un texto representa la estructura temática general del texto, sin tener en cuenta la sintaxis y los elementos individuales de este.

Un método que se utiliza actualmente es el análisis de redes, que se refiere a una familia de métodos que describen las relaciones entre unidades de análisis. Una red se compone de nodos y de aristas o conexiones entre ellos. En el análisis de textos, los nodos de la red son las palabras que aparecen en este y las aristas indican las relaciones entre las palabras y existen un nuevo conjunto de técnicas para crear redes que tienen en cuenta la estructura mesoscópica de los textos. Estas redes se generan conectando palabras que existen en el mismo contexto, que se define en términos de una longitud de ventana fija (por ejemplo, una oración). Este enfoque fue capaz de producir redes modulares, con cada comunidad relacionada con temas contextuales o subtemas del texto (F. de Arruda, da F. Costa, & R. Amancio, 2016).

2.1. Crítica al Estado del Arte

Las aplicaciones para medir la legibilidad de los textos en español son en general muy simples ya que solo incluyen medidas basadas en los algoritmos tradicionales y los métodos estadísticos y no utilizan los nuevos métodos basados en análisis de redes. En la tabla 3 se puede ver la caracterización de los métodos comentados:

Método	Dependencia del lenguaje	Enfoque
Algoritmos tradicionales	Sí	Sintáctico, léxico
Métodos estadísticos	No	Sintáctico, léxico
Métodos basados en SVM	Sí	Sintáctico, léxico
Métodos basados en RNN	No	Sintáctico, léxico
Métodos basados en análisis de redes	No	Mesoscópico, semántico

Tabla 3: Tabla con la caracterización de los métodos comentados.

De los artículos anteriores y las aplicaciones existentes se pueden realizar las siguientes críticas:

1. Falla en tener en cuenta la estructura mesoscópica del texto ya que los métodos utilizados solo tienen en cuenta la estructura sintáctica y léxica del lenguaje, dejando de lado la estructura semántica y temática del texto. Esto se debe a que analizan principalmente elementos del texto como el número de palabras, frases o sílabas, que no aportan ninguna información sobre el contenido real de este.

2. No ofrecen al usuario información que le sirva para hacerse una idea del contenido del texto.

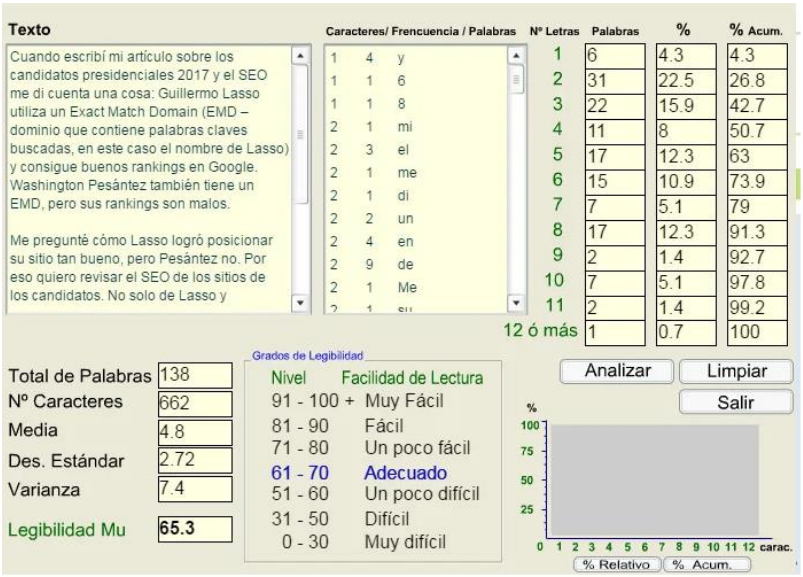


Ilustración 2:Análisis de la aplicación Legibilidad Mu

Ilustración 3: Análisis de Legibilidad de Word

Legibilidad del texto		
índice	**valor**	**dificultad**
Fernández Huerta	79.07	bastante fácil
Gutiérrez	46.88	normal
Szigriszt-Pazos	74.77	bastante fácil
INFLESZ	74.77	bastante fácil
legibilidad µ	62.93	adecuado

Ilustración 4: Análisis de Legible.es

La ilustración 2 muestra un ejemplo de una aplicación en la que se calcula e interpreta la legibilidad mu de un texto en concentro y se devuelven algunos valores que tienen que ver con la frecuencia de aparición de ciertas palabras en el texto. Este análisis es muy pobre ya que solo realiza el cálculo de un algoritmo para hacer su análisis. La ilustración 3, muestra las estadísticas de legibilidad de Word y en este caso, ni siquiera las utiliza para calcular algún valor o las interpreta. Por último, la ilustración 4 muestra el mismo análisis, pero realizado por la página *legible.es*, que utiliza un número mayor de algoritmos, pero se queda ahí, fallando en los problemas comentados anteriormente.

2.2. Propuesta

La propuesta de este trabajo es crear una aplicación que reciba como input un texto cualquiera en español y devuelva un análisis de la legibilidad utilizando varias de los métodos nombrados anteriormente, de forma que se pueda capturar la mayor información posible del texto. En concreto, se implementarán los algoritmos tradicionales y se interpretarán, como ya se hace en las aplicaciones actuales. Pero a esto se le introducirán los métodos basados en análisis de redes, generando una red a partir del texto y obteniendo e interpretando distintos parámetros de esta para obtener un análisis más rico y completo. Además, se incluye también una visualización de nube de palabras de los nombres que más aparecen en el texto para dar una visión rápida y general del posible tema del texto y una tabla en la que se mostrarán los 10 nombres, verbos y adjetivos que más aparecen en el texto.

Por tanto, el aporte de esta aplicación será aunar las técnicas tradicionales con las nuevas y así poder comprobar si ambas coinciden o no para los distintos textos, además de proporcionar una herramienta que se puede usar para crear textos más claros y fáciles de leer y realizar el experimento para comparar la legibilidad de distintas secciones de un periódico.

3. Análisis del Problema

3.1. Identificación y análisis de soluciones posibles

Una vez identificada la necesidad y creada la propuesta, el siguiente paso es realizar un análisis del problema a afrontar, en este caso crear una aplicación con interfaz gráfica que permita poder realizar todo lo discutido anteriormente.

Para esto tuvieron que tomarse varias decisiones, siendo una de las primeras y más importantes la elección de la plataforma en la que se iba a crear esta aplicación.

En primer lugar, mi idea fue desarrollarla para la web, ya que esta hubiera sido la mejor opción en cuanto a compatibilidad entre sistemas operativos, y además habría permitido a los usuarios no tener que descargar la aplicación para poder utilizarla. No obstante, esto habría varios problemas ya que implicaría la necesidad de tener conexión a internet, lo que además supone un problema para el almacenamiento, ya que sería necesario un sistema de almacenamiento en la nube. Por no mencionar que la implantación de la aplicación hubiera sido más compleja que en el resto de las plataformas sin aportar, como hemos visto, ningún beneficio.

Al final se optó por crear una aplicación de escritorio que se ejecuta en Python, por lo que debería poder ejecutarse y ser compatible con distintos sistemas operativos. La gran cantidad de información distinta mostrada en pantalla dificultaría que fuera usada en un dispositivo móvil, pero en un futuro se podría adaptar para poder ser usada en móviles también. De esta forma, no es necesario disponer de conexión a internet y usando la del ordenador, no habría problemas de almacenamiento.

La tabla 4 muestra las distintas plataformas consideradas y las respectivas ventajas y desventajas de cada una:

Plataforma	Pros	Contras
Web	• Compatibilidad total entre distintos dispositivos • No es necesario descargar la aplicación	• Conexión a internet necesaria para utilizar la aplicación • Dificultad extra de desarrollo e implantación
Aplicación de escritorio	• No necesita conexión a internet para funcionar • Compatibilidad total entre distintos sistemas operativos.	• Necesidad de descargar la aplicación • No sería válida para dispositivos móviles
Aplicación Móvil	• No necesita conexión a internet para funcionar • Portabilidad y posibilidad de usarla en cualquier sitio	• Necesidad de descargar la aplicación • Dificultad de integración entre distintos sistemas operativos móviles (Android, IOS)

Tabla 4: Tabla comparativa de plataformas posibles

3.2. Especificación de Requisitos

En cuanto a la especificación de requisitos de la aplicación, esta es muy sencilla, ya que la interacción de los usuarios con la aplicación será muy fácil. El usuario únicamente deberá introducir mediante un selector de archivos un fichero de texto en el que se encuentre el texto a ser analizado. A continuación, el programa mostrará por pantalla el análisis.

Así, existen varios requisitos funcionales:

RF1. – Selección de fichero.

RF2. – Cálculo del número de palabras, frases y sílabas del texto.

RF3. – Cálculo de los algoritmos tradicionales de medición de complejidad del texto.

RF4. – Creación de la red.

RF5. – Cálculo de los parámetros necesarios de la red para el análisis.

RF6. – Creación de la interfaz.

RF1: Seleccionar un fichero.

Descripción	El usuario debe ser capaz de seleccionar un fichero (únicamente de texto) mediante un selector de archivos que dependerá del sistema operativo que esté utilizando. A continuación, se mostrará el resultado del análisis por pantalla.
Precondición	

RF2: Cálculo del número de palabras, frases y sílabas del texto.

Descripción	Una vez el usuario haya seleccionado el fichero, se extraerá el texto de este y se procesará, calculando del número de palabras, frases y sílabas del texto.
Precondición	El usuario debe haber seleccionado un fichero de texto.

RF3: Cálculo de los algoritmos tradicionales de medición de complejidad del texto.

Descripción	Se calculan los valores de los algoritmos tradicionales y se interpretan.
Precondición	Número de palabras, frases y sílabas del texto deben haber sido calculados.

RF4: Creación de la red.

Descripción	Se recorren las frases del texto y se crea la red.
Precondición	Se debe haber procesado el texto.

RF5: Cálculo de los parámetros necesarios de la red para el análisis.

Descripción	Se utilizan las técnicas necesarias para obtener los parámetros de la red.
Precondición	Debe haberse creado la red.

RF6: Creación de la interfaz.

Descripción	Se utiliza todo lo calculado anteriormente para generar la interfaz de usuario.
Precondición	Deben haberse calculado todos los valores anteriores.

Además, la aplicación debe cumplir los siguientes requisitos no funcionales:

RNF1 – Interfaz sencilla e intuitiva

Descripción	La interfaz de la aplicación debe ser lo suficientemente sencilla e intuitiva para que cualquier usuario nuevo pueda ser capaz de usarla sin problemas.

RNF2 – Interfaz re escalable

Descripción	La interfaz de la aplicación debe poder escalarse para ajustarse de forma correcta a la pantalla del monitor en el que se está ejecutando.

3.3. Prototipado

Antes de pasar a crear la interfaz final de la aplicación creé un prototipo sencillo para saber cómo estructurar la información de la aplicación en pantalla para que esta pudiera ser entendida de la forma más fácil y rápida posible por el

usuario. Para la realización de este prototipo se ha usado la página de *Lucidcharts*.

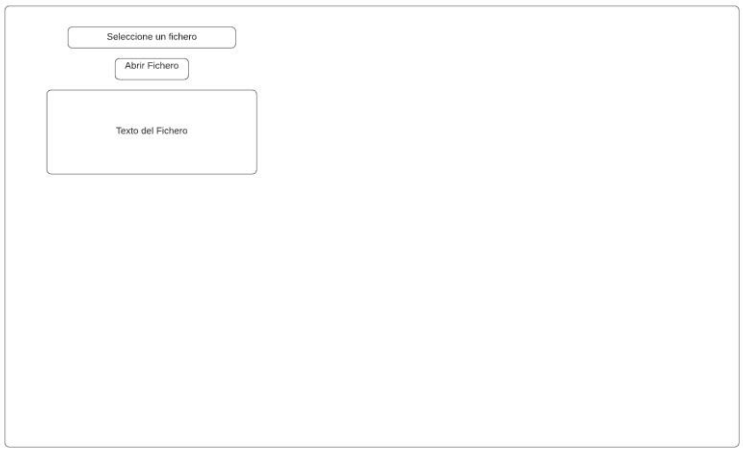


Ilustración 5: Prototipo de la interfaz inicial.

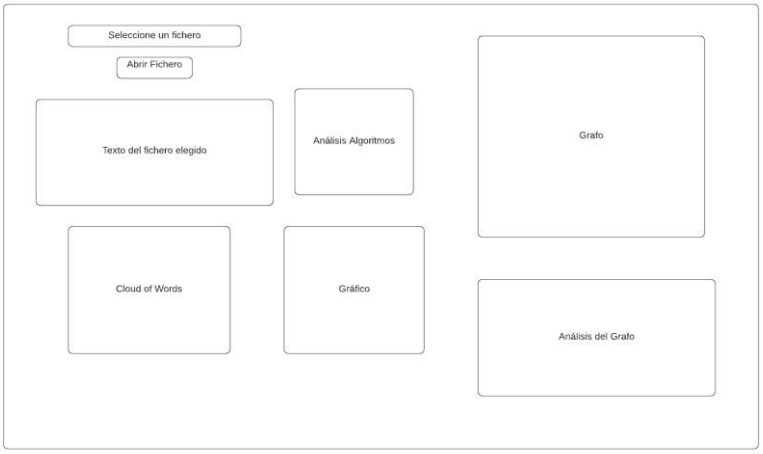


Ilustración 6: Prototipo de la interfaz después de seleccionar fichero

La ilustración 6 muestra la interfaz de la aplicación antes de que el usuario seleccione un fichero. En la ilustración 7 se ve la interfaz final después de haberlo elegido.

4. Diseño de la Solución

4.1. Arquitectura del Sistema

La arquitectura de este sistema es muy sencilla, como se puede ver en la ilustración 8, ya que no hay base de datos. Los datos se procesan directamente del fichero de texto y el resultado se calcula y se muestra al usuario dentro de la propia aplicación.

La ilustración 8 muestra un diagrama de bloques de la arquitectura general del sistema. En primer lugar, empezando desde abajo, tenemos el fichero *txt*, que introducirá el usuario mediante la interfaz gráfica, del que se extraerá el texto sobre el cual se producirá el análisis. Este texto se recibe en la capa de acceso de los datos, donde es enviado a la lógica del sistema. Aquí se aplicarán los cálculos necesarios para obtener todos los valores que se usarán para el análisis y se generarán los gráficos y las tablas que más tarde se le mostrarán al usuario a través de la interfaz de usuario.

El primer elemento de la arquitectura lo forma la interfaz de usuario, que es el elemento encargado de mostrar la información calculada, además de ser donde se encuentra el selector que este utilizará para elegir el fichero.

Los elementos 2,3,4,5 y 6 constituyen la lógica de la aplicación, que es donde se aplicarán todos los cálculos necesarios al texto para obtener las medidas, valores, tablas y gráficos que el usuario podrá visualizar. Esta es la parte más importante de la aplicación y la más compleja ya que incluye las numerosas funciones y librerías necesarias para el funcionamiento de la aplicación.

El elemento 7 es una capa intermedia entre los datos extraídos del fichero y la lógica del sistema. Esto sirve para que, si en un futuro se añade una nueva forma de introducir datos a la aplicación, ya sea mediante base de datos o introduciendo un enlace y descargando el texto de la web seleccionada, la lógica empresarial, que como he dicho anteriormente es la parte más complicada e importante de la aplicación, no se deba modificar.

Finalmente, el elemento 8 representa el fichero del que se extrae la información que se usará para realizar el análisis.

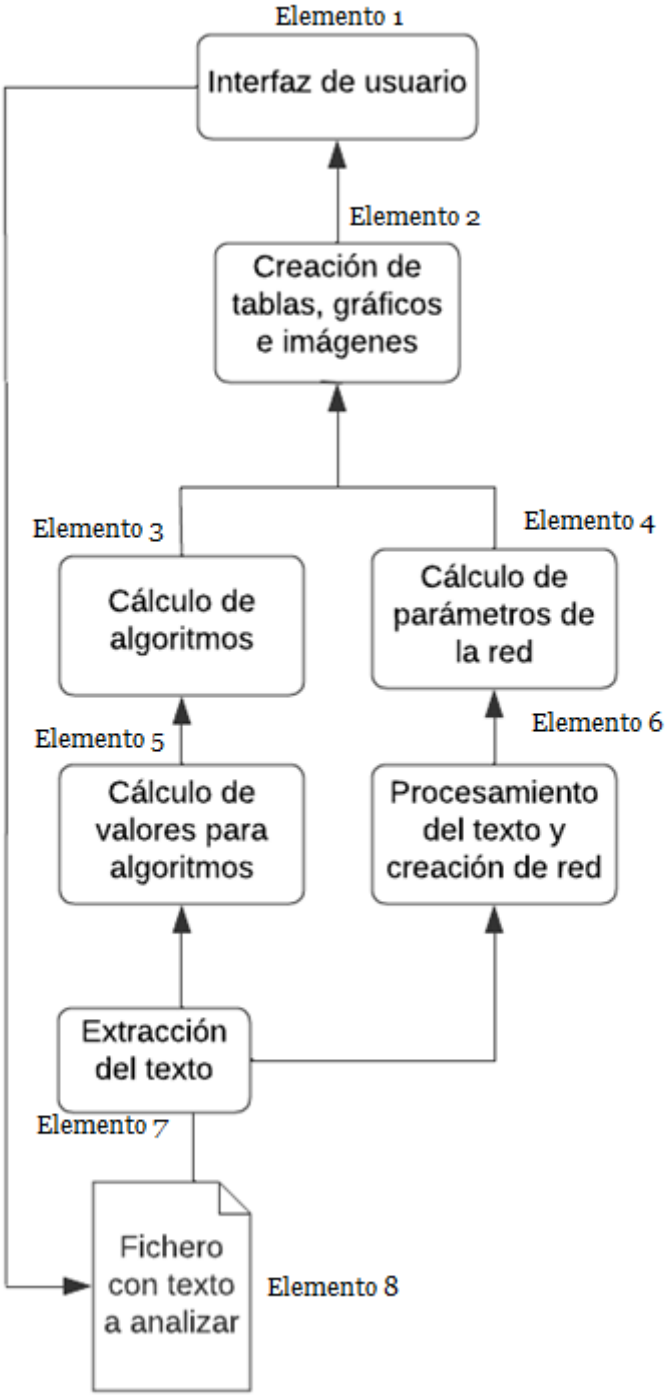


Ilustración 7: Arquitectura del Sistema

4.2. Tecnologías Utilizadas

Posteriormente, se nombrarán y explicarán las distintas tecnologías que se han utilizado para la creación de esta aplicación:

- *Python*

El lenguaje de programación elegido ha sido Python, ya que, debido a que es uno de los lenguajes más populares en la actualidad, existen multitud de librerías que permiten realizar varias de las funciones necesarias para la aplicación de forma sencilla, además de que es un lenguaje que me es familiar. Por tanto, no tenía ninguna alternativa importante a la hora de elegir un lenguaje.

- *Pycharm*

El IDE que he utilizado para este proyecto ha sido *Pycharm*, que es un entorno de la empresa *Jetbrains*. Lo he elegido por la comodidad para ofrece a la hora de programar con funciones como autocompletado, posibilidad de encontrar fácilmente el origen de las funciones o el depurador que permite la ejecución línea a línea del código. Además, es el entorno que utilizo en mis prácticas en empresa, por lo que me es bastante familiar.

- *Jupyter Note*

Jupyter Notebook es una librería de *Python* que ofrece plataforma interactiva basada en la web. El código de agrupa en cuadernos en los que puedes ejecutar el código de forma sencilla y lo he usado principalmente para probar y realizar experimentos de una manera simple.

- *Tkinter*

Tkinter es la librería de Python que he utilizado para crear la interfaz gráfica. Esta es sencilla de utilizar y permite una relativa flexibilidad que hizo que fuera la opción a elegir para este trabajo. Existían otras opciones como la librería de *PySimpleGUI* que es más simple de utilizar, pero las opciones estaban mucho más limitadas.

- *Spacy*

Spacy es la librería de Python que he utilizado para procesar el texto y realizar las distintas tareas necesarias con este. Esta librería permite separar el texto fácilmente en frases y palabras e incluso es capaz de eliminar las *stop-words* e identificar el tipo de palabra (verbo, sustantivo, adejtivo etc). Otras opciones que existen para este tipo de cosas son *Polyglot NLP o PyNLPl,* pero finalmente fueron descartadas ya que *Spacy* tiene mejor soporte para el español.

- *NetworkX*

Para la creación y análisis de las redes la librería que he utilizado ha sido *NetworkX,* ya que proporciona muchas funciones de generación y facilidades para leer y escribir gráficos en muchos formatos. Esta librería aprovecha los diccionarios de Python para almacenar las medidas de nodos y aristas. Otras opciones que existen para este propósito son *Pandas o NumPy*.

- *Pyvis*

Esta librería la he utilizado para la representación de los gráficos, ya que permite verlo en un navegador de forma cómoda y sencilla y proporciona una vista clara de la red que permite mover los nodos o marcar los más importantes entre otras cosas.

- *Beautiful Soup*

Beautiful Soup es una librería de *Python* para extraer datos de archivos *HTML y XML*. La he utilizado para extraer el texto de las noticias de forma sencilla de la página del periódico digital elDiario y Público que se utilizarán en el experimento final.

- Otros

Además de todas las librerías mencionadas anteriormente, también otras librerías que no considero que tuvieran la suficiente importancia en el proyecto para tener su propio apartado. Estas son:

- *Matplotlib*, utilizado para crear gráficas y mostrarlas por pantalla.

- Separasílabas, una librería que como su propio nombre indica sirve para separar las sílabas de las palabras en español de forma sencilla.
- *Itertools*, se ha utilizado para iterar sobre las palabras para generar la red de forma más sencilla.

5. Desarrollo de la Solución Propuesta

5.1. Extracción de texto para el análisis de legibilidad.

A la hora de desarrollar la aplicación, lo primero que se hizo fue crear una función que preparara el texto para ser analizado. Esto se hizo usando una expresión regular para eliminar todos los caracteres que no fueran válidos para el análisis y después se usó la librería *nlp* para iterar sobre el texto limpio y realizar los conteos de palabras y frases necesarios para calcular los algoritmos tradicionales.

Con esta librería, se separa el texto en *tokens* de forma que cada uno de ellos se corresponde con una de las palabras del texto. Estos *tokens*, incluyen además información sobre la palabra, como por ejemplo su categoría gramatical, lo que se usa para construir la tabla con los verbos, nombres y adjetivos que más aparecen en el texto.

A la hora de separar las palabras en sílabas, usé una librería de *Python* llamada separasilabas, que permitía realizar esta tarea de forma sencilla.

5.2. Análisis de la red

El siguiente desarrollo relevante fue la creación de la red a partir del texto. Para esto lo primero fue añadir cada palabra del texto como un nodo para la red. Luego separé el texto en frases y fui añadiendo una arista por cada par de palabras consecutivas en cada una de las frases.

Un problema a tener en cuenta fue considerar o no las *stopwords* para la creación de la red. Tras hacer pruebas, determiné que la información que la red aportaba con las *stopwords* era mejor para el análisis a realizar que sin incluirlas.

Otro problema importante fue la elección de valores de la red que fueran útiles para el análisis de legibilidad. Finalmente opté por utilizar los siguientes:

- Modularidad: es una medida de la fuerza de división de la red, es decir, de la cantidad y tamaño de las distintas comunidades o módulos en los que se divide la red. Las redes con alta modularidad tienen conexiones densas entre los nodos dentro de los módulos, pero conexiones escasas entre los nodos de diferentes módulos. Esta medida aplicada a la legibilidad de texto nos sirve para saber cómo de enfocado temáticamente está el texto, o, dicho de otra forma, lo enfocada que está la información en un tema concreto.

- Grado medio: el grado mide el número de aristas incidente a cada nodo y en este caso se calcula la media de ese valor. Esta medida indica el grado de conectividad de la red. Por tanto, cuanto mayor sea este valor, más conectadas estarán las palabras de la red.

  El la ilustración 9, se puede ver la distribución del grado de uno de los textos utilizados para probar los valores. Como se puede ver, sigue una ley de potencias, lo que es habitual en redes complejas y además tiene una correspondencia clara con la ley de *Zipf*.

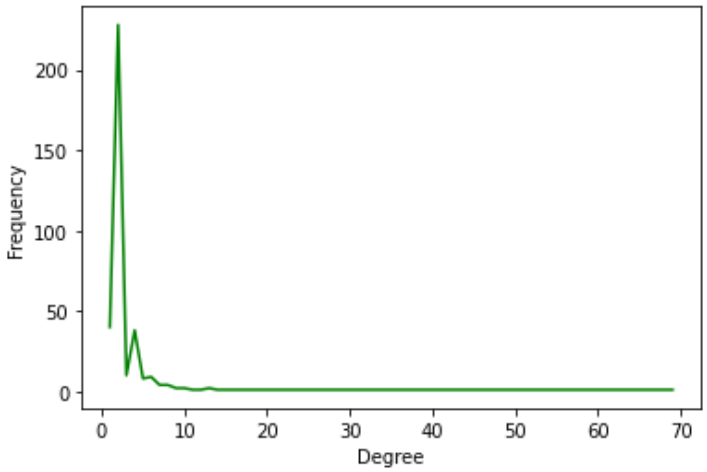


Ilustración 8: Distribución del grado

- Transitividad: esta medida se utiliza para saber si se da el fenómeno de mundo pequeño en la red, que es aquella red en la que se pueda llegar a la mayoría de los nodos desde cualquier otro mediante un pequeño número de saltos o pasos.
  De este modo en un texto con un valor alto de transitividad, las palabras están más conectadas, lo que intuitivamente debería aumentar la legibilidad del texto.

- Longitud media del camino más corto: en este caso el nombre es bastante descriptivo. El hecho de que la longitud de los caminos sea mayor implica que las frases del texto son más largas, lo que intuitivamente hace suponer que el texto será más difícil.

  Para calcular esta medida, hubo que generar el elemento conectado más grande del grafo y realizar esta medida sobre este, ya que esta medida se realiza sobre redes conectadas.

  La medida de la longitud media del camino más corto se utiliza junto a la transitividad para determinar si se da el fenómeno de mundo pequeño en la red. De esta forma, decimos que una red es un mundo pequeño si tiene caminos cortos y un alto coeficiente de transitividad.

5.3. Aplicación

Otro problema, con la implementación vino a la hora de crear la interfaz. El *software* desarrollado sigue un modelo vista controlador que es un estilo de arquitectura de software que separa los datos de una aplicación, la interfaz de usuario, y la lógica de control en tres componentes distintos. Como he dicho anteriormente, la interfaz se hizo utilizando la librería *Tkinter*. El principal problema fue cómo representar la red de forma que se pudiera ver de forma clara por el usuario. Finalmente opté por utilizar la librería llamada *pyvis network,* que representa la red en un navegador y permite al usuario interactuar con ella pudiendo marcar los nodos o moverlos para facilitar la comprensión del usuario.

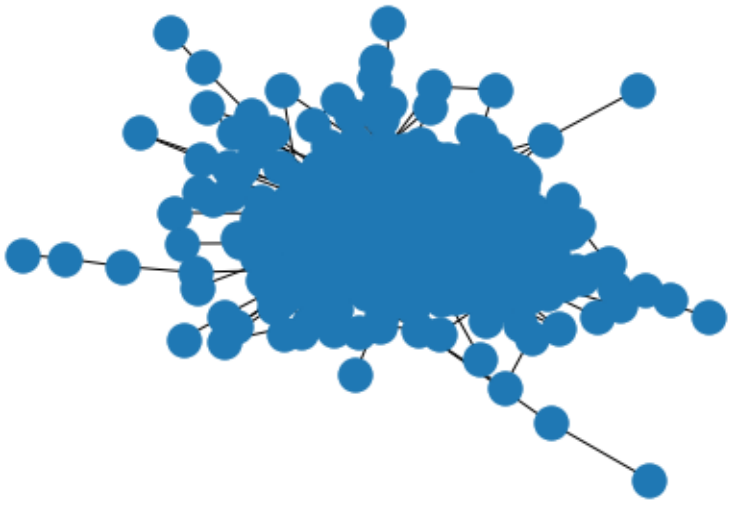


Ilustración 9: Red en matplotlib

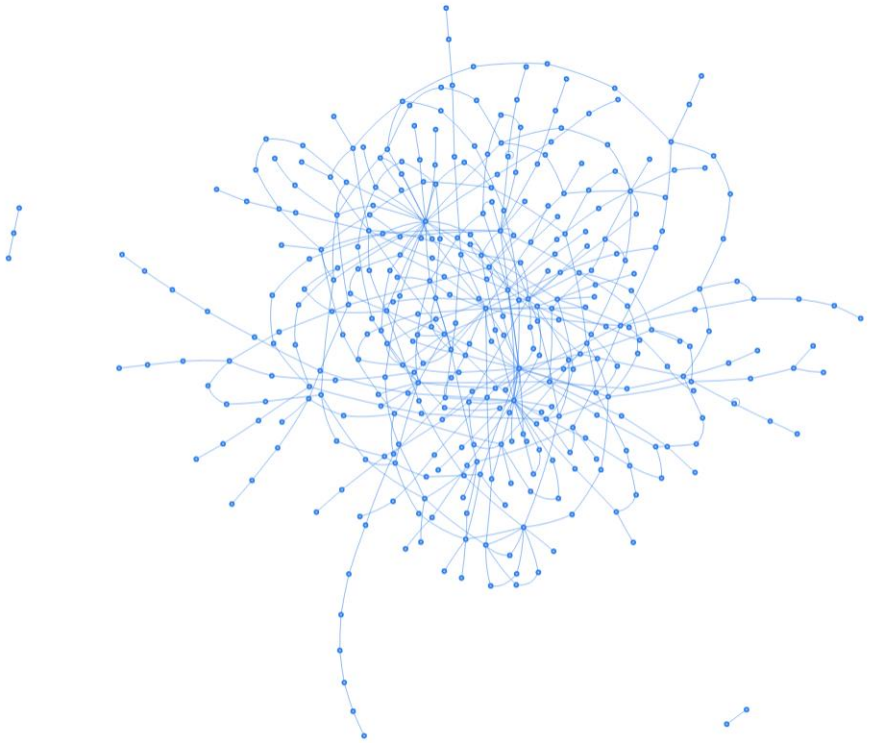


Ilustración 10: Red en pyvis

En la ilustración 10 se puede ver como se ve la red mostrada directamente desde *matplotlib* y en la ilustración 11 se puede ver cómo se ve la red utilizando la librería de *pyvis.network*.

5.4. Ejemplo de Funcionamiento

En la ilustración 12, se puede ver un ejemplo de cómo es la interfaz de la aplicación antes de seleccionar el fichero a analizar.

Una vez el usuario introduce el fichero, se muestra la siguiente pantalla, representada por la ilustración 13, en la que aparecen las tablas con la información además de una visualización en nube de palabras y un gráfico con la ley de Zipf. Además de la representación de la red que se puede ver en la figura ilustración 20.

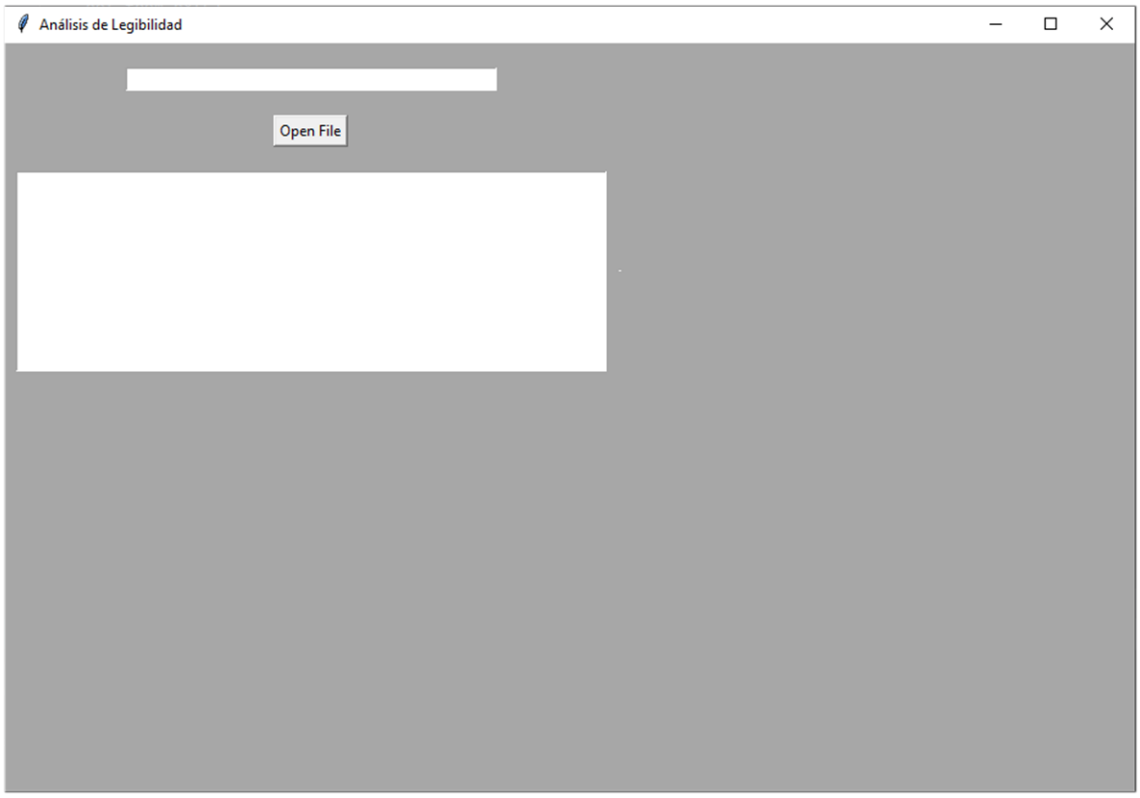


Ilustración 11: Interfaz antes de seleccionar fichero

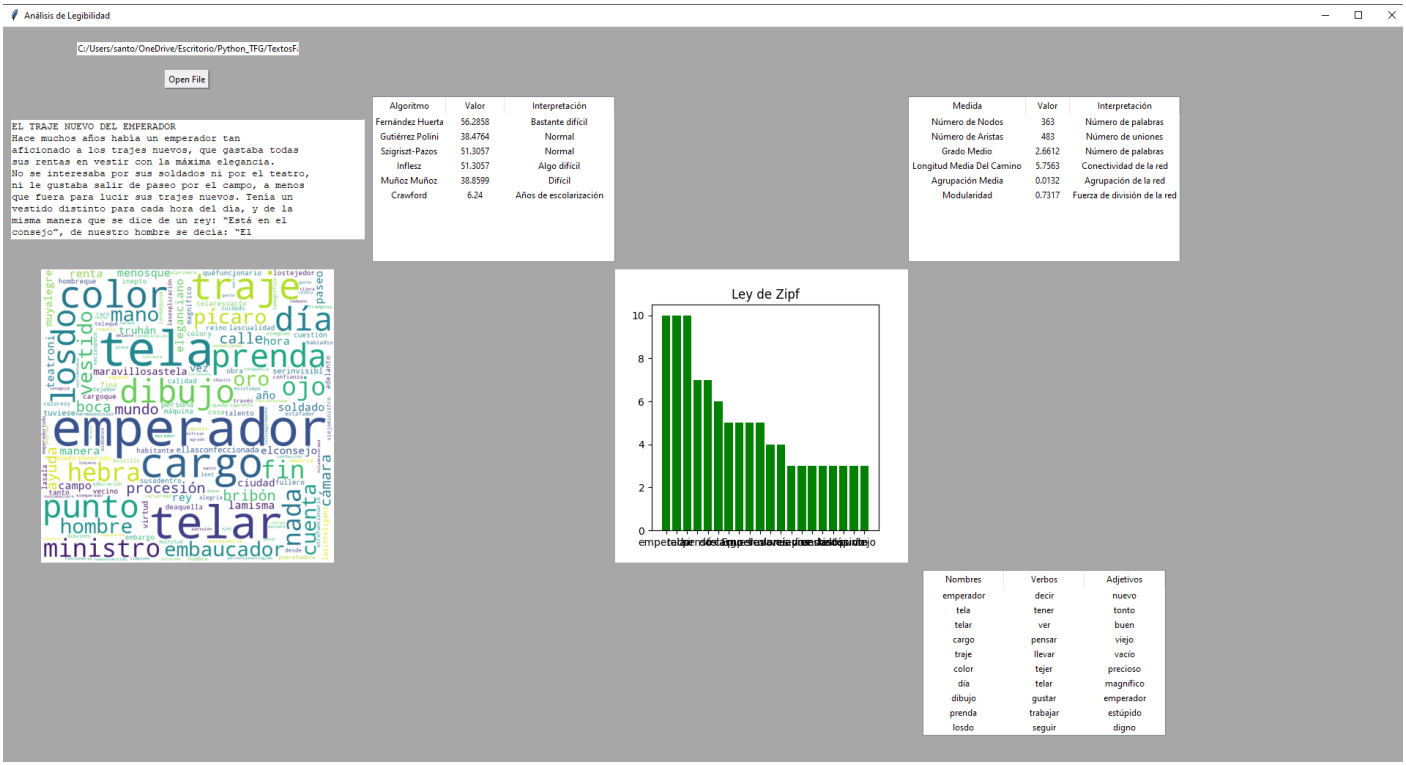


Ilustración 12: Interfaz tras seleccionar fichero que contiene el cuento El Traje Nuevo del Emperador

Ya que en la ilustración 13 se ve muy pequeño, a continuación incluyo capturas más detalladas de cada uno de los elementos de la aplicación con una breve explicación:

Ilustración 13: Representación por nube de palabras

Nombres	Verbos	Adjetivos
emperador	decir	nuevo
tela	tener	tonto
telar	ver	buen
cargo	pensar	viejo
traje	llevar	vacío
color	tejer	precioso
día	telar	magnífico
dibujo	gustar	emperador
prenda	trabajar	estúpido
losdo	seguir	digno

Ilustración 14: Tabla con los nombres, verbos y adjetivos que más aparecen

Los elementos de la aplicación representados por las ilustraciones 14 y 15 sirven para que el usuario pueda hacerse una idea del contenido del texto de forma rápida. Por ejemplo, en este caso, viendo estos dos elementos es fácil deducir que se trata del cuento El Traje Nuevo del Emperador.

Medida	Valor	Interpretación
Número de Nodos	363	Número de palabras
Número de Aristas	483	Número de uniones
Grado Medio	2.6612	Número de palabras
Longitud Media Del Camino	5.7563	Conectividad de la red
Transitividad Media	0.0132	Agrupación de la red
Modularidad	0.7317	Fuerza de división de la red

Ilustración 15: Tabla con los resultados del análisis de la red

Algoritmo	Valor	Interpretación
Fernández Huerta	56.2858	Bastante difícil
Gutiérrez Polini	38.4764	Normal
Szigriszt-Pazos	51.3057	Normal
Inflesz	51.3057	Algo difícil
Muñoz Muñoz	38.8599	Difícil
Crawford	6.24	Años de escolarización

Ilustración 16: Tabla con el resultado de los algortimos clásicos y su interpretación

Los elementos de la aplicación representados por las ilustraciones 16 y 17 son los que representan el propio análisis de la legibilidad del texto.

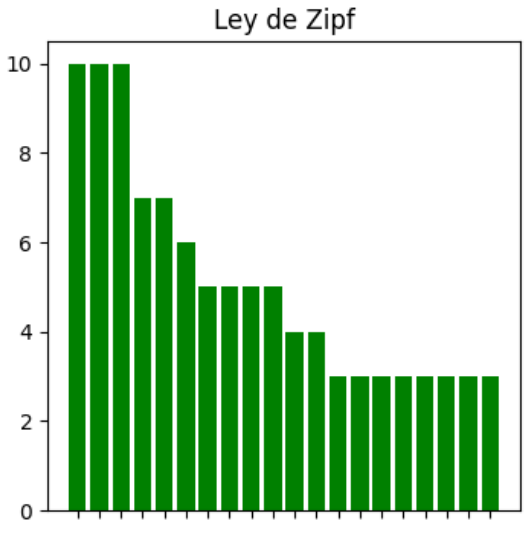


Ilustración 17: Representación de la Ley de Zipf

```
EL TRAJE NUEVO DEL EMPERADOR
Hace muchos años había un emperador tan
aficionado a los trajes nuevos, que gastaba todas
sus rentas en vestir con la máxima elegancia.
No se interesaba por sus soldados ni por el teatro,
ni le gustaba salir de paseo por el campo, a menos
que fuera para lucir sus trajes nuevos. Tenía un
vestido distinto para cada hora del día, y de la
misma manera que se dice de un rey: "Está en el
consejo", de nuestro hombre se decía: "El
```

Ilustración 18: Texto del fichero seleccionado

La ilustración 18 muestra la representación de la ley de *Zipf* del texto introducido y se puede usar para comprobar si el texto sigue una ley de potencias. La ilustración 19 muestra el texto del fichero.

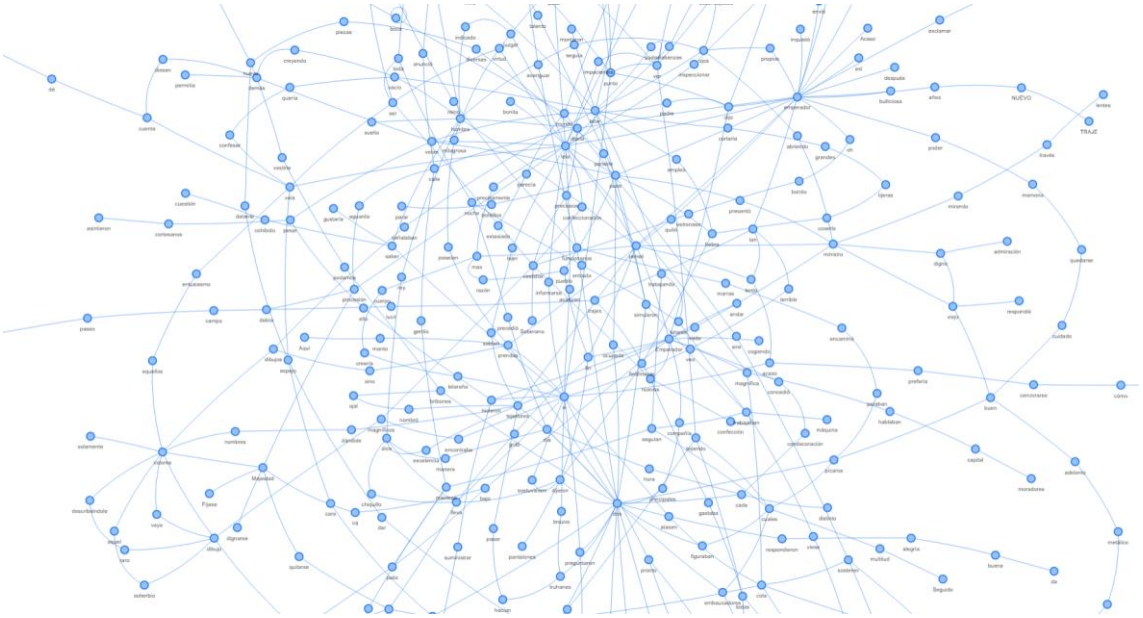


Ilustración 19: Red generada por la aplicación

6. Validación de la aplicación

Para probar la aplicación y determinar unos valores que sirvan como estimación para las medidas de la red y además comprobar cómo se relacionan estos con los valores devueltos por los algoritmos tradicionales, probé a analizar una serie de textos con la aplicación y apuntar los resultados.

Para realizar este experimento utilicé textos de alrededor de 800 palabras. Para obtener unos valores de referencia, usé 10 textos fáciles, compuestos de cuentos infantiles sencillos y 10 textos que vienen de fragmentos de textos académicos (buscando que a pesar de ser una parte tuvieran significado por ellos mismos).

Los resultados que se obtuvieron para cada uno de los atributos de la red se pueden ver representados en la ilustración 21 donde los puntos rojos representan los textos difíciles y las estrellas verdes los textos fáciles.

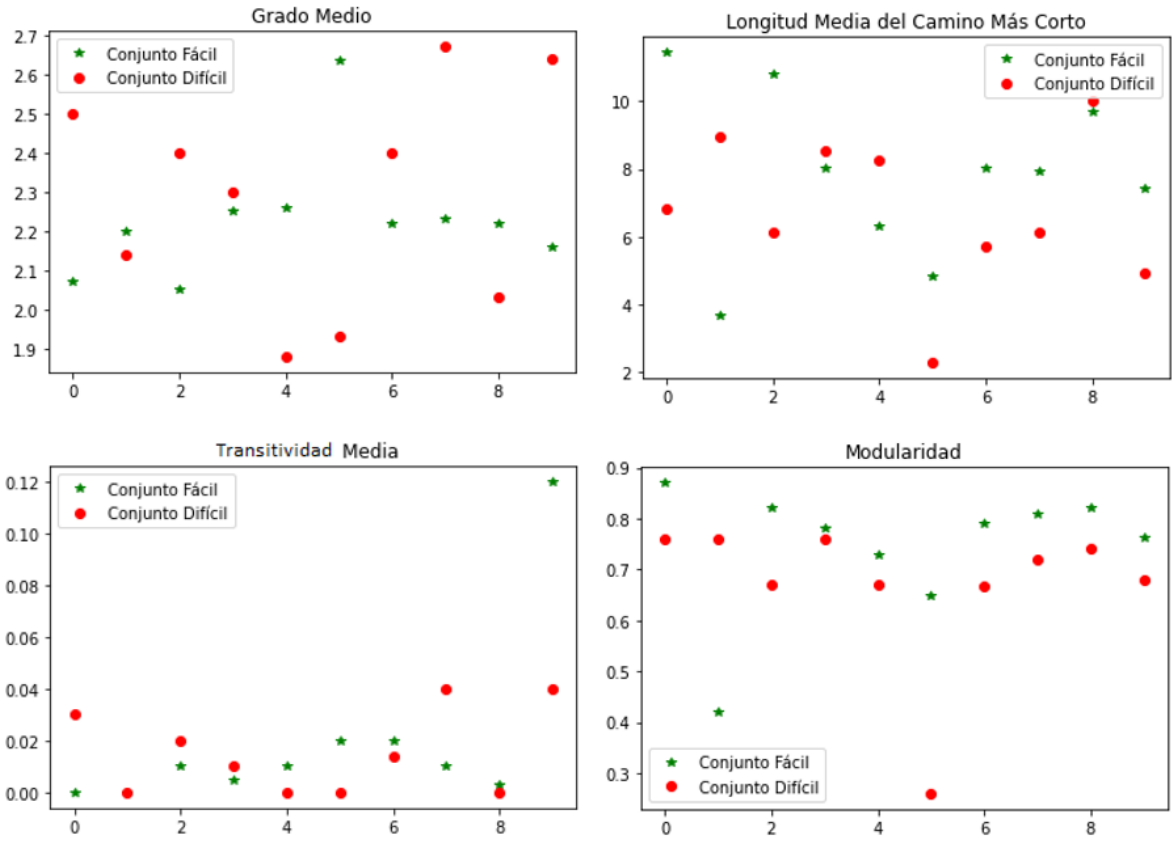


Ilustración 20: Análisis de los valores de la red para textos cortos

Como se puede ver, aparte de en la modularidad, no parece haber una regularidad en los valores, ya que a simple vista no emerge ningún patrón que permita afirmarlo.

Para verificar esto se ha usado la hipótesis de igualdad de medias para cada uno de los atributos de la red.

- Grado medio: se obtiene un *p_value* de 0.5709 que al ser mayor que 0.05 no se puede rechazar la hipótesis y por tanto se verifica lo dicho anteriormente.
- Longitud Media del Camino Más Corto: se obtiene un *p_value* de 0.3401 que como dije anteriormente, es mayor de 0.05 y verifica lo dicho anteriormente.

- Transitividad media: se obtiene un *p_value* de 0.7286 que, una vez más, es mayor de 0.05 y verifica lo dicho anteriormente.

Mi primera suposición al ver estos resultados fue pensar que esto se podía deber al uso de las *stop words* en la creación de la red y decidí realizar el mismo experimento pero eliminándolas previamente. Pero el resultado fue el mismo que usando *stop words* y se pueden ver los resultados obtenidos en el anexo.

Después pensé que tal vez se debía a que al ser textos relativamente cortos, las redes eran demasiado pequeñas y posiblemente no aportaban información suficiente, por tanto decidí repetir el experimento, pero esta vez usando textos más largos.

Para el caso de los textos difíciles, esta vez opté por cuentos de Borges que tienen una longitud media de 2673.2 palabras. Para el caso de los textos sencillos otra vez opté por cuentos de niños escritos por autores como Oscar Wilde o Charles Perrault, y estos tenían una longitud media de 2446.2 palabras.

En las siguientes ilustraciones se puede ver la misma representación que hice anteriormente:

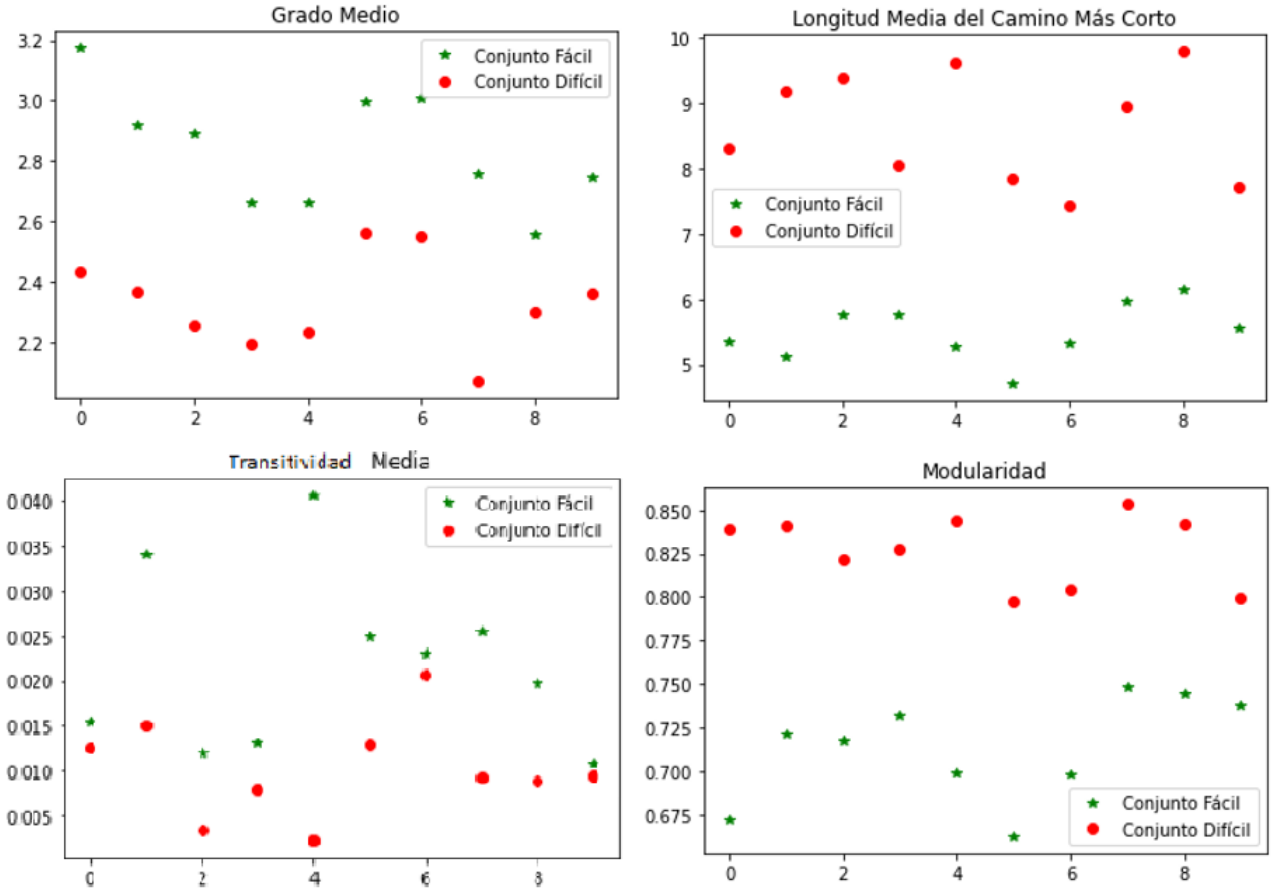


Ilustración 21: Análisis de los valores de la red para textos largos

Como se puede ver, en este caso sí existe una diferencia clara entre los textos fáciles y los textos difíciles.

Los valores medios obtenidos para ambos tipos de textos se pueden ver en la tabla 5:

Medida	Valores textos difíciles	Valores textos fáciles
Grado Medio	2.33242 ± 0.147	2.8365 ± 0.183
Longitud Media del Camino Más Corto	8.6294 ± 0.815	5.4942 ± 0.406
Transitividad Media	0.0102 ± 0.005	0.0219 ± 0.010
Modularidad	0.8269 ± 0.019	0.7132 ± 0.028

Tabla 5: Media de los valores de textos difíciles y fáciles

En vista de los resultados, como resulta intuitivo pensar, en los textos más complicados, los caminos de la red son más largos de media, lo que indica que las frases son más largas. En cuanto a la transitividad media, en los textos fáciles esta resulta ser el doble que en los difíciles, lo que junto al hecho de que los caminos son más cortos indica que es más fácil que se produzca el fenómeno del mundo pequeño, y esto, como comenté anteriormente, haría que el texto fuera más sencillo, ya que implicaría que todos los nodos están más conectados y por tanto hay menos palabras distintas.

Para verificar esto se ha usado la hipótesis de igualdad de medias para cada uno de los atributos de la red.

- Grado medio: se obtiene un *p_value* de 4.639e-06 que como se puede ver es inferior a 0.05.

- Longitud Media del Camino Más Corto: se obtiene un *p_value* de 5.465e-09 que como dije anteriormente, es inferior a 0.05.

- Transitividad media: se obtiene un *p_value* de 0.0039 que, una vez más, es menor de 0.05 y verifica lo dicho anteriormente.

- Modularidad: se obtiene un *p_value* de 9.5542e-09 que, una vez más, es mucho menor de 0.05 y se verifica lo dicho anteriormente.

Esto también se verifica al comprobar el grado del texto, ya que en los textos difíciles es ligeramente mayor, lo que implica que los nodos están menos conectados y eso podría significar que aparecen más palabras distintas.

Debido a que el resultado resultó verificado, voy a utilizar los valores medios obtenidos en cada atributo como referencia para experimentos futuros a realizar con la aplicación.

A continuación, adjunto una captura de dos redes de unos de los textos difíciles en las ilustraciones 23 y 24 y una de las redes de unos de los fáciles en las ilustraciones 25 y 26, para que se pueda apreciar la diferencia en las redes generadas:

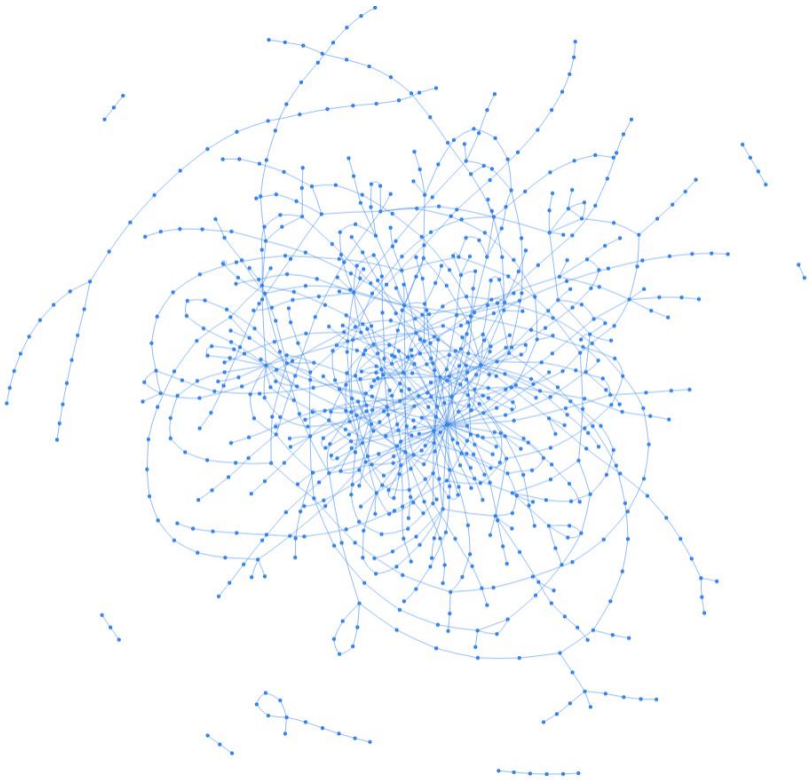


Ilustración 22: Gráfica del cuento El milagro secreto de Borges

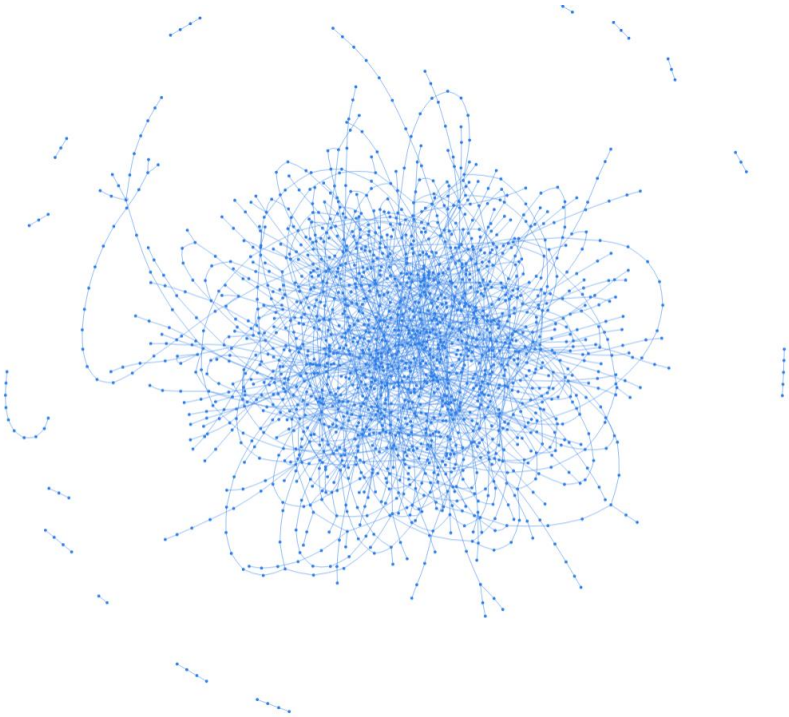


Ilustración 23: Gráfica del cuento El inmortal de Borges

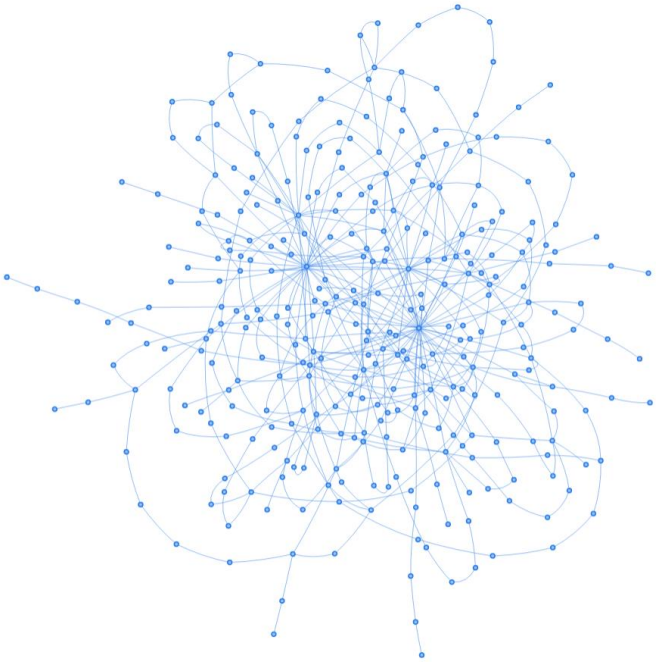


Ilustración 24: Red del Gato con botas

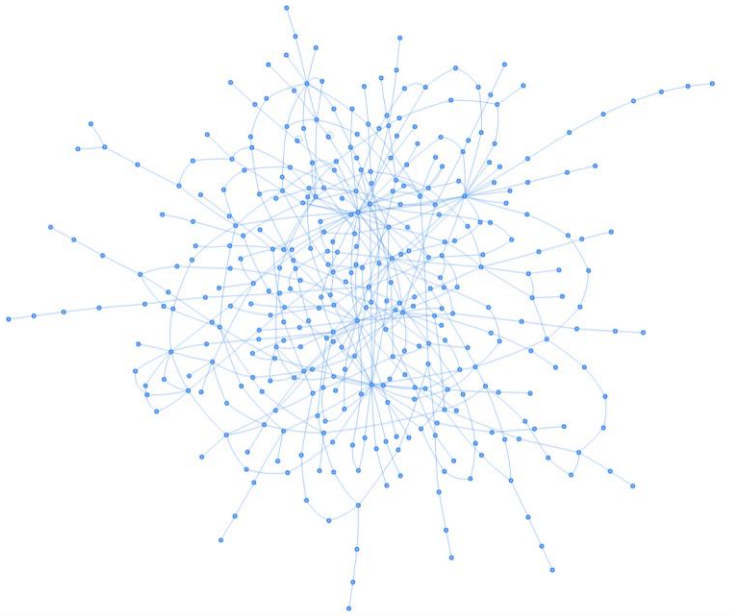


Ilustración 25: Red del cuento del Traje Nuevo del Emperador

En las ilustraciones 23, 24, 25 y 26, se puede ver que las redes de los textos difíciles son más complejas, como cabía esperar.

Para finalizar, mostraré el análisis devuelto por la aplicación de un texto sencillo, en este caso utilizaré el cuento del Intrépido Soldadito de Plomo:

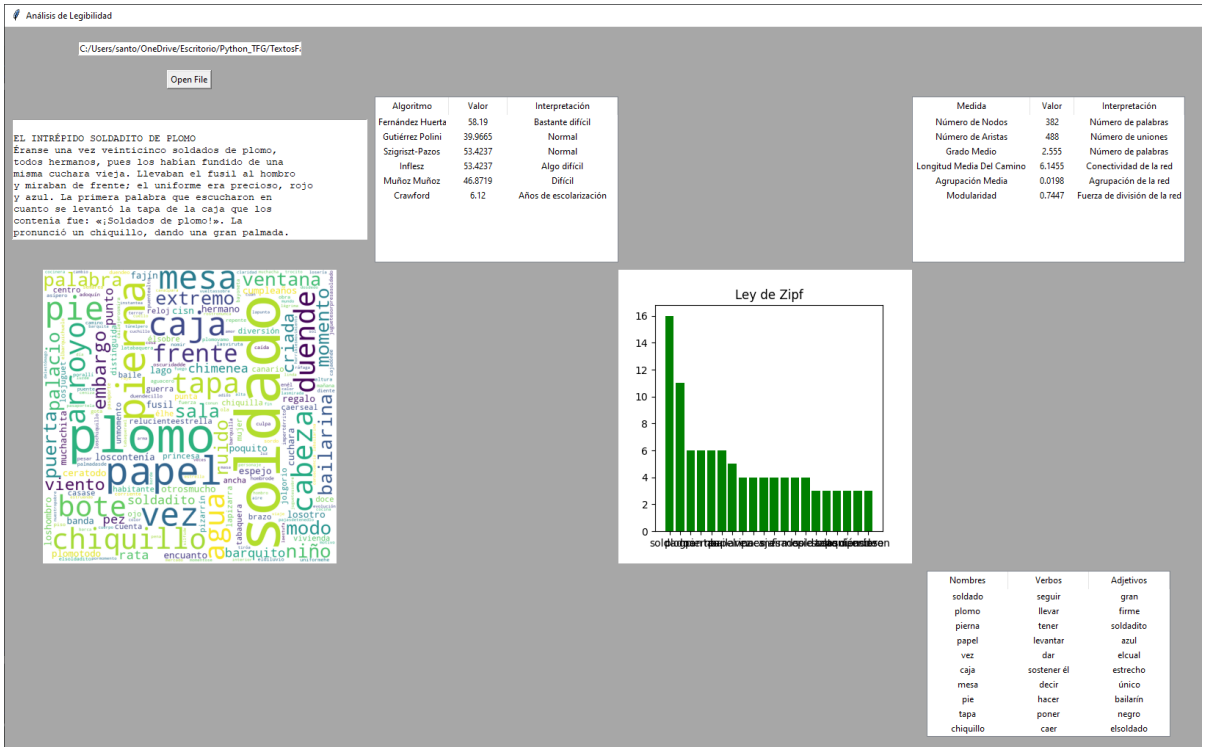


Ilustración 26: Interfaz tras analizar El Intrépido Soldadito de Plomo

Como se puede ver en la ilustración 27, el análisis de nube de palabras junto con la tabla en la que aparecen los nombres, verbos y adjetivos más populares, permiten hacerse una idea del contenido del texto de una manera sencilla y rápida.

Algoritmo	Valor	Interpretación
Fernández Huerta	58.19	Bastante difícil
Gutiérrez Polini	39.9665	Normal
Szigriszt-Pazos	53.4237	Normal
Inflesz	53.4237	Algo difícil
Muñoz Muñoz	46.8719	Difícil
Crawford	6.12	Años de escolarización

Ilustración 28: Resultado algoritmos

Medida	Valor	Interpretación
Número de Nodos	382	Número de palabras
Número de Aristas	488	Número de uniones
Grado Medio	2.555	Número de palabras
Longitud Media Del Camino	6.1455	Conectividad de la red
Agrupación Media	0.0198	Agrupación de la red
Modularidad	0.7447	Fuerza de división de la red

Ilustración 27: Resultado análisis de la red

Las ilustraciones 28 y 29 muestran las tablas de valores de forma ampliada para que se vean mejor. Curiosamente, se puede ver que el análisis de los algoritmos devuelve que es un texto complicado generalmente, lo cual no es cierto, ya que se trata de un texto infantil sencillo. El análisis de la red, en cambio devuelve valores que están más próximos a los de los textos sencillos que he obtenido anteriormente en la tabla 5.

Por tanto, aquí hay un ejemplo en el que el análisis de la red ofrece un resultado más acertado que los algoritmos tradicionales.

En la figura 31, se puede ver la red que se forma al analizar este texto:

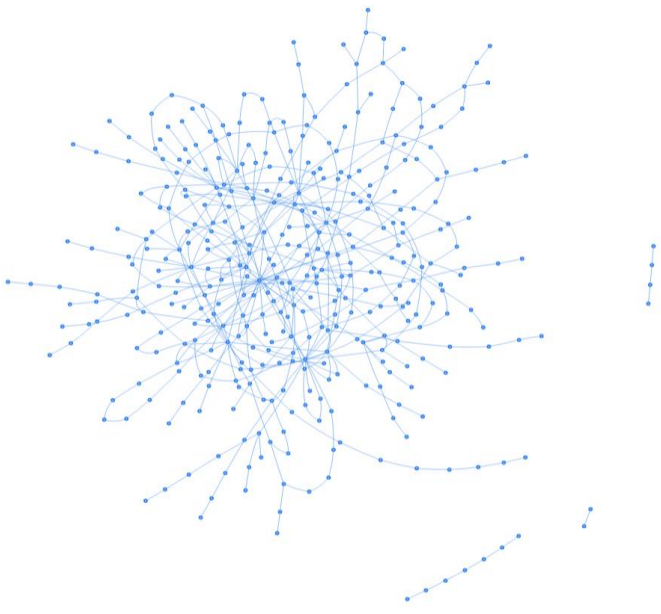


Ilustración 29:Red del Soldadito de Plomo

7. Experimento Medios de Comunicación

Para la realización de los siguientes experimentos se van a usar textos de dos periódicos digitales distintos, elDiario y Público. La elección de estos en concreto se debe a que los textos son públicos y tratan temas de actualidad que además están relacionados con algunos de los ods, como es la guerra de Ucrania o el día internacional del orgullo gay.

La primera prueba va a consistir en el análisis de distintas temáticas de elDiario para ver si existe una diferencia en cuanto a la legibilidad de los textos de cada uno de estos temas.

En la segunda prueba se van a analizar y comparar textos de la misma temática, pero de distintos periódicos con el mismo fin que en la primera prueba.

El análisis se va a realizar con la herramienta desarrollada.

7.1. Comparativa de diferentes temas de elDiario

En esta sección, voy a utilizar la aplicación para analizar si existe una diferencia en cuanto a la legibilidad en distintas categorías distintas de elDiario. Las categorías elegidas son Cultura, Política y Sociedad. Como se vio en el capítulo 6, el análisis de la red resultó ser más fiable que el que se realiza con los algoritmos tradicionales. Por tanto, usaré ese análisis de la red para compararlos.

La intuición inicial, es que los textos de política serán más complicados que los de opinión y cultura, debido a que la temática es más compleja. Por otro lado, los dos últimos serán parecidos en cuanto a este análisis.

Para realizar este análisis he utilizado 10 textos de cada una de las categorías que tienen una longitud mínima de 900 palabras ya que, como se vio anteriormente, para textos con una longitud menor a 800 el análisis no funcionaba bien. Para ello planteamos como hipótesis nula:

H_0: no hay diferencia en la complejidad de los textos de las áreas seleccionadas y buscaremos evidencia para confirmarla o rechazarla.

A continuación, incluyo los valores medios que he obtenido para cada uno de los temas en la tabla 7:

Medida	Cultura	Política	Sociedad
Grado Medio	2.4432 ± 0.0956	2.6477 ± 0.184	2.4455 ± 0.0824
Longitud Media del Camino Más Corto	7.9453 ± 0.9579	6.5467 0.7412	7.6042 ± 0.6963
Transitividad Media	0.0069 ± 0.0058	0.0198 ± 0.0095	0.0139 ± 0.0117
Modularidad	0.7828 ± 0.0268	0.7241 ± 0.0326	0.7628 ± 0.0177

Tabla 6: Resultados medios de elDiario, valor más menos desviación estándar

Para comprobar si son suficientemente similares, se realizado un anova unidireccional sobre los conjuntos de datos y los datos recogidos han sido los siguientes para las distintas medidas:

La hipótesis nula Ho, afirma que las medias de la población son todas iguales. Para esto, usaremos un p_valor de 0,05. Un nivel de significación de 0,05 indica un riesgo del 5% de concluir que existe una diferencia cuando en realidad no la hay.

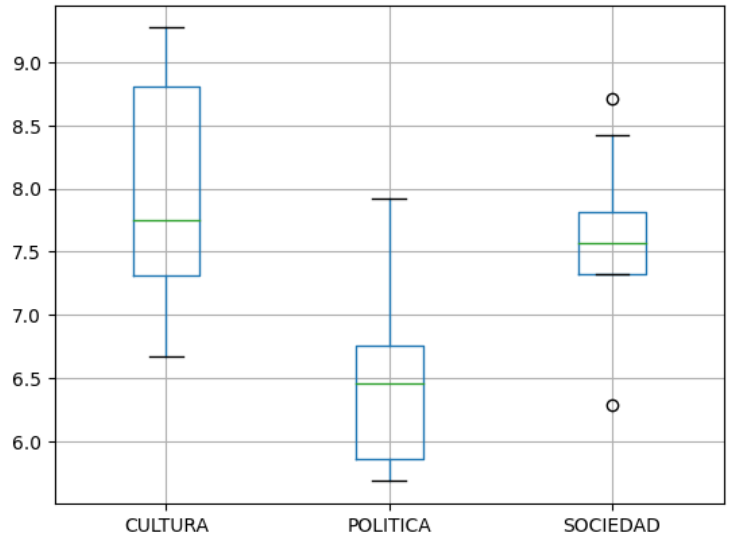


Ilustración 30: Diagrama de cajas y bigotes

	Grado Medio	Longitud Media del Camino Más Corto	Transitividad Media	Modularidad
P_valor	0.0030	0.0032	0.0259	0.0003
F	7.4735	7.3557	4.2687	11.4817

Tabla 7: Resultado del Anova del conjunto de textos de Cultura, Sociedad y Política

Ya que el valor p es menor o igual que el nivel de significación, se rechaza la hipótesis nula y se concluye que no todas las medias de la población son iguales.

Como se puede ver en la tabla 7, los valores de Cultura y Sociedad son muy parecidos, y a simple vista, comparando con valores anteriores que quedaron recogidos en las tablas 5 y 6, se puede ver que su nivel de dificultad es más alto que el de política, lo cual es contrario a la hipótesis inicial de que los textos de Política eran serían más difíciles.

Antes de verificarlo estadísticamente, representaré en una tabla la comparación de los valores de cada categoría con los textos fáciles o difíciles. Para el caso de cultura y sociedad al ser tan parecidos, se ha usado la media de los valores de ambos para compararlo con los difíciles y para el caso de política se comparará con los valores de los fáciles.

Política

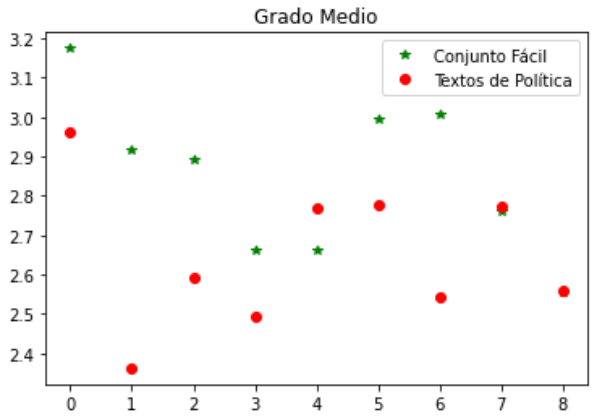


Ilustración 31: Grado de medio de textos de política

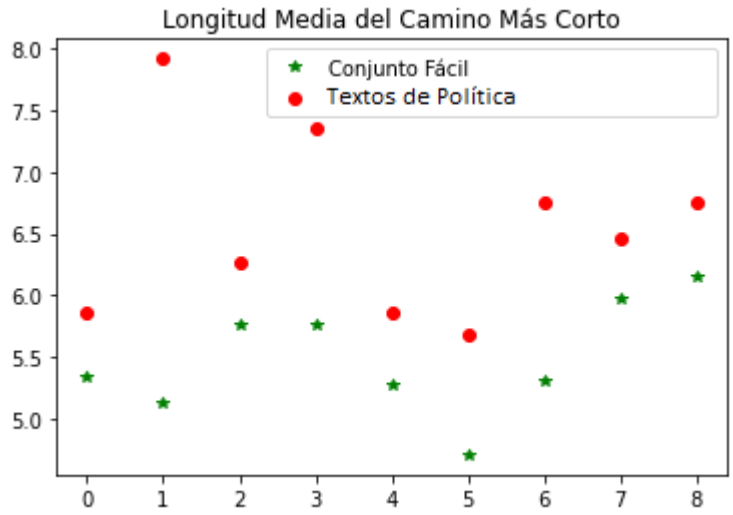


Ilustración 32: Longitud Media del Camino de textos de política

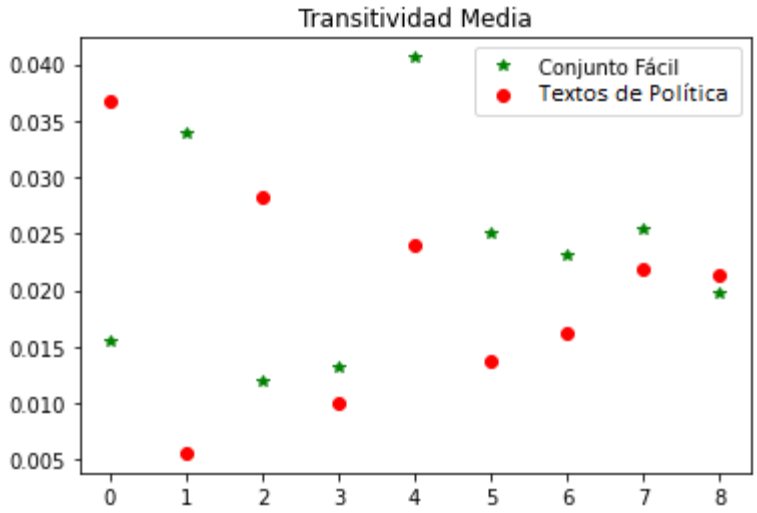
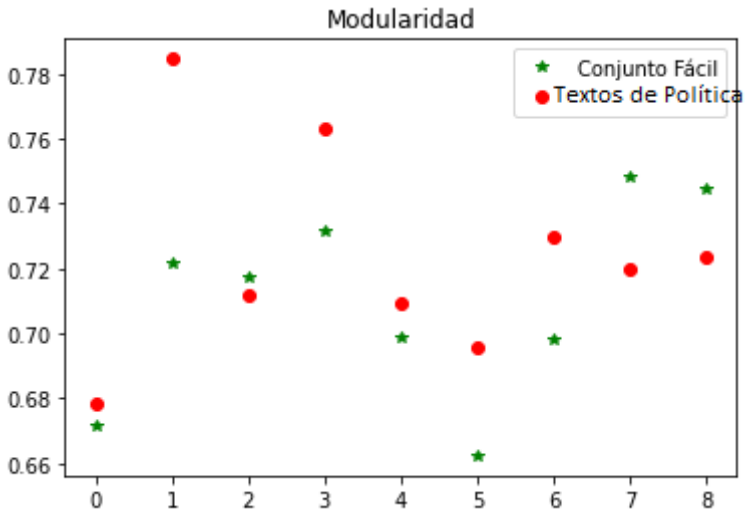


Ilustración 33: Transitividad Media de textos de política



Para verificar que los textos de política están próximos a los valores obtenidos para los textos fáciles, se ha usado la hipótesis de igualdad de medias para cada uno de los atributos de la red.

- Grado medio: se obtiene un *p_value* de 4.5530e-05 que como se puede ver es inferior a 0.05.

- Longitud Media del Camino Más Corto: se obtiene un *p_value* de 1.0059e-06 que como dije anteriormente, es inferior a 0.05.

- Transitividad media: se obtiene un *p_value* de 0.0823. En este caso, el valor ha sido superior a 0.05, por tanto no se puede rechazar la hipótesis y en este caso no verifica lo dicho.

- Modularidad: se obtiene un *p_value* de 0.0004 que, una vez más, es menor de 0.05 y se verifica lo dicho anteriormente.

Por tanto, a excepción de la transitividad que se pasa por 0.0323, podemos ver que para este conjunto de textos elegidos, estos son analizados como fáciles.

Ahora para terminar este experimento, se mostrará la comparación de los valores de Cultura y Sociedad con los de los textos fáciles:

Cultura y Sociedad

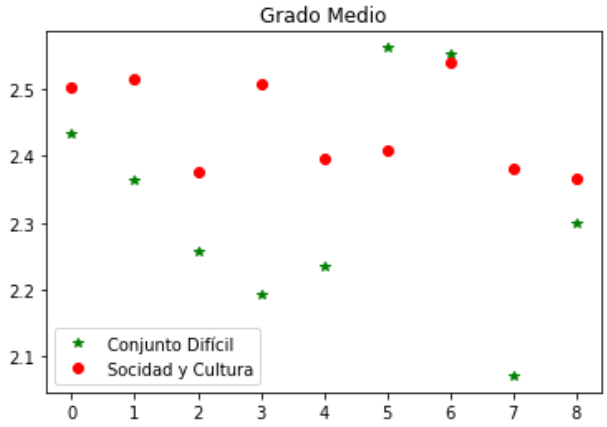


Ilustración 34: Grado Medio de textos de cultura y sociedad

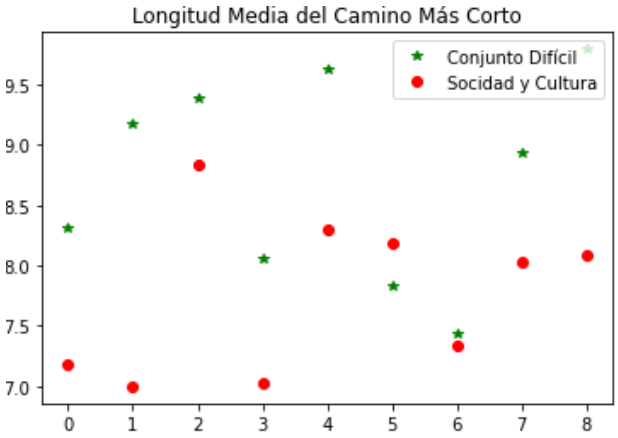


Ilustración 35: Longitud Media del Camino de textos de cultura y sociedad

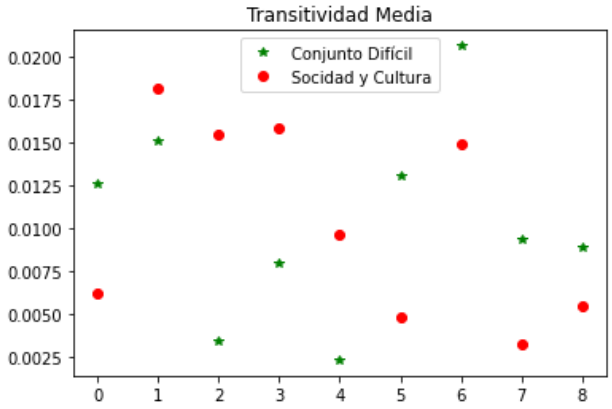


Ilustración 36: Transitividad Media de textos de cultura y sociedad

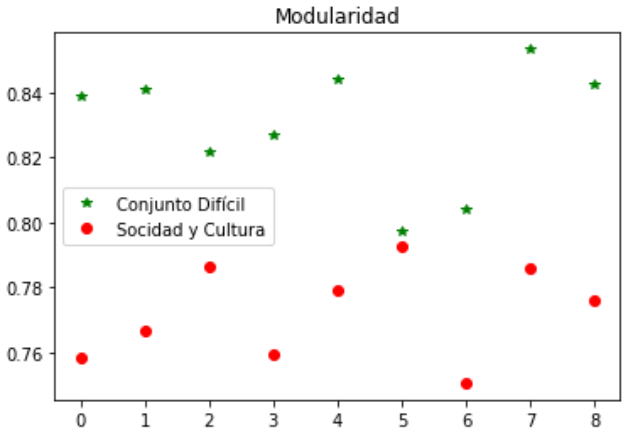


Ilustración 37: Modularidad de textos de cultura y sociedad

Como en el caso anterior, para verificar que los textos de sociedad y cultura están próximos a los valores obtenidos para los textos difíciles, se ha usado la hipótesis de igualdad de medias para cada uno de los atributos de la red.

- Grado medio: se obtiene un *p_value* de 0.0721 que como se puede ver es ligeramente superior a 0.05.

- Longitud Media del Camino Más Corto: se obtiene un *p_value* de 0.0165 que, en este caso, es inferior a 0.05.

- Transitividad media: se obtiene un *p_value* de 1. En este caso, el valor ha sido superior a 0.05, por tanto, no se puede rechazar la hipótesis y en este caso no verifica lo dicho.

- Modularidad: se obtiene un *p_value* de 2.4629e-06 que es menor de 0.05 y se verifica lo dicho anteriormente.

Por tanto, ya que en este caso dos de los valores no se han verificado, no podemos asumir que sean difíciles, pero en cualquier caso se puede ver que según este análisis estos dos tipos de textos han resultado ser más difíciles que los de política.

Finalmente la conclusión que obtenemos de la realización de este experimento, es que los textos de política han resultado ser más sencillos que los textos de cultura y sociedad.

7.2. Comparativa de la sección de opinión de elDiario y Público

En este segundo experimento, se comparará la misma sección de dos periódicos diferentes. En este caso se ha usado la sección de opinión de los periódicos elDiario y Público. Una vez más se utilizarán los valores de la red proporcionados por la aplicación para obtener realizar este análisis. Se analizarán 10 textos de cada periódico de esta categoría que tienen al menos 900 palabras.

En la tabla 8 se muestra el valor medio de cada uno de las medidas utilizadas:

Medidas	**elDiario**	**Público**	**P_valor**
Grado Medio	2.4143 ± 0.1118	2.4097 ± 0.2695	0.9626
Longitud Media del Camino Más Corto	7.9277 ± 1.1531	8.0642 ± 1.4519	0.8283
Transitividad Media	0.0099 ± 0.0039	0.0112 ± 0.0052	0.5641
Modularidad	0.7778 ± 0.0192	0.7773 ± 0.0290	0.9647

Tabla 8: Valores de elDiario y Público

Debido a que los valores eran tan parecidos he añadido una columna con el p_valor que sale de la comparación de ambos para ver si ambos se pueden considerar iguales en cuanto a la legibilidad y como se puede el valor es ampliamente superior a 0.05 por tanto no se puede rechazar la hipótesis de que sean iguales.

Esta conclusión es la esperable ya que se trata de las noticias de un mismo día y de una misma sección y aunque sean distintos periódicos, la mayoría de los textos trataban de los mismos temas, principalmente de la guerra en Ucrania y sobre el mes del orgullo LGBT.

8. Conclusiones

En este trabajo, se ha desarrollado una aplicación que permite un análisis de los textos que va más allá de utilizar los algoritmos tradicionales. Para esto, se hizo primero de todo un estudio de los posibles métodos a utilizar e inicialmente se optó por incorporar la creación y análisis de una red formada a partir de los textos.

Utilizando estos métodos, se ha creado una aplicación con una interfaz de usuario que permite seleccionar un fichero y realizar un análisis del contenido del texto de este utilizando lo nombrado anteriormente además de incorporar distintos elementos que permiten al usuario hacerse una idea del contenido del texto como un análisis de nube de palabras y una tabla en la que incluyen los nombres, verbos y adjetivos que más aparecen.

Además de esto, se muestra en el navegador una versión interactiva de la red generada para que el usuario pueda visualizarla.

Con el desarrollo de esta aplicación, se han satisfecho los requisitos no funcionales nombrados en el capítulo 3, que eran que la interfaz sencilla e intuitiva además de re escalable.

Mediante algunos experimentos, se ha verificado el funcionamiento de la aplicación y se ha visto que, para el caso de los textos largos de alrededor de 2700 palabras, los resultados devueltos por el análisis de la red eran más fiables a los de los algoritmos tradicionales, mientras que para los textos cortos de alrededor de 800 palabras sucedió lo contrario.

Más tarde, en el experimento realizado, usando el análisis de la aplicación, se ha visto que no había una diferencia notable entre la dificultad de los textos de sociedad de elDiario y Público. Además, se ha comprobado que para una selección de textos de elDiario de 3 secciones distintas: política, cultura y sociedad si existía una diferencia de dificultad entre ellos, y se ha llegado a la conclusión de que los textos de política eran los más fáciles de los 3 elegidos.

Por tanto, se han conseguido solucionar todos los objetivos planteados al principio del trabajo, pero durante la realización de este, se me han ocurrido varias mejoras para la aplicación que se dejan como trabajo futuro.

8.2. Relación del trabajo desarrollado con los estudios cursados

Para la realización de este TFG he usado muchos de los conocimientos que he aprendido durante mis estudios, además de otros nuevos que he aprendido a utilizar para la realización de este trabajo.

Entre los que había aprendido, se incluye toda la parte del análisis estadístico (anova, p_valor), la parte de creación de interfaz y la parte de ingeniería del software (diagrama UML, casos de uso, prototipo).

Las cosas nuevas que he tenido que aprender incluyen el uso de las librerías específicas que he tenido que gastar y sobre todo el tema del análisis de la red y aprender la interpretación de todos los valores.

8.3. Trabajos futuros

Como trabajos futuros, se podrían añadir funcionalidades a la aplicación, como, por ejemplo, añadir la posibilidad de elegir dos ficheros y realizar una comparación de ambos.

Otra extensión que se podría a hacer a la aplicación sería la posibilidad de introducir un enlace y que la aplicación realice el análisis sobre el texto la página.

Finalmente, también se podrían realizar análisis más exhaustivos con otros tipos de textos y un rango mayor de longitudes.

9. Bibliografía

Alsumait, L. (2008). On-Line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking.

Barrio Cantalejo, I. M., & Simón Lorda, P. (2003). ¿Pueden leer los pacientes lo que pretendemos que lean? Un análisis de la legibilidad de materiales escritos de educación para la salud. Atención Primaria.

Blanco, A. (2004). A propósito de la legibilidad de lectura.

Cantalejo, B. &. (2003). Atención primaria: Publicación oficial de la Sociedad Española de Familia y Comunitaria. 104-108.

F. de Arruda, H., da F. Costa, L., & R. Amancio, D. (2016). Topic segmentation via community detection in complex networks.

Ferrando Belart, V. (2004). La legibilidad: un factor fundamental para comprender un texto. 143-146.

Hotho, A. (2005). A Brief Survey of Text Mining.

Madrazo Azpiazu, I., & Soledad Pera, M. (2019). Multiattentive Recurrent Neural Network Architecture for Multilingual Readability Assessment. Cambridge: MIT Press.

Martinc, M., Pollak, S., & Robnik-Šikonja, M. (2021). Supervised and Unsupervised Neural Approaches to Text Readability. Computational Linguistics. 141–179.

Peñaranda Cortés, R. (2015). Análisis y evolución de la legibilidad web en entidades públicas.

Petersen, S. E., & Ostendorf, M. (2009). A machine learning approach to reading level assessment. Computer Speech & Language. 89–106.

Porras-Garzón, J., & Estopà, R. (2020). Escalas de legibilidad aplicadas a informes médicos: límites de un análisis cuantitativo formal. Círculo de Lingüística Aplicada a la Comunicación. 205-216.

Ríos, I. (2009). Influencias del lenguaje y origen de un lector en la comprensión de mensajes de comunicación en salud y en la formación de actitud e intención hacia la realización de una conducta preventiva.

Schwarm, S. E., & Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics. 523–530.

Spaulding, S. (1956). A Spanish Readability Formula.

Szigriszt Pazos, F. (1993). Sistemas predictivos de legibilidad del mensaje escrito: fórmula de perspicuidad.

Szigriszt, F. (1993). Sistemas predictivos de legibilidad del mensaje escrito: Fórmula de Perspicuidad.

Todirascu, A. (2016). Are Cohesive Features Relevant for Text Readability Evaluation? Osaka: The COLING 2016 Organizing Committee.

Vajjala, S., & Lučić., I. (2018). OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications.

Wikipedia. (2020). Nivel Mesoscópico.

Wikipedia, W. (2 de Julio de 2022). *Wikipedia*. Obtenido de https://en.wikipedia.org/w/index.php?title=Flesch%E2%80%93Kincaid_readability_tests&oldid=1091166383

10. Anexo

Valores de los textos fáciles devueltos por la aplicación con *stop words*.

Texto 1

Algoritmo	Valor	Interpretación
Fernández Huerta	68.5924	Normal
Gutiérrez Polini	43.7605	Normal
Szigriszt-Pazos	64.4126	Normal
Inflesz	64.4126	Normal
Muñoz Muñoz	65.7001	Adecuado
Crawford	5.22	Años de escolarización

Medida	Valor
Número de Nodos	531
Número de Aristas	868
Grado Medio	3.2693
Longitud Media Del Camino	3.9703
Agrupación Media	0.0712
Modularidad	0.5828

Texto 2

Algoritmo	Valor	Interpretación
Fernández Huerta	72.5759	Bastante fácil
Gutiérrez Polini	45.2877	Normal
Szigriszt-Pazos	68.3247	Bastante fácil
Inflesz	68.3247	Bastante fácil
Muñoz Muñoz	62.9459	Adecuado
Crawford	4.94	Años de escolarización

Medida	Valor
Número de Nodos	482
Número de Aristas	813
Grado Medio	3.3734
Longitud Media Del Camino	3.9485
Agrupación Media	0.0542
Modularidad	0.5706

Texto 3

Algoritmo	Valor	Interpretación
Fernández Huerta	70.9718	Bastante fácil
Gutiérrez Polini	44.8545	Normal
Szigriszt-Pazos	66.8	Bastante fácil
Inflesz	66.8	Bastante fácil
Muñoz Muñoz	58.9601	Un poco difícil
Crawford	5.06	Años de escolarización

Medida	Valor
Número de Nodos	409
Número de Aristas	664
Grado Medio	3.2469
Longitud Media Del Camino	3.9648
Agrupación Media	0.0935
Modularidad	0.5759

Texto 4

Algoritmo	Valor	Interpretación
Fernández Huerta	73.9316	Bastante fácil
Gutiérrez Polini	46.4702	Normal
Szigriszt-Pazos	70.0542	Bastante fácil
Inflesz	70.0542	Bastante fácil
Muñoz Muñoz	71.7404	Un poco fácil
Crawford	4.73	Años de escolarización

Medida	Valor
Número de Nodos	278
Número de Aristas	468
Grado Medio	3.3669
Longitud Media Del Camino	3.869
Agrupación Media	0.0486
Modularidad	0.5789

Texto 5

Algoritmo	Valor	Interpretación
Fernández Huerta	65.0135	Normal
Gutiérrez Polini	43.7604	Normal
Szigriszt-Pazos	60.3554	Normal
Inflesz	60.3554	Normal
Muñoz Muñoz	55.3442	Un poco difícil
Crawford	5.52	Años de escolarización

Medida	Valor
Número de Nodos	213
Número de Aristas	372
Grado Medio	3.493
Longitud Media Del Camino	3.6425
Agrupación Media	0.0399
Modularidad	0.5478

Texto 6

Algoritmo	Valor	Interpretación
Fernández Huerta	77.839	Bastante fácil
Gutiérrez Polini	46.5713	Normal
Szigriszt-Pazos	73.7132	Bastante fácil
Inflesz	73.7132	Bastante fácil
Muñoz Muñoz	58.2562	Un poco difícil
Crawford	4.5	Años de escolarización

Medida	Valor
Número de Nodos	368
Número de Aristas	609
Grado Medio	3.3098
Longitud Media Del Camino	3.8266
Agrupación Media	0.0615
Modularidad	0.574

Texto 7

Algoritmo	Valor	Interpretación
Fernández Huerta	76.8823	Bastante fácil
Gutiérrez Polini	47.6563	Normal
Szigriszt-Pazos	72.8889	Bastante fácil
Inflesz	72.8889	Bastante fácil
Muñoz Muñoz	65.8144	Adecuado
Crawford	4.59	Años de escolarización

Medida	Valor
Número de Nodos	269
Número de Aristas	450
Grado Medio	3.3457
Longitud Media Del Camino	3.8125
Agrupación Media	0.0602
Modularidad	0.5828

Texto 8

Algoritmo	Valor	Interpretación
Fernández Huerta	70.6105	Bastante fácil
Gutiérrez Polini	45.4939	Normal
Szigriszt-Pazos	66.3024	Bastante fácil
Inflesz	66.3024	Bastante fácil
Muñoz Muñoz	65.5679	Adecuado
Crawford	5.1	Años de escolarización

Medida	Valor
Número de Nodos	252
Número de Aristas	392
Grado Medio	3.1111
Longitud Media Del Camino	4.0724
Agrupación Media	0.0381
Modularidad	0.5848

Texto 9

Algoritmo	Valor	Interpretación
Fernández Huerta	67.528	Normal
Gutiérrez Polini	43.9485	Normal
Szigriszt-Pazos	63.2372	Normal
Inflesz	63.2372	Normal
Muñoz Muñoz	60.0256	Un poco difícil
Crawford	5.34	Años de escolarización

Medida	Valor
Número de Nodos	231
Número de Aristas	362
Grado Medio	3.1342
Longitud Media Del Camino	4.1257
Agrupación Media	0.0397
Modularidad	0.5721

Texto 10

Algoritmo	Valor	Interpretación
Fernández Huerta	68.7985	Normal
Gutiérrez Polini	44.1096	Normal
Szigriszt-Pazos	64.6517	Normal
Inflesz	64.6517	Normal
Muñoz Muñoz	59.4637	Un poco difícil
Crawford	5.19	Años de escolarización

Medida	Valor
Número de Nodos	351
Número de Aristas	608
Grado Medio	3.4644
Longitud Media Del Camino	3.6353
Agrupación Media	0.1141
Modularidad	0.5401

Valores de los textos difíciles devueltos por la aplicación con *stop words*.

Texto 1

Algoritmo	Valor	Interpretación
Fernández Huerta	52.958	Bastante difícil
Gutiérrez Polini	37.4585	Normal
Szigriszt-Pazos	48.698	Bastante difícil
Inflesz	48.698	Algo difícil
Muñoz Muñoz	46.6552	Difícil
Crawford	6.11	Años de escolarización

Medida	Valor
Número de Nodos	283
Número de Aristas	512
Grado Medio	3.6184
Longitud Media Del Camino	3.6406
Agrupación Media	0.1348
Modularidad	0.5168

Texto 2

Algoritmo	Valor	Interpretación
Fernández Huerta	49.534	Difícil
Gutiérrez Polini	35.665	Normal
Szigriszt-Pazos	44.9749	Bastante difícil
Inflesz	44.9749	Algo difícil
Muñoz Muñoz	45.7089	Difícil
Crawford	6.54	Años de escolarización

Medida	Valor
Número de Nodos	195
Número de Aristas	322
Grado Medio	3.3026
Longitud Media Del Camino	3.7496
Agrupación Media	0.0943
Modularidad	0.5336

Texto 3

Algoritmo	Valor	Interpretación
Fernández Huerta	44.5014	Difícil
Gutiérrez Polini	33.5292	Normal
Szigriszt-Pazos	39.5879	Bastante difícil
Inflesz	39.5879	Muy difícil
Muñoz Muñoz	37.2352	Difícil
Crawford	7.08	Años de escolarización

Medida	Valor
Número de Nodos	156
Número de Aristas	256
Grado Medio	3.2821
Longitud Media Del Camino	3.4519
Agrupación Media	0.0918
Modularidad	0.5228

Texto 4

Algoritmo	Valor	Interpretación
Fernández Huerta	35.603	Difícil
Gutiérrez Polini	29.1068	Difícil
Szigriszt-Pazos	31.3833	Árido
Inflesz	31.3833	Muy difícil
Muñoz Muñoz	45.4455	Difícil
Crawford	6.75	Años de escolarización

Medida	Valor
Número de Nodos	110
Número de Aristas	174
Grado Medio	3.1636
Longitud Media Del Camino	3.6068
Agrupación Media	0.0731
Modularidad	0.5225

Texto 5

Algoritmo	Valor	Interpretación
Fernández Huerta	32.615	Difícil
Gutiérrez Polini	27.8933	Difícil
Szigriszt-Pazos	27.452	Árido
Inflesz	27.452	Muy difícil
Muñoz Muñoz	31.3896	Difícil
Crawford	7.88	Años de escolarización

Medida	Valor
Número de Nodos	231
Número de Aristas	354
Grado Medio	3.0649
Longitud Media Del Camino	4.4678
Agrupación Media	0.0712
Modularidad	0.5651

Texto 6

Algoritmo	Valor	Interpretación
Fernández Huerta	42.0227	Difícil
Gutiérrez Polini	32.8809	Difícil
Szigriszt-Pazos	37.2028	Bastante difícil
Inflesz	37.2028	Muy difícil
Muñoz Muñoz	34.0697	Difícil
Crawford	7.13	Años de escolarización

Medida	Valor
Número de Nodos	124
Número de Aristas	197
Grado Medio	3.1774
Longitud Media Del Camino	3.4277
Agrupación Media	0.1229
Modularidad	0.5184

Texto 7

Algoritmo	Valor	Interpretación
Fernández Huerta	24.9542	Muy difícil
Gutiérrez Polini	24.5969	Difícil
Szigriszt-Pazos	20.1721	Árido
Inflesz	20.1721	Muy difícil
Muñoz Muñoz	37.1957	Difícil
Crawford	7.8	Años de escolarización

Medida	Valor
Número de Nodos	124
Número de Aristas	197
Grado Medio	3.1774
Longitud Media Del Camino	3.8162
Agrupación Media	0.1178
Modularidad	0.5468

Texto 8

Algoritmo	Valor	Interpretación
Fernández Huerta	44.1313	Difícil
Gutiérrez Polini	32.7517	Difícil
Szigriszt-Pazos	39.8012	Bastante difícil
Inflesz	39.8012	Muy difícil
Muñoz Muñoz	44.0701	Difícil
Crawford	6.55	Años de escolarización

Medida	Valor
Número de Nodos	240
Número de Aristas	424
Grado Medio	3.5333
Longitud Media Del Camino	3.5914
Agrupación Media	0.1275
Modularidad	0.5067

Texto 9

Algoritmo	Valor	Interpretación
Fernández Huerta	52.7006	Bastante difícil
Gutiérrez Polini	35.4649	Normal
Szigriszt-Pazos	48.2163	Bastante difícil
Inflesz	48.2163	Algo difícil
Muñoz Muñoz	42.6541	Difícil
Crawford	6.32	Años de escolarización

Medida	Valor
Número de Nodos	147
Número de Aristas	232
Grado Medio	3.1565
Longitud Media Del Camino	3.8744
Agrupación Media	0.0877
Modularidad	0.5423

Texto 10

Algoritmo	Valor	Interpretación
Fernández Huerta	29.846	Muy difícil
Gutiérrez Polini	28.0807	Difícil
Szigriszt-Pazos	25.2401	Árido
Inflesz	25.2401	Muy difícil
Muñoz Muñoz	39.2461	Difícil
Crawford	7.41	Años de escolarización

Medida	Valor
Número de Nodos	143
Número de Aristas	253
Grado Medio	3.5385
Longitud Media Del Camino	3.4909
Agrupación Media	0.159
Modularidad	0.4926

Valores de los textos fáciles devueltos por la aplicación sin *stop words*.

Texto 1

Medida	Valor	Interpretación
Número de Nodos	421	Número de palabras
Número de Aristas	436	Número de uniones
Grado Medio	2.0713	Número de palabras
Longitud Media Del Camino	11.4354	Conectividad de la red
Agrupación Media	0.0	Agrupación de la red
Modularidad	0.8734	Fuerza de división de la red

Texto 2

Medida	Valor	Interpretación
Número de Nodos	386	Número de palabras
Número de Aristas	425	Número de uniones
Grado Medio	2.2021	Número de palabras
Longitud Media Del Camino	3.6667	Conectividad de la red
Agrupación Media	0.0	Agrupación de la red
Modularidad	0.4259	Fuerza de división de la red

Texto 3

Medida	Valor	Interpretación
Número de Nodos	309	Número de palabras
Número de Aristas	318	Número de uniones
Grado Medio	2.0583	Número de palabras
Longitud Media Del Camino	10.781	Conectividad de la red
Agrupación Media	0.0132	Agrupación de la red
Modularidad	0.8274	Fuerza de división de la red

Texto 4

Medida	Valor	Interpretación
Número de Nodos	203	Número de palabras
Número de Aristas	229	Número de uniones
Grado Medio	2.2562	Número de palabras
Longitud Media Del Camino	8.0394	Conectividad de la red
Agrupación Media	0.0052	Agrupación de la red
Modularidad	0.7838	Fuerza de división de la red

Texto 5

Medida	Valor	Interpretación
Número de Nodos	164	Número de palabras
Número de Aristas	186	Número de uniones
Grado Medio	2.2683	Número de palabras
Longitud Media Del Camino	6.291	Conectividad de la red
Agrupación Media	0.0143	Agrupación de la red
Modularidad	0.7368	Fuerza de división de la red

Texto 6

Medida	Valor	Interpretación
Número de Nodos	154	Número de palabras
Número de Aristas	203	Número de uniones
Grado Medio	2.6364	Número de palabras
Longitud Media Del Camino	4.8588	Conectividad de la red
Agrupación Media	0.0217	Agrupación de la red
Modularidad	0.6505	Fuerza de división de la red

Texto 7

Medida	Valor	Interpretación
Número de Nodos	277	Número de palabras
Número de Aristas	308	Número de uniones
Grado Medio	2.2238	Número de palabras
Longitud Media Del Camino	8.0431	Conectividad de la red
Agrupación Media	0.0174	Agrupación de la red
Modularidad	0.7986	Fuerza de división de la red

Texto 8

Medida	Valor	Interpretación
Número de Nodos	200	Número de palabras
Número de Aristas	223	Número de uniones
Grado Medio	2.23	Número de palabras
Longitud Media Del Camino	7.9238	Conectividad de la red
Agrupación Media	0.0063	Agrupación de la red
Modularidad	0.8102	Fuerza de división de la red

Texto 9

Medida	Valor	Interpretación
Número de Nodos	184	Número de palabras
Número de Aristas	195	Número de uniones
Grado Medio	2.1196	Número de palabras
Longitud Media Del Camino	9.6984	Conectividad de la red
Agrupación Media	0.0028	Agrupación de la red
Modularidad	0.8169	Fuerza de división de la red

Texto 10

Medida	Valor	Interpretación
Número de Nodos	168	Número de palabras
Número de Aristas	182	Número de uniones
Grado Medio	2.1667	Número de palabras
Longitud Media Del Camino	7.4446	Conectividad de la red
Agrupación Media	0.0151	Agrupación de la red
Modularidad	0.7648	Fuerza de división de la red

Valores de los textos difíciles devueltos por la aplicación sin *stop words*.

Texto 1

Medida	Valor	Interpretación
Número de Nodos	226	Número de palabras
Número de Aristas	282	Número de uniones
Grado Medio	2.4956	Número de palabras
Longitud Media Del Camino	6.8014	Conectividad de la red
Agrupación Media	0.0355	Agrupación de la red
Modularidad	0.7646	Fuerza de división de la red

Texto 2

Medida	Valor	Interpretación
Número de Nodos	150	Número de palabras
Número de Aristas	161	Número de uniones
Grado Medio	2.1467	Número de palabras
Longitud Media Del Camino	8.9612	Conectividad de la red
Agrupación Media	0.0	Agrupación de la red
Modularidad	0.7614	Fuerza de división de la red

Texto 3

Medida	Valor	Interpretación
Número de Nodos	118	Número de palabras
Número de Aristas	142	Número de uniones
Grado Medio	2.4068	Número de palabras
Longitud Media Del Camino	6.1541	Conectividad de la red
Agrupación Media	0.0192	Agrupación de la red
Modularidad	0.669	Fuerza de división de la red

Texto 4

Medida	Valor	Interpretación
Número de Nodos	183	Número de palabras
Número de Aristas	215	Número de uniones
Grado Medio	2.3497	Número de palabras
Longitud Media Del Camino	8.5435	Conectividad de la red
Agrupación Media	0.0056	Agrupación de la red
Modularidad	0.7587	Fuerza de división de la red

Texto 5

Medida	Valor	Interpretación
Número de Nodos	84	Número de palabras
Número de Aristas	79	Número de uniones
Grado Medio	1.881	Número de palabras
Longitud Media Del Camino	8.2462	Conectividad de la red
Agrupación Media	0.0	Agrupación de la red
Modularidad	0.6691	Fuerza de división de la red

Texto 6

Medida	Valor	Interpretación
Número de Nodos	94	Número de palabras
Número de Aristas	91	Número de uniones
Grado Medio	1.9362	Número de palabras
Longitud Media Del Camino	2.3333	Conectividad de la red
Agrupación Media	0.0	Agrupación de la red
Modularidad	0.26	Fuerza de división de la red

Texto 7

Medida	Valor	Interpretación
Número de Nodos	88	Número de palabras
Número de Aristas	105	Número de uniones
Grado Medio	2.3864	Número de palabras
Longitud Media Del Camino	5.6996	Conectividad de la red
Agrupación Media	0.014	Agrupación de la red
Modularidad	0.6671	Fuerza de división de la red

Texto 8

Medida	Valor	Interpretación
Número de Nodos	182	Número de palabras
Número de Aristas	243	Número de uniones
Grado Medio	2.6703	Número de palabras
Longitud Media Del Camino	6.1543	Conectividad de la red
Agrupación Media	0.0395	Agrupación de la red
Modularidad	0.7248	Fuerza de división de la red

Texto 9

Medida	Valor	Interpretación
Número de Nodos	110	Número de palabras
Número de Aristas	112	Número de uniones
Grado Medio	2.0364	Número de palabras
Longitud Media Del Camino	10.0181	Conectividad de la red
Agrupación Media	0.0	Agrupación de la red
Modularidad	0.7388	Fuerza de división de la red

Texto 10

Medida	Valor	Interpretación
Número de Nodos	103	Número de palabras
Número de Aristas	136	Número de uniones
Grado Medio	2.6408	Número de palabras
Longitud Media Del Camino	4.933	Conectividad de la red
Agrupación Media	0.0366	Agrupación de la red
Modularidad	0.6819	Fuerza de división de la red