Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

Análisis de expresiones de búsqueda relacionadas con industrias culturales en un motor de búsqueda

Trabajo Fin de Máster

Máster Universitario en Gestión de la Información

Autor: Asier Dasí Osca

Tutor: Jorge Ignacio Serrano-Cobos
María de los Ángeles Calduch Losa

Curso 2019-2020

Resumen

En este En este trabajo se ha desarrollado un análisis sobre un listado de expresiones de búsqueda relacionas con industrias culturales en un motor de búsqueda. El objetivo principal de este estudio es encontrar una metodología de trabajo efectiva sobre conjuntos masivos de datos consistentes en expresiones de búsqueda o palabras clave extraídas de motores de búsqueda, con independencia de su temática. Para llevarlo a cabo se partió de un listado de expresiones de búsqueda relacionas con industrias culturales extraídas de Google mediante el uso de la herramienta Google Keyword Planner. Tras probar diferentes conjuntos de datos y herramientas especializadas en análisis de grafos, finalmente se decidió trabajar con una muestra aleatoria del 5% de los datos originales y el programa Gephi. A partir de esta muestra se creó una matriz que enfrentaba cada búsqueda de la muestra con el resto de búsquedas y recogía el número de palabras que coincidían en cada caso. Además, se eliminó la diagonal de la matriz y los conectores más comunes de las búsquedas para evitar sesgos y ruido. Con esta matriz y mediante el algoritmo Fruchterman Reingold se obtuvo un grafo formado por 1.506 nodos y 28.242 aristas que contenía 27 comunidades, siendo la comunidad más grande y céntrica, la correspondiente al conjunto formado por las expresiones contenedoras de la palabra clave “libros”. Dados los resultados, les puede considerar que la metodología final propuesta es efectiva y cabría tenerla en cuenta para poder replicarla en el futuro a una escala mayor.

Palabras clave: expresiones de búsqueda, palabras clave, posicionamiento SEO, análisis de grafos, literatura, análisis textual, análisis de industrias culturales, cibermetría.

Abstract

In this paper an analysis has been developed on a list of queries related to cultural industries in a search engine. The main objective of the study is to find an effective working methodology on massive datasets consisting of queries or keywords extracted from search engines, regardless of their subject matter. To carry it out, we started from a list of search expressions related to cultural industries extracted from Google using the Google Keyword Planner tool. After testing different data sets and specialized tools in graph analysis, it was decided to work with a random sample of a 5% of the original dataset and the Gephi software. From this sample, a matrix was created that compared each query of the sample with the rest of the queries and collected the number of words that matched in each case. In addition, the diagonal of the matrix and the most common connectors of the searches were eliminated to avoid bias and noise. With this matrix and working with the Fruchterman Reingold algorithm, a graph formed by 1,506 nodes and 28,242 edges was obtained that contained 27 communities, where the largest and most central community being the one corresponding to the set formed by the expressions containing the keyword “libros”. Given the results, it can be considered that the final proposed methodology is effective and should be taken into account to be able to replicate it in the future on a larger scale.

Keywords: queries, keywords, SEO positioning, graph analysis, literature, textual analysis, analysis of cultural industries, cybermetrics.

Tabla de contenidos

Tabla de ilustraciones

1. Introducción

El conjunto de la sociedad de hoy en día vive en un mundo en el que cada vez más gente, sin importar la edad ni cualquier otro factor que pueda llegar a chocar con las tecnologías, decide informarse por su parte navegando en Internet, donde todo tipo de información se pone a disposición de este, casi como en escaparates a los que se accede a través de los Sistemas de Recuperación de Información (en adelante SRI), concretamente de los SRI llamados motores de búsqueda o buscadores de Internet. Ejemplos de estos son los clásicos Google, Yahoo! Search, Bing, etc.

Pero antes de continuar, se va explicar brevemente lo que es la Recuperación de Información (en adelante RI). Diversos autores han dado su propia definición del concepto, mostrándose a continuación una breve selección de estos (**Bordignon y Tolosa Chacón**, 2007):

- "La Recuperación de Información trata con la representación, el almacenamiento, la organización y el acceso a ítems de información" (**Baeza-Yates y Ribeiro-Neto**, 1999).

- "El conjunto de tareas mediante las cuales el usuario localiza y accede a los recursos de información que son pertinentes para la resolución del problema planteado. En estas tareas desempeñan un papel fundamental los lenguajes documentales, las técnicas de resumen, la descripción del objeto documental, etc." (**Croft**, 1987).

- "La localización y presentación a un usuario de información relevante a una necesidad de información expresada como una pregunta" (**Korfhage**, 1997).

Por lo tanto, de estas definiciones, se puede concluir que los SRI se dedican a recuperar aquellos documentos que, teniendo en cuenta la expresión de búsqueda o consulta (*query*) llevada a cabo por un usuario, puedan adaptarse a esta, pudiendo ser tanto de importancia como irrelevantes para la necesidad de información que ha llevado al individuo a utilizar un SRI.

Para entender mejor la definición explicada en el párrafo anterior, a continuación se muestran las definiciones de los principales conceptos contenidos en ella de forma sintetizada (**Vilares Ferro**, 2005):

- **Documento:** Unidad de texto almacenado por el sistema y disponible para la recuperación.
- **Colección:** Repositorio de documentos disponible para resolver las necesidades de información del usuario.
- **Término:** Cada una de las unidades léxicas que componen un documento.
- **Consulta (*Query*):** El término más importante para esta investigación, es la necesidad de información del usuario, expresada en términos que el sistema pueda comprender.
- **Ordenación (*Ranking*):** Ordenación por grado de similaridad o relevancia de los resultados obtenidos respecto a la consulta.
- **Relevancia:** Concepto subjetivo, ya que depende del usuario en sí, que hace referencia a la relevancia del documento devuelto respecto a su necesidad de información original.

De la facilidad de acceso a todo tipo de información mencionada anteriormente, nace el problema de la infoxicación o sobrecarga de información (*information overload*), utilizado por autores como **Alvin Toffler** (1973) en su libro *El "shock" del futuro*, es decir, que dicha facilidad excede las capacidades de raciocinio humanas. Esto significa que el usuario actual de Internet no es capaz de asimilar ni filtrar la información que pasa ante sus ojos de forma continua y, mucho menos, de profundizar en ella e informarse auténticamente. Este hecho, en palabras del *Informe Ciber*, aunque estudiado en jóvenes, se declara como que la velocidad y facilidad de búsqueda de estos en la Web es inversamente proporcional a la calidad de la información recabada, ya que implica una escasa evaluación de esta (**British Library y Joint Information Systems Committee [JISC]**, 2008).

Para esto se intenta adaptar los motores de búsqueda a los usuarios estudiando el comportamiento de estos en aquellos, estudiando sus necesidades de información. Dichas necesidades de información, ya mencionadas anteriormente, hacen referencia a las preguntas, dudas o problemas que se formulan los usuarios en su mente (ideas, percepciones, acciones, sucesos, datos, etc.), en cualquier tipo de situación, y que desencadenan el proceso que les llevará a satisfacerlas (investigación), comenzando por definir la respectiva necesidad, es decir, comprenderla y darle forma (**Medina Hernández**, 2012).

Una vez precisada la necesidad de información y establecidos los criterios correspondientes, el usuario formula una expresión de búsqueda o consulta (*query*), que es la cadena de palabras clave que el usuario inserta en cualquier, en el caso de este trabajo, motor de búsqueda, esperando resultados que satisfagan su necesidad de información, aunque esta no se vea reflejada de forma exacta en su consulta, pudiendo llegar a utilizar más de una expresión de búsqueda para una única necesidad. Estas expresiones de búsqueda de los usuarios se clasifican en tres tipos, dependiendo de la intención con las que se realizan (**Ordoñez**, 2016):

- **Navegacionales:** el usuario está intentando llegar a un sitio determinado. Busca páginas concretas, nombres, marcas, etc.
- **Informacionales:** el usuario está buscando información.
- **Transaccionales:** el usuario está buscando comprar.

A la consulta realizada, el SRI, en este caso Google, ofrecerá como resultado una página "personalizada" con un conjunto de enlaces a documentos que se adapten a ella. Esta página de resultados recibe el nombre de *Search Engine Result Page* (en adelante, SERP).

> "Cuando el usuario hace clic en un botón de búsqueda, el algoritmo entonces examina la información almacenada en la base de datos de fondo y recupera enlaces a las páginas web que parecen coincidir con los términos de búsqueda que el usuario ha ingresado" (**Ledford**, 2008).

A continuación, ya explicados los conceptos anteriores, cabe destacar que se deben evitar confusiones bastante comunes entre algunos términos. Principalmente, no se tiene que confundir el concepto de expresión de búsqueda o consulta (*query*) con el de necesidad de información, que, como se ha visto, el primero se refiere a la necesidad de información plasmada en palabras clave para su utilización en el motor de búsqueda y, el segundo, a las formulaciones mentales de los usuarios. También suele errarse la definición de consulta con la de palabra clave, siendo esta última el concepto ideal para representar la necesidad. Entre estos dos conceptos, se debe aclarar, igualmente, que la expresión de consulta siempre está escrita en lenguaje natural, es decir, con faltas de ortografía, diferentes conjugaciones e interpretaciones erróneas, mientras que la palabra clave es la representación exacta de una idea, objeto, concepto, etc.

El lenguaje natural, cómo lo define el profesor y lingüista **John Lyons** (1991), es "cualquier lenguaje que ha evolucionado de forma natural en los seres humanos, a través de su uso y la repetición sin planificación consciente o premeditación". Cada una de sus múltiples variantes tiene sus propias normas, gramática, usos, formalidades y coloquialismos, finalidades, etc. Es por esto que, por ejemplo, el filósofo lingüista **Ludwig Wittgenstein** (1988) lo ha llegado a definir, en su investigación filosófica número 18, como "una vieja ciudad: una maraña de callejas y plazas, de viejas y nuevas casas, y de casas con anexos de diversos períodos; y esto rodeado de un conjunto de barrios nuevos con calles rectas y regulares y con casas uniformes".

Un lenguaje artificial se diferencia del anterior, principalmente, por el hecho de haber sido creado específicamente para fines técnicos. Esto significa que no es innato del ser humano y que puede conllevar ciertas dificultades, dada su utilización metódica y el carácter altamente restrictivo que suelen tener los medios receptores de estos tipos de lenguajes, como por ejemplo, en el caso de los lenguajes de consulta, pueden ser SQL (*Structured Query Language*), SPARQL (*SPARQL Protocol and RDF Query Language*), XQuery o OPath, entre otros.

Otra de las principales características diferenciales a destacar entre ellos, es que en los lenguajes artificiales, cada símbolo se ha generado para un significado específico invariable. Por el contrario, en el lenguaje natural, un mismo símbolo o signo lingüístico puede llegar a tener varios significados muy distintos, es decir, que sean términos de condición polisémica. Esta característica permite entender que los lenguajes artificiales pierdan la espontaneidad del lenguaje natural, por lo tanto, siguiendo la metáfora de **Wittgenstein** mencionada anteriormente, estos vendrían a ser los "barrios nuevos con calles rectas y regulares y con casas uniformes" que rodean a la vieja ciudad principal, derivadas de ella, representando esta al lenguaje natural.

Por mucho que la característica presentada en el párrafo anterior pueda parecer una ventaja del lenguaje natural frente a los lenguajes artificiales, para los motores de búsqueda que trabajan con este lenguaje, resulta ser todo lo contrario. Dicha polisemia mencionada, junto con los también referidos anteriormente errores ortográficos y posibles interpretaciones erróneas del propio usuario, derivan en situaciones de ambigüedad.

Es así como se llega a la conclusión de que la ambigüedad del lenguaje natural es el verdadero reto para los motores de búsqueda, por lo que, actualmente, el hecho de que los SRI puedan llegar a "entender" y procesar de forma correcta y precisa

cualquier tipo de consulta de este tipo realizada por los usuarios sigue siendo un campo de considerable importancia.

Es en este punto donde entra la investigación presente, en la cual se ha trabajado sobre un pequeño conjunto de las definidas anteriormente expresiones de búsqueda contenidas en el ámbito de las industrias culturales, más concretamente, en el de la literatura. Dicho conjunto de consultas han sido realizadas en el motor de búsqueda Google, uno de los más importantes en la actualidad, si no el más importante, gracias a funcionalidades enfocadas a la RI como su búsqueda predictiva (*Google Instant*), la información siempre actualizada o las distintas visualizaciones gráficas de información que implementa, ya sean los resúmenes en el lateral, las imágenes o los videos (**Viñas**, 2015).

Sobre este conjunto de datos se laboró en la búsqueda de una metodología de trabajo para localizar relaciones entre las mencionadas consultas que lo componen, buscando encontrar intenciones de búsqueda de los usuarios en dicho navegador y realizar algún tipo de clasificación para ellas. Con la metodología resultante se pretende obtener resultados (entender el comportamiento de usuarios) que, en un futuro, puedan ser aplicados, en el campo del marketing digital, a la mejora del posicionamiento SEO (*Search Engine Optimization*) de páginas web, es decir, lograr que páginas web, donde se apliquen cambios a partir de resultados extraídos con dicha metodología, aparezcan en posiciones más altas en la página SERP de Google en cuanto usuarios realicen consultas sobre temas relacionados y así se consigan más visitas (Alonso, 2020), presentando también, resultados de mejor calidad desde el punto de vista de los usuarios.

2. Objetivos

2.1. Objetivo principal

Encontrar una metodología de trabajo efectiva sobre conjuntos masivos de datos consistentes en expresiones de búsqueda o palabras clave extraídas de motores de búsqueda, con independencia de su temática. Dado el escaso estado del arte encontrado en cuanto al análisis de redes sociales de palabras clave, se estima necesaria una iniciativa de este tipo.

2.2. Objetivos específicos

2.2.1. Objetivo específico #1

Analizar de forma básica, de las seleccionadas en un primer momento tras diversas búsquedas de comparativas en cuanto a requisitos y funcionalidades, la mejor herramienta para trabajar en el campo del presente trabajo, concretamente, en la sección de la generación y análisis de grafos, siendo estas: Pajek, SocNetV, Gephi, Neo4j y UCINET.

2.2.2. Objetivo específico #2

Clasificar las expresiones de la muestra a analizar según los tipos de intención de búsqueda del usuario descritos en la introducción, a recordar:

- Informacionales: el usuario busca información.
- Navegacionales: el usuario intenta llegar a una página.
- Transaccionales: el usuario busca comprar.

2.2.3. Objetivo específico #3

Examinar la tipología de las palabras (verbos, sustantivos, nombres propios, etc.) que conforman las expresiones analizadas y en qué formas se utilizan con más frecuencia dentro de la muestra.

2.2.4. Objetivo específico #4

Encontrar algún tipo de relación, o vínculo común, entre los clústeres más grandes aparecidos del análisis de la muestra seleccionada.

3. Estado del arte

3.1. Marketing, SEO y SEM

Para empezar, el campo más amplio y general dónde se puede ver integrada la presente investigación sería el campo del marketing. Para definir el marketing como concepto, una de las definiciones más acertadas es la aportada por los profesores **Philip Kotler y Gary Armstrong** (2013), cuyos libros y artículos sobre el ámbito se utilizan en gran cantidad de escuelas económicas y, concretamente los de **Kotler**, han sido premiados en múltiples ocasiones. Dicha definición reza de la siguiente forma: "proceso mediante el cual las empresas crean valor para sus clientes y generan fuertes relaciones con ellos para, en reciprocidad, captar el valor de sus clientes". En dicho libro, se describe, como uno de los pasos iniciales del proceso de marketing, una separación de conceptos a identificar en cuanto a los clientes entre necesidades, deseos y demandas, definiéndolos, el primero como estados de carencia percibida (necesidades físicas, sociales e individuales), el segundo como las necesidades ya procesadas por el individuo y, por último, el tercero como los deseos descritos en términos de los objetos y, adaptados a los recursos de los individuos, apoyados por el poder de compra. Como se puede observar, se establece un cierto paralelismo, aunque no exacto, entre estas definiciones y las planteadas anteriormente, en el apartado de Introducción (ver **1. Introducción**) del presente trabajo, para el proceso de transformación de las necesidades de información de los usuarios en expresiones de búsqueda o consultas (*queries*). Los pasos que, junto al que incluye los conceptos anteriores, completan el proceso de marketing descritos en el libro son los siguientes, sin extenderse mucho en ellos:

1. Comprender las necesidades del mercado y de los clientes.
2. Plantear y crear una estrategia de marketing de cara al cliente.
3. Crear un plan de marketing.
4. Crear y administrar relaciones provechosas con los clientes.

Estas explicaciones se han planteado desde el punto de vista del marketing como disciplina general, mostrando la importancia del campo en sí para todo tipo de

empresas, industrias, instituciones, negocios, proyectos, etc. Actualmente, se debe puntualizar que, gracias al crecimiento de las nuevas tecnologías de la información y la comunicación (TICs), y principalmente, de Internet, una gran parte de los objetivos de dicho campo se han tenido que adaptar a lo que se denomina el marketing digital. El marketing digital se puede definir como "el conjunto de herramientas y estrategias digitales que ayudan a solucionar una necesidad de mercado generando beneficios" (**Membiela-Pollán y Pedreira Fernández**, 2019) y no se tiene que confundir con el concepto de comercio electrónico (*e-commerce*), dado que este aparece con la realización de cualquier intercambio, ya sea económico o de datos, entre el cliente y la empresa (**Chaffey y Ellis-Chadwick**, 2014). En su artículo, **Membiela-Pollán y Pedreir-Fernández** (2019) desglosan el conjunto de herramientas y estrategias digitales mencionado en su definición de marketing digital de la siguiente forma:

- Web corporativa y tienda online.
- Blogs.
- Redes sociales.
    - RRSS basadas en el perfil.
    - RRSS visuales.
- E-mail Marketing.
- SEO
- SEM
- Publicidad digital.
- Otras.

Para la investigación presente, de esta lista de herramientas y técnicas, se prosigue a profundizar en el estado del arte alrededor del proceso SEO (*Search Engine Optimization*), no sin antes aclarar sus diferencias con las técnicas SEM (*Search Engine Marketing*), ya que son conceptos que suelen ser mencionados en conjunto. La principal diferencia radica en la implicación monetaria del implementador de estas técnicas. Así como el SEO se centra en el posicionamiento de páginas web en las SERP de los motores de búsqueda mediante la optimización de distintas formas gratuitas centradas en los algoritmos de los motores de búsqueda, el SEM "te permite posicionarte mucho más rápido pero requiere una fuerte inversión" (**Moreno Pila**,

2017), es decir, lo hace gracias a la inversión de dinero. Dicha inversión se realiza, según Romero (2015), en pujas de Adwords en las que, dependiendo de la cantidad monetaria destinada, se determina la posición de las páginas y esto, sumado a factores compartidos con el SEO que se describen más adelante, hace del SEM la opción de posicionamiento y visibilidad más rápida. Esta característica define, a su vez, las dos áreas de las SERP, respectivamente, la de los resultados de la búsqueda u orgánicos (SEO) y la de los enlaces publicitados (SEM). Como ejemplo, en la **Ilustración 1** se pueden observar estas dos áreas diferenciadas en el buscador Google. Se ha elegido Google para el ejemplo por el hecho de que la presente investigación se centra en su motor de búsqueda en concreto, como ya se ha explicado en la Introducción (ver **1. Introducción**).

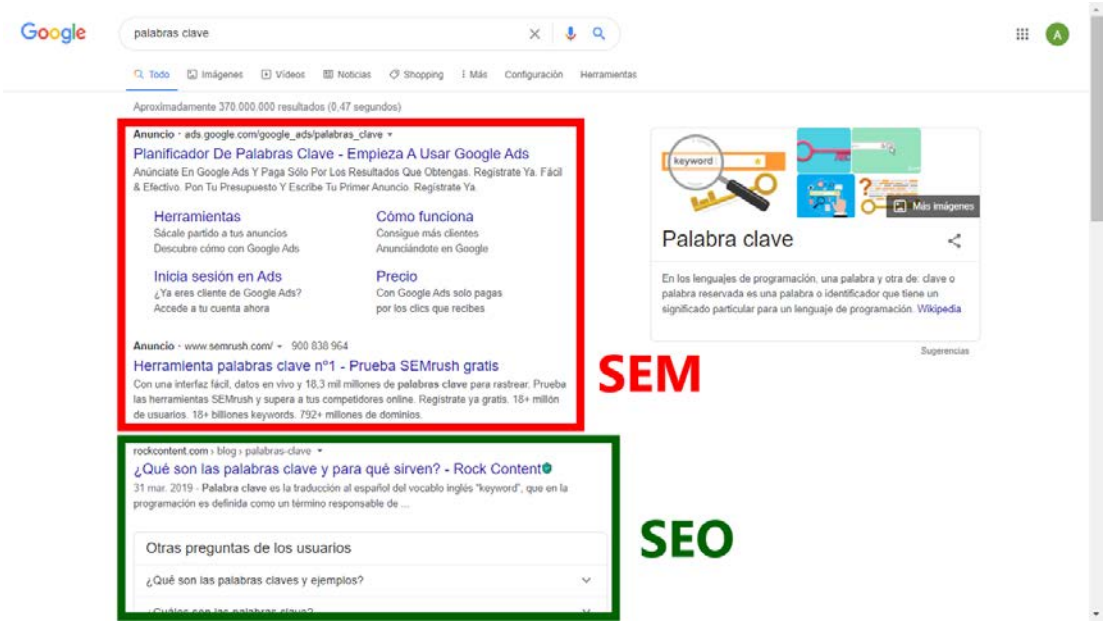


Ilustración 1. Áreas SEO y SEM de la SERP de Google. Fuente: elaboración propia

Según comentan **Sabaté Garriga, Berbegal Mirabent, Consolación y Cañabate Carmona** (2009) en su artículo sobre el SEO y la venta de libros, "conviene aclarar que toda estrategia de marketing en buscadores, sea de pago o SEO, debe partir de la identificación de las palabras clave que los clientes potenciales utilizarían cuando buscan información relacionada con la oferta de la empresa" ya sea optimizando las páginas o pagando, como se ha explicado anteriormente. El posicionamiento en las SERP resulta de vital importancia para las páginas web debido al comportamiento de los usuarios, los cuáles, casi en su totalidad, en el proceso de navegación sólo llegan a acceder a las primeras páginas que se muestran en dicha SERP (**Olaru**, 2019).

Dejando a un lado las técnicas SEM y centrando las explicaciones en el posicionamiento SEO, el motor de búsqueda Google tiene en consideración una cantidad de factores superior a doscientos para los resultados orgánicos. Aunque ya se ha definido en la Introducción, para continuar, el funcionamiento básico de Google, que consiste en la recepción de una expresión de consulta de parte de un usuario a la que responde generando la página SERP con los resultados, como apunta **Carreras Lario** (2012), más cercanos textualmente, dado que Google no entiende significados semánticos. Poco más se puede añadir sobre el funcionamiento de este motor de búsqueda dada su hermeticidad en cuanto a los algoritmos que usa o deja de usar, hecho entendible de cara a la competencia que pueda tener, por lo que los criterios que se enlistan a continuación, han sido determinados por expertos ajenos a la empresa en su mayoría.

Los más de doscientos factores mencionados, los cuales no se han incluido al completo para evitar extensión innecesaria, se pueden encontrar explicados brevemente uno por uno en la lista publicada y actualizada por **Dean** (2020). Dichos factores se encuentran divididos en los siguientes nueve grupos:

1. **Factores de dominio:** corresponde a los factores que versan sobre las características del dominio en el que encuentra la página web.
2. **Factores a nivel de página (contenido):** corresponde a los factores de contenido de las distintas etiquetas de la página web.
3. **Factores a nivel de sitio (alojamiento):** corresponde a los factores que atañen a la arquitectura, el servidor, los certificados, la privacidad y accesibilidad de la página web.
4. **Factores de *backlinks* (enlaces entrantes):** corresponde a los factores referentes a los enlaces entrantes de la página web y sus tipologías.
5. **Interacción del usuario (comportamiento):** corresponde a los factores que incumben al comportamiento del usuario en Internet, la navegación, el tráfico, el CTR (*Click Through Rate*) y páginas de interacción entre usuarios.
6. **Reglas especiales de algoritmos de Google:** corresponde a los factores más específicos tocantes a los diversos historiales del navegador, la diversidad de las SERPs y geolocalización, entre otros.
7. **Señales de marcas (*branding*):** corresponde a los factores sobre marcas y empresas.

8. **Factores de spam en el sitio web:** corresponde a los factores relacionados con prácticas fraudulentas (redirecciones, *content farms*, pop-ups, spam, etc.) y contenido de poca calidad.

9. **Factores de spam fuera del sitio web:** corresponde a los factores relacionados con prácticas fraudulentas externas a la web (hackeos, enlaces, etc.).

Analizando los factores de la lista de **Dean** (2020) junto con los factores elegidos por **Carreras Lario** (2012) o **Morato, Sánchez-Cuadrado, Moreno y Moreiro** (2012), entre otros autores, y los factores expuestos en la metodología de recomendación propuesta por **Injante y Mauricio** (2020), se puede comprobar la gran importancia que tienen las palabras clave para el posicionamiento SEO, siendo uno de los principales conceptos para la definición de páginas web.

Dado que ya se ha hablado, aunque de manera breve, de lo referente al comportamiento de los usuarios, las necesidades de información y las consultas en la Introducción y en párrafos anteriores de este mismo apartado, se procede a explicar en el apartado siguiente, con el estado del arte correspondiente, el campo del análisis de palabras clave.

3.2. Análisis de redes y palabras clave

Como se ha explicado ya la definición de las palabras clave y de las expresiones de búsqueda y sus tipos en la introducción (ver **1. Introducción**), este apartado se va a dedicar al estado del arte existente sobre el campo del análisis de palabras clave o expresiones de búsqueda y sus métodos.

Investigando sobre el tema, se encuentran estudios e investigaciones de este tipo sobre conjuntos de datos, redes sociales y similares que se acercan al tema del trabajo presente, como por ejemplo, en la comunidad hispanohablante, el estudio de **Russell, Madera Jaramillo y Ainsworth** (2009) sobre la comunicación científica entre países, el de **Molina** (2001) sobre la cultura en las organizaciones o el artículo de **Miguel, Caprile y Jorquera-Vidal** (2008) sobre la generación de mapas temáticos, entre otros.

En cuanto a la comunidad de habla inglesa, algunos artículos se acercan más que los anteriores, examinando tendencias de investigación en el artículo de **Kho, Cho y Cho** (2013) o temas y tendencias en Australia y Nueva Zelandia en cuanto al sector del turismo en el de **Benckendorff** (2009). Otro estudio que también se acerca al tema presente es el de **Chen, Chen, Wu, Xie y Li** (2016), en el que, como el primer artículo mencionado de habla inglesa, muestran grafos de tendencias de investigación.

Si se ordenan las búsquedas por fecha, de más recientes a más antiguos, aparecen estudios como el de **Spanou y Bekiari** (2020) analizando comportamientos destructivos en las universidades o el análisis de flujo de jugadores entre equipos, como en **Silva, Santos, Pinheiro da Silveira y Reis Mourao** (2020) entre otros.

A raíz de este estado del arte encontrado sobre este tipo de análisis de palabras clave, se decidió utilizar una metodología de análisis de redes sociales con grafos y aplicarla al ámbito de la literatura, correspondiente a la presente investigación, sobre el que no se encontraron prácticamente investigaciones.

Finalmente, con la elección de este tipo de metodología, se procedió a investigar el estado del arte correspondiente a la teoría de grafos, para entender sus fundamentos básicos y proceder con el trabajo.

3.3. Teoría de grafos

A continuación, se procede a explicar, dado que esta disciplina bebe de diversas áreas teóricas complejas y extensas englobadas dentro de las matemáticas discretas y las matemáticas aplicadas, la teoría de grafos según el estado del arte encontrado de una forma básica y sin profundizar en algoritmos. Igualmente como proviene de diversos campos matemáticos, los métodos y teoremas de la teoría de grafos se pueden utilizar en una gran variedad de sectores como son “teoría de la información, planificación de la producción, transportes, programación lineal, redes de conexión, mecánica estadística, genética y química, encontrándosele ahora un nuevo campo de aplicación: la Contabilidad” (**Millet Luaces, Beyris Bringuez y Rosales Almaguer**, 2012). Para la definición de conceptos, seguidamente se unifica, y complementa entre sí, la información encontrada en diversas fuentes, ya que suelen ser definiciones muy similares con cambios escasos aunque notables (**Chirinos**, 2010; **Montes**, s.f.; **Combariza**, 2003; **Croft**, 1987; **Villacañas Velasco**, 2014).

Para empezar, se entiende como grafo (*G*), cuya fórmula definitoria es *G*={*V, E*}, un conjunto de nodos o vértices (*V*) relacionados unos con otros, en el que dichas relaciones se encuentran representadas por líneas denominadas aristas (*E*) cuyos dos extremos son nodos. En la **Ilustración 2**, se observa una representación básica de un grafo formado por cinco nodos (*V*={V_1, V_2, V_3, V_4 y V_5}) unidos por cinco aristas (*E*={{V_1, V_3}, {V_1, V_4}, {V_2, V_4}, {V_3, V_5}, {V_4, V_5}}).

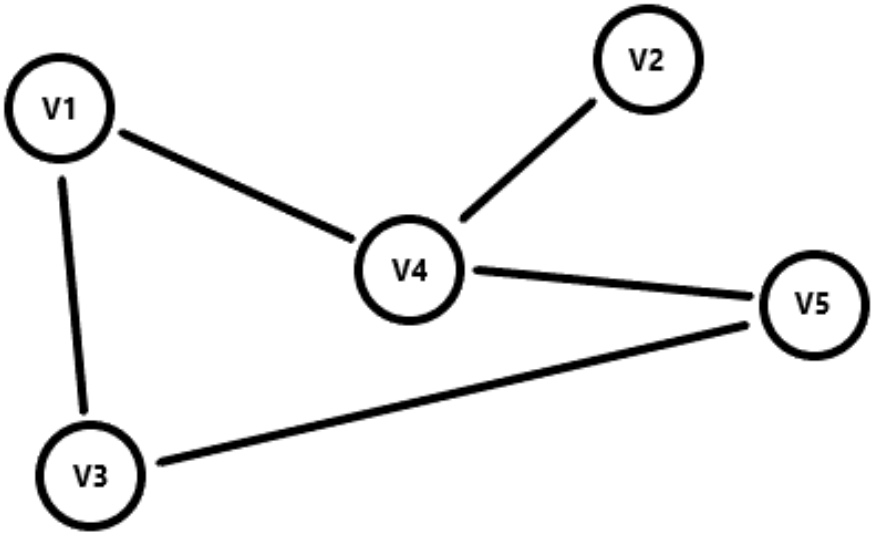


Ilustración 2. Representación básica de grafo. Fuente: elaboración propia

Los nodos (*V*) simplemente representan los datos o valores que se les puedan haber adjudicado, mientras que las aristas, que son las que verdaderamente definen los grafos, se pueden dividir en diversos tipos según sus características, cuyos ejemplos se observan en la **Ilustración 3**, igual a la anterior pero con un nodo (V_6) y dos aristas nuevas ({V_1, V_4} y {V_5, *V5*}):

- **Aristas adyacentes:** convergen en el mismo vértice. En la figura, son ejemplos de este tipo todas las aristas presentes, ya que todas convergen con alguna arista en alguno de sus vértices.
- **Aristas múltiples o paralelas:** comparten los mismos vértices inicial y final. En la figura, son ejemplos de este tipo las dos aristas *E*={V_1,V_4}.
- **Aristas cíclicas, lazos o bucles:** sus extremos son el mismo vértice. En la figura, se observa como ejemplo la arista *E*={V_5, V_5}.

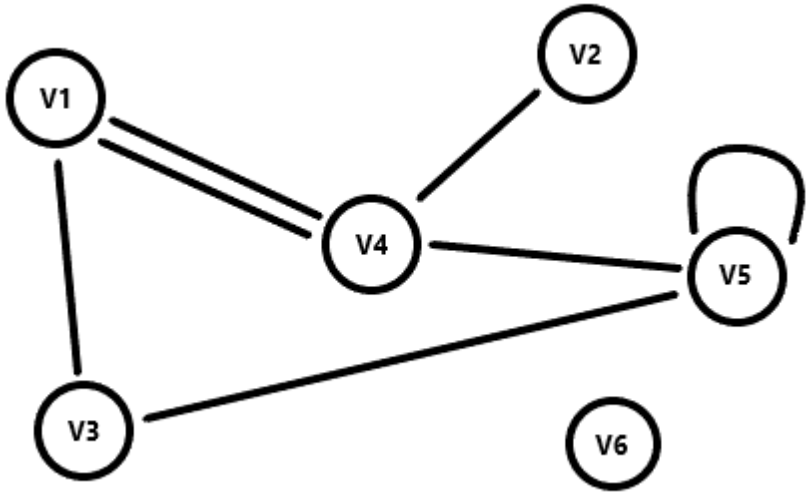


Ilustración 3. Grafo con diferentes tipos de aristas. Fuente: elaboración propia

De estas definiciones se pueden extraer las de los siguientes conceptos. Dos nodos son adyacentes si, al igual que las aristas adyacentes están relacionadas por un nodo, están relacionados por una arista. Cuando una arista une dos nodos, se denomina incidencia y, cuando diversos nodos están conectados en sucesión de aristas, se llama camino y el número de estas participantes en él determina su longitud. Cabe destacar que, cuando un camino supera una longitud de 3 aristas, empezando en un nodo y acabando en el mismo sin repetir nodos ni aristas, se le denomina ciclo. El número de aristas incidentes en un nodo definen su grado y, según esta cualidad, un nodo puede ser:

- **Nodo adyacente (Ilustración 3: V_1, V_2, V_3, V_4, V_5)**: como ya se ha dicho, están unidos a otros nodos mediante aristas. Su grado puede variar de 1 a más.

- **Nodo terminal o pendiente (Ilustración 3: V_2):** se trata de un nodo adyacente con solo una arista o relación con otro nodo dentro de un grafo, es decir, de grado 1.

- **Nodo aislado (Ilustración 3: V_6)**: se trata de un nodo sin ningún tipo de relación con otros nodos, es decir, de grado 0.

En cuanto a los tipos de grafos, ya se ha visto que la definición más básica de un grafo (Grafo simple) implica solo una relación entre dos nodos, pero se pueden encontrar también las variaciones los siguientes:

- **Multigrafo:** por el contrario que el grafo simple, este tipo de grafo acepta más de una arista entre dos nodos.
- **No dirigido (Ilustración 3):** Las aristas consisten en relaciones simétricas, sin dirección, por lo que sus dos vértices actúan a la vez como inicio y final de esta.
- **Grafo dirigido o digrafo (Ilustración 4):** Las aristas entre nodos tienen una única dirección (unidireccional) u orden asignada según los datos, es decir, van de un vértice a otro, que actúan como inicio y final de esta.

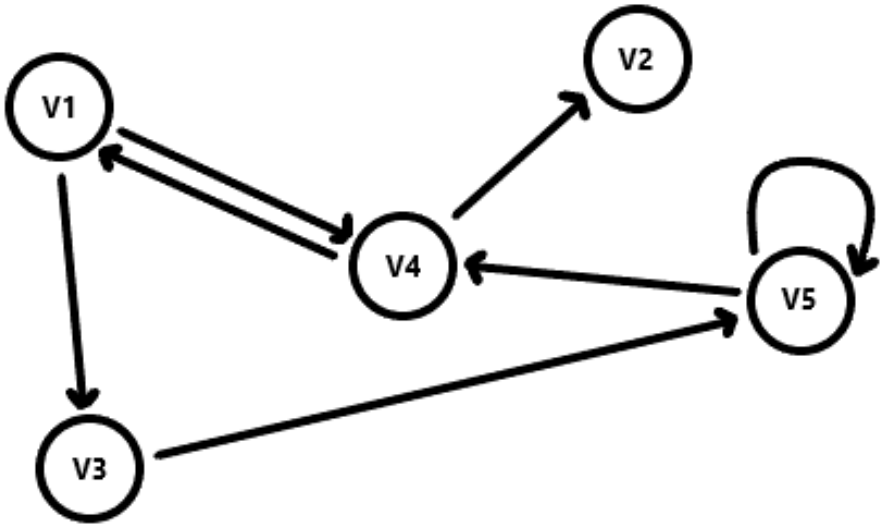


Ilustración 4. Grafo dirigido. Fuente: elaboración propia

En este tipo de grafos, las aristas reciben también el nombre de arcos y, gráficamente, se define la dirección con una flecha, por lo que, según el nodo que se observe, una arista será entrante o saliente.

- **Grafo etiquetado, ponderado o dirigido con pesos:** en este tipo de grafo los nodos y aristas contienen información adicional, es decir, los nodos tienen una etiqueta o las aristas tienen un peso (número entero).
- **Grafo bipartito:** en este tipo de grafo, los nodos pueden estar separados en conjuntos diferentes, pero estar unos con nodos de los otros conjuntos.
- **Grafo conexo:** en él, cada par de nodos, sea cual sea su situación en el grafo, se encuentra conectado mínimo por un camino posible.

- **Grafo completo:** es como el anterior, pero todos los nodos se encuentran unidos por aristas de todas las formas posibles.

- **Grafo vacío:** grafo sin aristas.

Dentro de cualquier tipo de grafo se encuentran los subgrafos, que son partes de un mismo grafo (subconjuntos) que pueden formar asimismo grafos dentro de este, teniendo características similares pero de dimensiones más reducidas.

Existen más tipologías de grafos pero, para la investigación presente, no se estima profundizar más en ellas y se procede con las estructuras de datos para la representación de grafos. Este concepto se refiere a las estructuras (textos, tablas, listas, matrices, etc.) que contienen los datos que los grafos representan. Dichas estructuras implican tanto los archivos que se podrían extraer de grafos creados como los archivos que se deberían importar para visualizar grafos, como es el caso del trabajo presente. La división más básica y extendida de estas estructuras se encuentra entre listas o matrices y, dentro de ellas, se encuentran las siguientes:

- **Estructuras de datos en listas:**

  - Lista de incidencia o de aristas: se listan las aristas dentro de un vector, en el que se contienen los extremos de cada una de ellas. Como ejemplo, el vector utilizado anteriormente para describir el grafo de la **Ilustración 2**: E= {{V_1, V_3}, {V_1, V_4}, {V_2, V_4}, {V_3, V_5}, {V_4, V_5}}.

  - Lista de adyacencia: se listan los nodos de forma que cada uno de ellos contiene los vértices con los que tiene adyacencia (coloquialmente, vecinos). Como ejemplo, el vector correspondiente a la **Ilustración 2**: V= {{V_3, V_4}, {V_4}, {V_1, V_5}, {V_1, V_2, V_5}, {V_3, V_4}}.

  - Lista de grados: se listan los grados de los nodos que componen el grafo. Como ejemplo, el vector correspondiente a la **Ilustración 2**: V= {2, 1, 2, 3, 2}.

- **Estructuras de datos en matrices:**

  - Matriz de adyacencia: consiste en una matriz binaria conformada por los nodos (V) del grafo en la primera fila por los mismos nodos (V) del grafo en la primera columna, indicando 1 si existe una arista y 0

si esta no existe. Como ejemplo, a continuación (**Ilustración** 5) se observa la matriz de adyacencia del grafo de la **Ilustración 2**:

	V_1	V_2	V_3	V_4	V_5
V_1	0	0	1	1	0
V_2	0	0	0	1	0
V_3	1	0	0	0	1
V_4	1	1	0	0	1
V_5	0	0	1	1	0

Ilustración 5. Matriz de adyacencia. Fuente: elaboración propia

- Matriz de incidencia: consiste en una matriz binaria conformada por las aristas (*E*) en la primera fila por los nodos (*V*) del grafo en la primera columna, indicando 1 si el nodo se encuentra en la arista y 0 si no se encuentra. Como ejemplo, a continuación (**Ilustración 6**) se observa la matriz de incidencia del grafo de la **Ilustración 2**:

	$\{V_1, V_3\}$	$\{V_1, V_4\}$	$\{V_2, V_4\}$	$\{V_3, V_5\}$	$\{V_4, V_5\}$
V_1	1	1	0	0	0
V_2	0	0	1	0	0
V_3	1	0	0	1	0
V_4	0	1	1	0	1
V_5	0	0	0	1	1

Ilustración 6. Matriz de incidencia. Fuente: elaboración propia

Una vez explicados conocimientos básicos de las disciplinas que implican a la presente investigación a partir del estado del arte, se procede a pasar a explicar la metodología seguida.

4. Metodología

4.1. Datos

El conjunto de expresiones de búsqueda que se va a utilizar consiste originalmente en un fichero de tipo Hoja de cálculo de Microsoft Office Excel, de extensión ".xlsx". Dicho fichero fue extraído por **Jorge Serrano-Cobos** (2019) en el transcurso de su Tesis doctoral *Hábitos de recuperación de información en motores de búsqueda sobre lectura, libro y bibliotecas en España (2004-2016)* sobre la validez de los procesos, y expresiones de búsqueda, llevados a cabo por usuarios de motores de búsqueda, como indicadores para la investigación de hábitos de búsqueda de lectura en Google entre los años 2004 y 2016.

La génesis de la lista de expresiones, siguiendo con la tesis citada en el párrafo anterior, se llevó a cabo utilizando la herramienta gratuita *Google Keyword Planner*, como su nombre indica, de *Google*. Esta herramienta sirve para realizar análisis de palabras clave para adaptar páginas web al comportamiento de los usuarios que se refleja en los resultados que aporta el servicio para así mejorar el posicionamiento SEO de las páginas (**Olivier Peralta**, 2020).

El conjunto en sí, está formado por un total de 30.865 expresiones de búsqueda, cada una con su ID respectiva, estructuradas en 2 columnas, la A para las IDs y la B para las expresiones completas. En la **Ilustración 7** se pueden observar, como ejemplo, las primeras 33 expresiones del fichero con sus IDs correspondientes. Cabe destacar el hecho de que la totalidad de las expresiones de búsqueda se halla escrita en letra minúscula y sin acentos, pudiendo contener faltas de ortografía que, en este trabajo, se reproducen con total fidelidad.

	A	B
1	Idkeyword	Keyword
2	2	1964 borges
3	3	1964 borges analisis
4	4	1964 jorge luis borges
5	5	3 mujeres escritoras
6	6	6 personajes en busca de un autor
7	7	a christmas carol de charles dickens
8	8	a la busca del tiempo perdido
9	9	a la busqueda del tiempo perdido
10	10	a los amigos borges
11	11	a mi manera autor
12	12	a puerta cerrada sartre
13	13	a puerta cerrada sartre descargar
14	14	a puerta cerrada sartre pdf
15	15	a puertas cerradas sartre
16	16	a que movimiento literario pertenece gabriel garcia marquez
17	17	a un gato borges
18	18	a un gato jorge luis borges
19	19	abel hernandez escritor
20	20	abigail borges
21	21	abogados derechos de autor
22	24	absalon absalon faulkner
23	28	adios a las armas ernest hemingway
24	29	adios a las armas hemingway
25	31	adriana ortemberg
26	32	aforismo nietzsche
27	33	aforismos nietzsche
28	34	afterglow borges
29	35	agatha christi
30	36	agatha christie
31	37	agustina roca escritora
32	38	ajedrez borges
33	39	ajedrez borges analisis

Hoja1

Ilustración 7. Primeras 33 expresiones de búsqueda del conjunto de datos original. Fuente: elaboración propia

El conjunto completo de datos se puede consultar y descargar en el siguiente enlace.

4.2. Hardware

En total, se han utilizado tres ordenadores, dadas las distintas complicaciones surgidas según qué programas y limitaciones. Estos dispositivos se describen a continuación.

El dispositivo propio y principalmente utilizado (en adelante, **Dispositivo 1**), es un ordenador portátil HP que se caracteriza por las siguientes especificaciones:

- **Sistema operativo:** Windows 10 Home 64 bits.
- **Procesador:** Intel® Core™ i5-1035G1 CPU @ 1.00GHz (8CPUs), ~1.2GHz.
- **Memoria RAM:** 8.192MB.
- **Versión de DirectX:** DirectX 12.
- **Pantalla:**
    - **Nombre:** Intel® UHD Graphics.
    - **Memoria total aprox.:** 4.100 MB.
    - **Memoria VRAM:** 128 MB.
    - **Modo de presentación:** 1.366 x 768 (32 bit) (60Hz).

La siguiente computadora, se trata de un ordenador de mesa (en adelante, **Dispositivo 2**), el más potente de los tres dispositivos, con las siguientes propiedades:

- **Sistema operativo:** Windows 10 Home 64 bits.
- **Procesador:** Intel® Core™ i7-6700.
- **Memoria RAM:** 16.383MB.
- **Versión de DirectX:** DirectX 12.
- **Pantalla:**
    - **Nombre:** Sapphire Radeon Rx480 8GB.
    - **Memoria total aprox.:** 16.383MB.
    - **Memoria VRAM:** 8.192MB.
    - **Modo de presentación:** Dato no disponible.

Finalmente, el último dispositivo a describir de los utilizados, un MacBook Air (en adelante, **Dispositivo 3**) con las siguientes especificaciones:

- **Sistema operativo**: macOS Catalina version 10.15.6.
- **Pocesador:** 1.1 GHz Intel® Core™ i3 de doble núcleo.
- **Memoria RAM:** 8.192MB 3733 MHz LPDDR4X.
- **Pantalla:**
    - **Nombre:** Intel Iris Plus Graphics**.**
    - **Memoria total aprox.:** 1.536 MB.
    - **Memoria VRAM:** Dato no disponible.
    - **Modo de presentación:** 13.3 pulgadas (2560 x 16000).

4.3. Software

En este apartado se describe el software recopilado, testado y, finalmente, utilizado para llevar a cabo el proceso de tratamiento de los datos iniciales (ver **4.1. Datos**) y la posterior visualización de estos en forma de grafos.

Para la observación del fichero original de expresiones de búsqueda, en formato ".xlsx" (ver **4.1. Datos**), se utilizó el programa básico Microsoft Office Excel 2007.

Después de elaborar un plan de tratamiento de los datos para la creación de un nuevo archivo en alguno de los diferentes formatos de archivo para grafos, se utilizó, para la realización de este proceso, el software RStudio (**RStudio Team**, 2020) para el lenguaje de programación R (**R Core Team**, 2017).

Seguidamente, para la generación de los grafos, se instalaron en el **Dispositivo 1** (ver **4.2. Hardware**) los siguientes software: Pajek (Pajek, Pajek XXL y Pajek 3XL) (**Batagelj y Mrvar**, 1996), SocNetV (**Kalamaras**, 2015), Gephi (**Bastian, Heymann y Jacomy**, 2009), Neo4j Desktop (**Neo4j Engineering**, 2010) y, por último, UCINET

(**Borgatti, Everett y Freeman**, 2002). A continuación se muestra una pequeña descripción de cada uno de estos cinco:

- **Pajek v5.09 (Pajek, Pajek XXL y Pajek 3XL):** Software libre de uso no comercial especializado en el análisis, de forma cómoda, de grandes redes sociales para cuya finalidad dispone diversos algoritmos, por ejemplo, la división de redes grandes en varias más pequeñas (**Alonso Moreno, González Hernández y Lavera Ulloa**, 2012). Las versiones del programa mencionadas en el paréntesis aumentan el número de vértices de los que tiene capacidad el programa original. Pajek acepta los siguientes formatos de archivo: Pajek networks (".net"), Pajek matrices (".mat"), Vega graphs (".vgr"), GEDCOM files (".ged"), UCINET DL files (".dat"), Ball and Stick files (".bs"), Mac Molecule files (".mac"), MDL MOL files (".mol"), Pajek partitions (".clu"), Pajek vectors (".vec"), Pajek permutations (".per"), Pajek cluster (".cls") y Pajek hierarchy (".hie").

- **SocNetV v2.5 (Social Network Visualizer):** Herramienta de análisis y visualización de redes sociales que permite tanto la creación de nuevas redes como la modificación de redes cargadas (**Alonso Moreno, González Hernández y Lavera Ulloa**, 2012), en diversos formatos: GML (".gml"), Pajek (".net", ".paj", ".pajek"), Adjacency (".csv", ".sm", ".adj", ".txt"), Two-Mode Sociomatrix (".2sm", ".aff"), Weighted y Simple Edge List (".txt", ".list", ".edgelist", ".lst", ".wlst"), UCINET (".dl", .dat") y GraphViz (".dot").

- **Gephi v0.9.2:** Software de código abierto desarrollado en Java para la visualización, exploración, manipulación y análisis de gráficos de redes sociales. Los formatos soportados por este programa son los siguientes: Archivos Gephi (".gephi"), UCINET (".dl"), GraphViz (".dot", ".gv"), GDF (".gdf"), GEXF (".gexf"), GML (".gml"), GraphML (".graphml"), Pajek (".net"), TGF (".tgf"), TLP (".tlp"), VNA (".vna"), CSV/Spreadsheet (".csv", ".tsv", ".edges", ".xls", ".xlsx") e incluso, archivos comprimidos (".zip", ".gz", ".bz2").

- **Neo4j Desktop v1.3.4:** Se trata de un software libre de bases de datos NoSQL (Not only SQL) orientado a grafos con gran cantidad de datos y relaciones entre estos. Está capacitado para soportar hasta un total de 34.000.000.000 nodos y el mismo número de aristas.

- **UCINET v6.714:** Al igual que todos los anteriores, se trata de un software de análisis de redes sociales construido específicamente para la velocidad de

trabajo, funcionando con una estructura basada en menús y la elección de algoritmos (**Alonso Moreno, González Hernández y Lavera Ulloa**, 2012). UCINET admite los siguientes formatos: Excel (".xls", tanto DL como matriz) y archivos de texto (".txt", DL, VNA, Pajek, Krackplot, Negopy, etc.) (**Borgatti, Everett y Freeman**, 2002).

4.3.1. Justificación de la elección

La elección del lenguaje de programación R se debe a su gran capacidad para la manipulación de datos masivos de forma sencilla, por lo que es bastante adecuado para el campo de los análisis estadísticos. Otra característica a destacar es su adaptabilidad a casi cualquier tipo de formato de archivo que se pretenda importar, por lo que facilita el trabajo evitando procesos de transformación previos de archivos.

Seguidamente, la elección del entorno RStudio frente a la consola de trabajo de R se resulta de la comodidad de trabajo que ofrece su diseño principalmente frente a la interfaz rudimentaria de R. Dadas las características de RStudio como entorno de desarrollo integrado (IDE), es decir, su interfaz separada en módulos totalmente distinguibles y manejables, que permiten un uso más intuitivo y guiado para usuarios no expertos de dicho lenguaje de programación, como es el caso de esta investigación. Dicho entorno facilita también bastante el proceso de trabajo gracias a la posibilidad de visualizar los datos modificados o creados de una forma rápida con cada cambio que se realiza.

En cuanto a los programas de análisis de grafos, siguiendo el orden en el que han sido citados en el apartado en el apartado Software (ver **4.3. Software**), a continuación se procede a la justificación de su elección según criterios de adaptabilidad a formatos, interfaz gráfica y dificultad de uso y justificando, primero, los descartes, y segundo, la elección.

Pajek (v5.09), se descartó dada la poca variedad de formatos de archivos sobre los que trabajar. Dichos formatos ofrecidos resultaban demasiado propios de Pajek, resultando demasiado estricto, y complicaban o limitaban el proceso de trabajo, también influido por la dificultad de uso del programa, dada su interfaz también un tanto tosca (**Ilustración 8**).

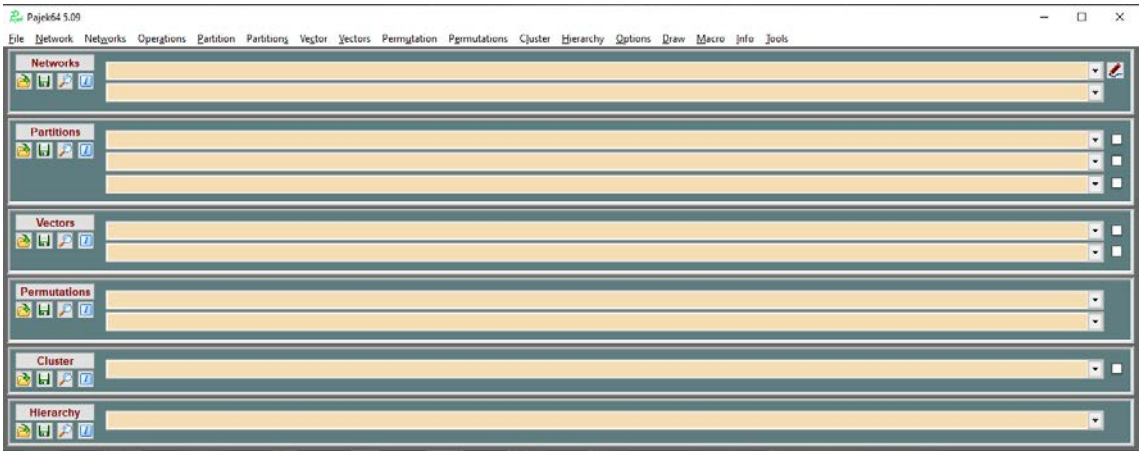


Ilustración 8. Interfaz de Pajek. Fuente: elaboración propia

SocNetV (v2.5), pese a tener, junto a Gephi, una de las interfaces más agradables, cómodas e intuitivas y admitir una gran variedad de formatos, fue prácticamente imposible de utilizar ya que, con cada intento de importar matrices o archivos, saltaban errores de compatibilidad (**Ilustración 9**) aunque los archivos estuviesen correctamente adaptados.

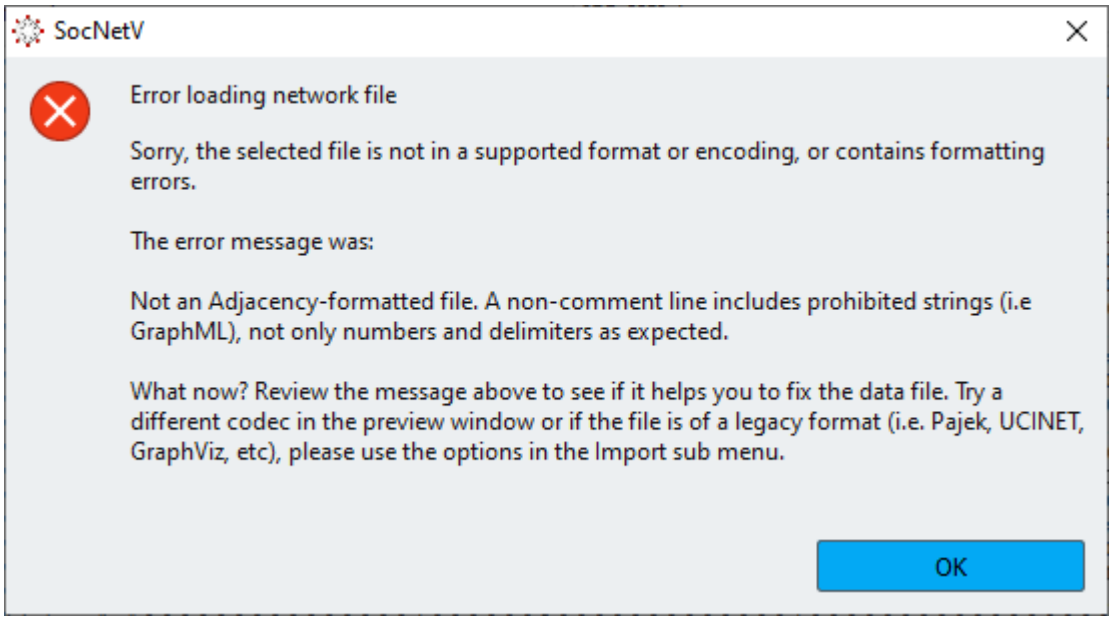


Ilustración 9. Error de compatibilidad en SocNetV. Fuente: elaboración propia

Neo4j Desktop (v1.3.4), siendo una buena alternativa gráfica y atractiva frente a todos los programas mencionados, pierde con el hecho de exigir, al ser una base de datos orientada a grafos (BDOG) NoSQL (Not Only SQL), una preparación necesaria

del usuario previa a su uso, aumentando así la dificultad para trabajar con este programa. Otro inconveniente que propició el descarte de Neo4j Desktop fueron sus requisitos de sistema, más altos en comparación con el resto de programas testados.

Finalmente, en cuanto a **UCINET (v6.714)** el programa en se encuentra un tanto limitado a la visualización gráfica de datos ya que, en cuanto la red sobrepasa cierto número de nodos, la capacidad de procesamiento de UCINET se ve ralentizada en exceso. Este hecho, sumado a la lectura de la desventaja siguiente señalada por **Miceli** (2008): "La cantidad de columnas de la matriz de adyacencia transformada en matriz de incidencia será igual a la cantidad total de lazos de la matriz. Para redes de alta densidad que tengan esta estructura de vínculos entre nodos del mismo tipo, la interpretación de datos puede ser torturantemente difícil.", propició el descarte de UCINET.

En cuanto al programa elegido definitivamente para la investigación, **Gephi (v0.9.2)**, fue prácticamente el único programa que no dio problemas ni de compatibilidad de formatos, ni de capacidad, desde un principio. Ya con el primer conjunto de datos que se le importó (ver **4.3.2.2.1. Script 1**), Gephi fue capaz de generar con bastante rapidez un grafo, incompleto por falta de memoria VRAM dadas las elevadas dimensiones de dicho conjunto inicial, totalmente manejable. Ésta característica permite decidir si descartar o seguir trabajando con los datos respectivos de una manera bastante rápida y así ver que se puede modificar o no de estos para mejorar la visualización. Aunque su manejo en principio puede resultar un tanto confuso, se han encontrado artículos y guías bastante claras que han ayudado en este proceso, como ha sido, por ejemplo, el artículo de **Grandjean** (2015), al contrario del resto de programas, cuya bibliografía encontrada resultaba un tanto confusa.

4.3.2. R & RStudio

Por una parte, R hace referencia tanto al lenguaje de programación interpretado como a su propia consola o entorno de trabajo. Lenguaje de programación interpretado significa que interpreta y lleva a cabo las órdenes recibidas de forma directa en forma de comandos en la mencionada anteriormente consola (**Ilustración 10**).

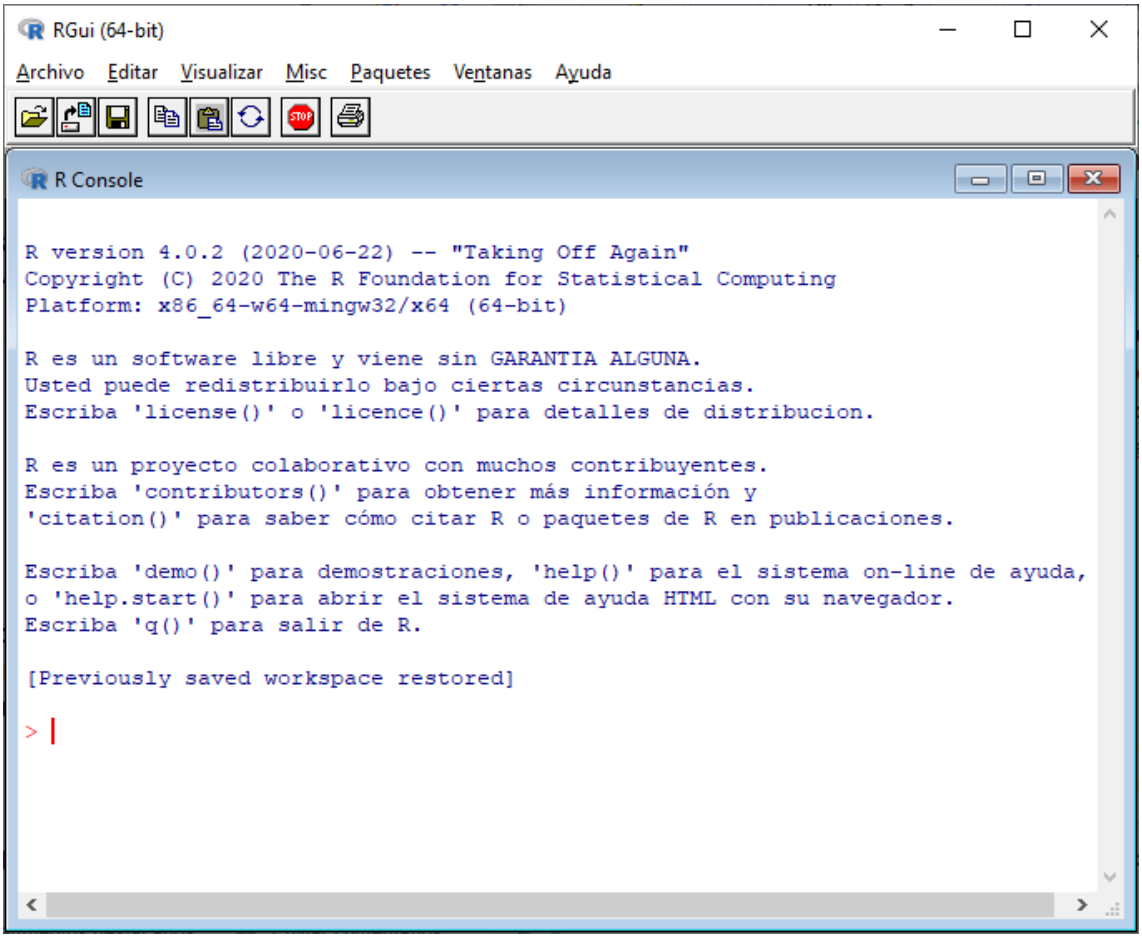


Ilustración 10. Entorno de trabajo de R. Fuente: elaboración propia

R está dedicado al análisis estadístico y gráfico de datos y permite, aparte de la visualización de cualquier archivo cargado, la modificación de estos e incluso, la generación de información nueva a partir de ellos. Esta capacidad de modificación y generación es la que se ha utilizado en la presente investigación, como se muestra más adelante. Otra utilidad bastante común de R es la de herramienta de cálculo matemático, ya que posee una gran potencia para llevar a cabo operaciones y de forma sencilla.

Por otra parte, RStudio es un entorno de desarrollo integrado o *Integrated Drive Electronics* (IDE) que, básicamente, sirve como una interfaz de trabajo más confortable, intuitiva y sencilla que la consola descrita anteriormente para utilizar el lenguaje de programación R.

En la **Ilustración 11**, se puede observar una captura de la interfaz de RStudio, cuya estructura se encuentra dividida en cuatro paneles principales, siendo estos (**Boccardo Bosoni y Ruiz Bruzzone**, 2019):

1. ***R Script*:** Este panel sirve para escribir, sin ejecución, sólo escritura, las distintas órdenes que más adelante llevará a cabo el programa en la siguiente ventana. Se puede tener más de un *Script* abierto al mismo tiempo.

2. ***Console / Terminal / Jobs*:** En ella se ejecutan las órdenes que se han escrito previamente en el panel anterior (*R Script*) o las que se hayan escrito directamente en ella misma, llevando a la directa ejecución de estas con solo pulsar una tecla. Básicamente, es el mismo entorno de trabajo de R (**Ilustración 10**).

3. ***Environment / History / Connections / Tutorial*:** Aquí se almacenan los datos, matrices y variables, entre otros, que se van generando a medida que se ejecutan las órdenes en la consola.

4. ***Files / Plots / Packages / Help / Viewer*:** Por último, en este panel, respectivamente, se puede:

    a. ***Files*:** Seleccionar el directorio de trabajo en el dispositivo y ver los archivos a utilizar o generados.

    b. ***Plots*:** Visualizar gráficos.

    c. ***Packages*:** Descargar, instalar y seleccionar las librerías que se necesiten para trabajar.

    d. ***Help*:** Acceder a información sobre cualquier aspecto de R (CRAN).

    e. ***Viewer*:** Visualizar informes que se generen.

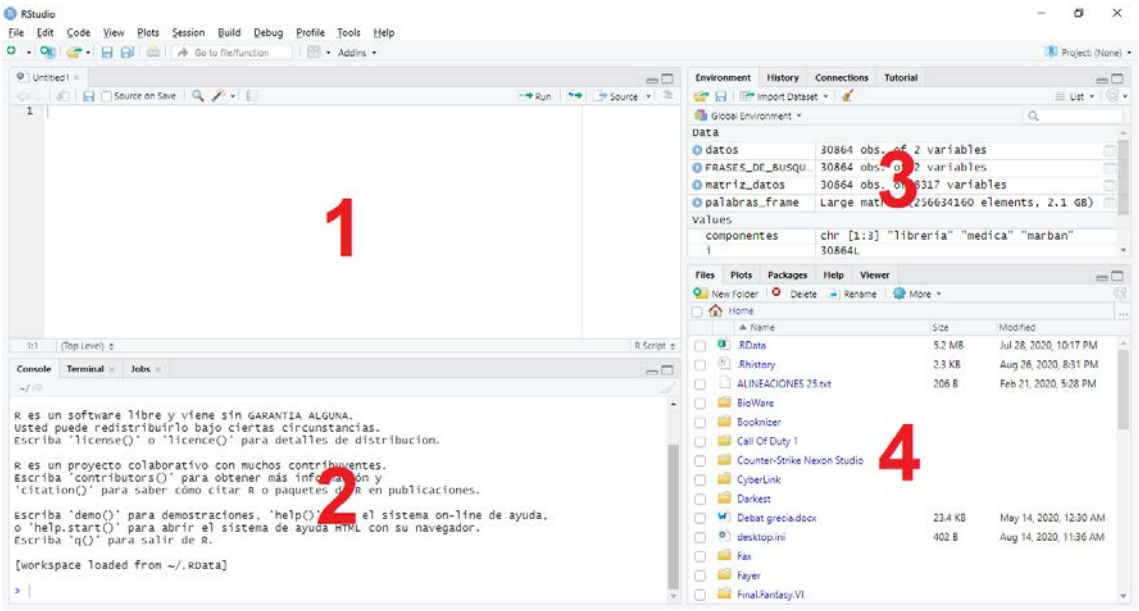


Ilustración 11. Interfaz de RStudio con paneles numerados. Fuente: elaboración propia

4.3.2.1. Requisitos mínimos específicos

RStudio, en Windows, requiere una versión de Sistema Operativo Windows 9x/ME/NT4.0/2000/XP/2003/Vista/7/8/2012 Server/8.1 compatible con Intel. En cuanto a memoria RAM, exige 32MB (**R-Tools Technology Inc.**, 2000-2020).

4.3.2.2. Scripts utilizados en el proceso de trabajo

Los procesos de modificación de los datos con RStudio, debido a las dimensiones de los datos, se realizaron, como se ha explicado en el apartado Hardware, en hasta tres dispositivos distintos. A continuación, se dividen, en cuatro subapartados, los diferentes *scripts* de código que se crearon con la correspondiente explicación de cada uno.

4.3.2.2.1. Script 1: Del conjunto de datos original a la matriz

A partir del examen del conjunto de datos, se precisó la estructura de datos que debería servir para la representación en grafo. Debido a que entre los objetivos específicos del trabajo se encontraban ver la tipología de las expresiones de búsqueda y en que temáticas se podían localizar conjuntos o comunidades, se decidió generar a partir de la lista original una matriz que, teniendo las expresiones de búsqueda en los ejes de forma simétrica, contase las palabras coincidentes entre ellas.

Para esto, en primer lugar, se creó la estructura de la matriz. Para ello, se utilizó el número de expresiones de búsqueda como número de columnas y filas de la matriz. Después se utilizaron los términos de la lista como nombre de las filas y columnas, como se muestra en el código fuente siguiente (**Ilustración 12**):

```
datos <- FRASES_DE_BUSQUEDA_TFM_copia

Matriz_Datos_3<-matrix(ncol = length(datos$Idkeyword), nrow =
length(datos$Idkeyword))
rownames(Matriz_Datos_3)<-datos$Keyword
colnames(Matriz_Datos_3)<-datos$Keyword
```

Ilustración 12. Script 1, parte 1. Fuente: elaboración propia

Seguidamente, se creó un bucle (**Ilustración 13**) para completar los campos de la matriz. En este bucle se compara cada elemento del nombre de las columnas con cada elemento del nombre de las filas. Al realizar dicha comparación y, como se quería, se cuenta cuántas palabras coinciden en cada caso.

No obstante, antes de realizar la comparación fue necesario dividir cada elemento de los nombres de las filas y columnas en las palabras que lo componían. Esto es así porque en R se reconocen los nombres de las filas y columnas como un único elemento, independientemente de que esté conformado por distintas palabras. Así pues, dentro del bucle se crearon dos vectores: uno para el nombre de las columnas y otro para el de las filas. En estos vectores se utilizó la función *strsplit* para evaluar los nombres y separarlos por los espacios, obteniendo así un vector con las palabras que conformaban cada uno. Finalmente, se comparó las palabras de cada nombre de las columnas y filas y se registró el número de palabras que coincidían en cada caso.

```
for (i in 1:dim(Matriz_Datos_3)[1]){
  componentes_fila <- strsplit(rownames(Matriz_Datos_3)[i], ' ')[[1]]
  long_fila <- length(componentes_fila)
  for (j in 1:dim(Matriz_Datos_3)[2]){
    componentes_col <- strsplit(colnames(Matriz_Datos_3)[j], '
')[[1]]
    long_col <-length(componentes_col)
    igual <-0
    for (k in 1:long_fila){
      for (l in 1:long_col){

        if (componentes_fila[k]==componentes_col[l]){
          igual<-igual+1
        }
      }
    }
    Matriz_Datos_3[i,j]<-igual
  }
}
```

Ilustración 13. Script 1, parte 2. Fuente: elaboración propia

La matriz obtenida se exportó (**Ilustración 14**) a formato CSV y, el archivo resultante, es la matriz original de la cual han surgido el resto de matrices con las que se ha trabajado a lo largo del estudio.

```
write.csv2(Matriz_Datos_3, file = 'Matriz_Datos_3.csv',
           row.names = TRUE,
           fileEncoding = 'UTF-8')
```

Ilustración 14. Script 1, parte 3. Fuente: elaboración propia

Este proceso resultó ser el más complejo debido al tamaño del conjunto de datos originales y de la matriz generada (*Matriz_datos_3*), ya que esta última estaba conformada (**Ilustración 15**) por 30.864 expresiones multiplicadas por ellas mismas, generando así 952.586.496 celdas y tenía un peso de 7.6 GB. En total, requirió un coste temporal de entre 5 y 6 horas de procesamiento y se llevó a cabo en el **Dispositivo 2** (ver **4.2. Hardware**).

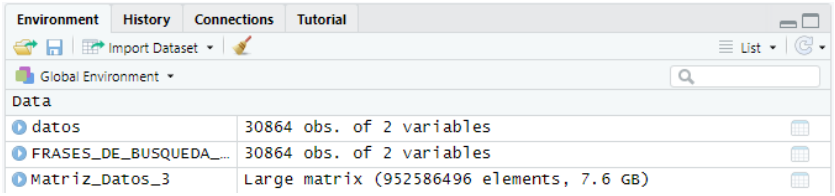


Ilustración 15. Pestaña Environment de Rstudio resultante del Script 1. Fuente: elaboración propia

El archivo CSV extraído llegó a pesar 1.77 GB, por lo que fue totalmente imposible tanto abrir el archivo en cualquiera de los tres dispositivos utilizados, como importarlo en Gephi por problemas de memoria dedicada del programa, por lo que se llevó a cabo la extracción de muestras aleatorias de la matriz.

4.3.2.2.2. Script 2: Extracción de muestras aleatorias de la matriz

Dados los problemas de potencia insuficiente para trabajar con la matriz completa mencionados en el anterior subapartado, se decidió crear nuevas matrices que contuvieran un porcentaje menor de observaciones. De esta forma se crearon matrices aleatorias con un 1%, 2%, 3%, 5%, 10%, 15% y 20% de observaciones de la matriz completa. Esta variedad de porcentajes se eligió para, una vez generados los CSV correspondientes a cada porcentaje, ir probándolos de menor a mayor porcentaje en Gephi hasta llegar a su límite de potencia en el **Dispositivo 1** (ver **4.2. Hardware)**.

El procedimiento empleado (**Ilustración 16**) fue el mismo para todas las matrices. Primero, se calculó el número de observaciones que correspondían a cada porcentaje. Seguidamente, se creó un vector que contenía una muestra aleatoria de la posición de cada búsqueda del listado, utilizando el número calculado para determinar la longitud del vector. Este vector se utilizó para seleccionar las correspondientes filas y columnas de la matriz completa.

```
colnames(Matriz_Datos_3)<-FRASES_DE_BU_SQUEDA_TFM_copia$Keyword
rownames(Matriz_Datos_3)<-FRASES_DE_BU_SQUEDA_TFM_copia$Keyword

pos_1 <- sample(1:dim(Matriz_Datos_3)[1], 309)
Matriz_1_Porc <-Matriz_Datos_3[pos_1,pos_1]

pos_2 <- sample(1:dim(Matriz_Datos_3)[1], 618)
Matriz_2_Porc <-Matriz_Datos_3[pos_2,pos_2]

pos_3 <- sample(1:dim(Matriz_Datos_3)[1], 926)
Matriz_3_Porc <-Matriz_Datos_3[pos_3,pos_3]

pos_5 <- sample(1:dim(Matriz_Datos_3)[1], 1544)
Matriz_5_Porc <-Matriz_Datos_3[pos_5,pos_5]

pos_10 <- sample(1:dim(Matriz_Datos_3)[1], 3087)
Matriz_10_Porc <-Matriz_Datos_3[pos_10,pos_10]

pos_15 <- sample(1:dim(Matriz_Datos_3)[1], 4630)
Matriz_15_Porc <-Matriz_Datos_3[pos_15,pos_15]

pos_20 <- sample(1:dim(Matriz_Datos_3)[1], 6173)
Matriz_20_Porc <-Matriz_Datos_3[pos_20,pos_20]
```

Ilustración 16. Script 2, parte 1. Fuente: elaboración propia

Finalmente (**Ilustración 17**), se exportaron las nuevas muestras obtenidas en formato CSV, destacando el cambio de formato de codificación de caracteres de los archivos de UTF-8 a ISO 8859-1, la codificación correspondiente al alfabeto latino.

```
write.csv2(Matriz_1_Porc, file = 'Matriz_1_porc.csv',
           fileEncoding = 'ISO-8859-1',
           row.names = TRUE)

write.csv2(Matriz_2_Porc, file = 'Matriz_2_porc.csv',
           fileEncoding = 'ISO-8859-1',
           row.names = TRUE)

write.csv2(Matriz_3_Porc, file = 'Matriz_3_porc.csv',
           fileEncoding = 'ISO-8859-1',
           row.names = TRUE)


write.csv2(Matriz_5_Porc, file = 'Matriz_5_porc.csv',
           fileEncoding = 'ISO-8859-1',
           row.names = TRUE)

write.csv2(Matriz_10_Porc, file = 'Matriz_10_porc.csv',
           fileEncoding = 'ISO-8859-1',
           row.names = TRUE)

write.csv2(Matriz_15_Porc, file = 'Matriz_15_porc.csv',
           fileEncoding = 'ISO-8859-1',
           row.names = TRUE)

write.csv2(Matriz_20_Porc, file = 'Matriz_20_porc.csv',
           fileEncoding = 'ISO-8859-1',
           row.names = TRUE)
```

Ilustración 17. Script 2, parte 2. Fuente: elaboración propia

Este último cambio se debió al hecho de que, con el formato de codificación UTF-8, RStudio no detectaba letras propias del alfabeto latino como la “ñ”, dando problemas en cuanto al reconocimiento de dicho carácter.

Todo este proceso se llevó a cabo en el **Dispositivo 3** (ver **4.2. Hardware**), suponiendo un coste de tiempo casi nulo comparado con el paso del subapartado anterior. En el apartado de Resultados (ver **5. Resultados**), se muestra el porcentaje elegido finalmente y el primer grafo resultante. Cabe destacar que los CSV generados en este proceso ocuparon de 202 KB (1%) a 72.9 MB (20%) y se pueden obtener en este enlace.

4.3.2.2.3. Script 3: Eliminación de la diagonal de la matriz

Habiendo cargado la matriz con el 5% de las observaciones originales, generada en el subapartado anterior, y dados los motivos que se indican y explican en el

subapartado Grafo 1 del apartado Resultados (ver **5.1. Grafo 1**), se procedió a eliminar los valores de la diagonal de la matriz. Al ser una matriz simétrica, dichos valores representaban la coincidencia del número de palabras de cada expresión de búsqueda de la lista frente a ella misma.

Para ello, se creó un bucle (**Ilustración 18**) en el que se indicó que, cuando el número de columna y el número de filas coincidiesen, se introdujese NA (No disponible, del inglés *Not Available*) en la casilla correspondiente de la matriz.

```
Matriz_5_porc<-as.data.frame(Matriz_5_porc)
rownames(Matriz_5_porc)<-Matriz_5_porc[,1]
Matriz_5_porc<-Matriz_5_porc[,-1]

Matriz_5_porc_sin_diag <- Matriz_5_porc

for (i in 1:dim(Matriz_5_porc_sin_diag)[1]){
  for (j in 1:dim(Matriz_5_porc_sin_diag)[2]){
    if (i==j){
      Matriz_5_porc_sin_diag[i,j]<-NA
    }
  }
}

write.csv2(Matriz_5_porc_sin_diag, file =
'Matriz_5_porc_sin_diag.csv',
           row.names = TRUE,
           fileEncoding = 'ISO-8859-1')
```

Ilustración 18. Script 3. Fuente: elaboración propia

Este proceso se llevó a cabo en el **Dispositivo 1** sin coste alguno de tiempo, generando una matriz de las mismas dimensiones, pero sin la diagonal (**Ilustración 19**).

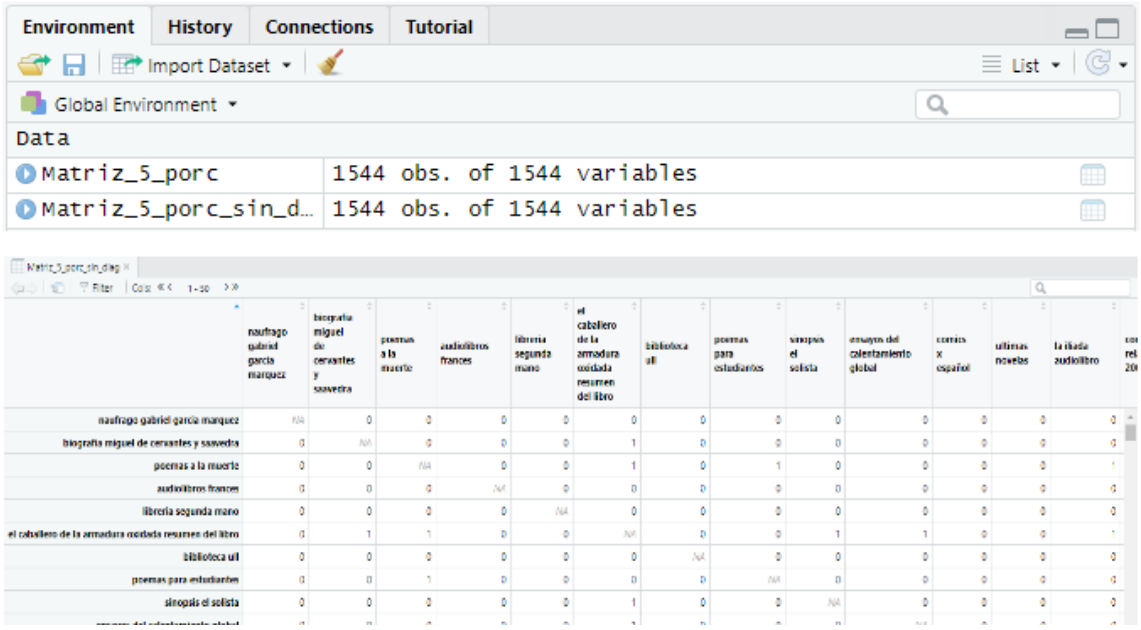


Ilustración 19. Pestaña Environment resultante del Script 3 y visualización de la matriz. Fuente: elaboración propia

El CSV extraído en este proceso ocupó un tamaño de 4.62 MB y se puede descargar en el siguiente enlace.

4.3.2.2.4. Script 4: Eliminación de ruido de la matriz

Al igual que con el subapartado anterior, por los motivos extraídos de los resultados del subapartado anterior y explicados en el subapartado Grafo 3 de Resultados (ver **4.3. Grafo 3**), se procedió a eliminar las palabras clave conectoras que más ruido generaban en el grafo. Para ello, se procedió como se explica a continuación.

En primer lugar, se creó un vector (**Ilustración 20**) con todos los conectores a eliminar con un espacio antes y después de cada uno de ellos. La inclusión de este espacio se debe a que, sin él, simplemente se eliminarían todas sus letras de las expresiones de búsqueda y no solo los conectores requeridos, es decir, que los espacios permiten asegurarse de que se está seleccionando el conector.

```
datos <- Matriz_5_porc

Matriz_5_porc_sin_conectores <- datos

conectores <- c(" y ",
                " el ",
                " la ",
                " los ",
                " las ",
                " por ",
                " para ",
                " a ",
                " de ",
                " del ",
                " en ")
```

Ilustración 20. Script 4, parte 1. Fuente: elaboración propia

En segundo lugar, se añadió un espacio al principio y al final de cada expresión de búsqueda para después poder englobar para la eliminación también todos los conectores que se encontraran al inicio o al final de la expresión búsqueda (**Ilustración 21**).

```
for (i in 1:dim(Matriz_5_porc_sin_conectores)[1]){
  Matriz_5_porc_sin_conectores$X1[i] <- paste('
',Matriz_5_porc_sin_conectores$X1[i],' ')
}
```

Ilustración 21. Script 4, parte 2. Fuente: elaboración propia

Seguidamente, se eliminaron las palabras clave (**Ilustración 22**) contenidas en el vector anterior de cada expresión de búsqueda de la matriz mediante la función *gsub* de R, la cual permite seleccionar un determinado patrón de un *storing* y sustituirlo por lo que se desee. En este caso no se aportó ningún patrón de sustitución, lo que dio lugar a la eliminación de todas las palabras clave conectoras detectadas.

```
for (i in 1:dim(Matriz_5_porc_sin_conectores)[1]){
  for (j in 1:length(conectores)){
    Matriz_5_porc_sin_conectores$X1[i] <- gsub(conectores[j],'
',Matriz_5_porc_sin_conectores$X1[i])
  }
}
```

Ilustración 22. Script 4, parte 3. Fuente: elaboración propia

Una vez eliminadas las palabras clave seleccionadas de la matriz, fue necesario volver a calcular (**Ilustración 23**) la coincidencia de palabras dentro de las expresiones de búsqueda en cada caso, puesto que los valores originales de dicha matriz ya no eran válidos. Para ello se siguió el mismo procedimiento que se utilizó para crear la matriz original.

```
busquedas <- unique(Matriz_5_porc_sin_conectores$X1)
Matriz_5_porc_sin_conectores_sin_diagonal <- matrix(nrow =
length(busquedas), ncol = length(busquedas))
colnames(Matriz_5_porc_sin_conectores_sin_diagonal)<-busquedas
rownames(Matriz_5_porc_sin_conectores_sin_diagonal)<-busquedas

for (i in 1:dim(Matriz_5_porc_sin_conectores_sin_diagonal)[1]){
  componentes_fila <-
strsplit(rownames(Matriz_5_porc_sin_conectores_sin_diagonal)[i], '
')[[1]]
  componentes_fila <- componentes_fila[which(componentes_fila!='')]
  long_fila <- length(componentes_fila)
  for (j in 1:dim(Matriz_5_porc_sin_conectores_sin_diagonal)[2]){
    componentes_col <-
strsplit(colnames(Matriz_5_porc_sin_conectores_sin_diagonal)[j], '
')[[1]]
    componentes_col <- componentes_col[which(componentes_col!='')]
    long_col <-length(componentes_col)
    igual <-0
    for (k in 1:long_fila){
      for (l in 1:long_col){

        if (componentes_fila[k]==componentes_col[l]){
          igual<-igual+1
        }
      }
    }
    Matriz_5_porc_sin_conectores_sin_diagonal[i,j]<-igual
  }
}
```

Ilustración 23. Script 4, parte 4. Fuente: elaboración propia

Finalmente, se eliminó la diagonal (**Ilustración 24**) de la matriz tal y como se ha explicado anteriormente (ver **4.3.2.2.3. Script 3**).

```
for (i in 1:dim(Matriz_5_porc_sin_conectores_sin_diagonal)[1]){
  for (j in 1:dim(Matriz_5_porc_sin_conectores_sin_diagonal)[2]){
    if (i==j){
      Matriz_5_porc_sin_conectores_sin_diagonal[i,j]<-NA
    }
  }
}

write.csv2(Matriz_5_porc_sin_conectores_sin_diagonal, file =
'Matriz_5_porc_sin_conectores_sin_diagonal.csv',
           row.names = TRUE,
           fileEncoding = 'ISO-8859-1')
```

Ilustración 24. Script 4, parte 5. Fuente: elaboración propia

El CSV resultante de este proceso ocupa 4.60 MB y se puede descargar en el siguiente enlace.

4.3.3. Gephi

Gephi ha sido la herramienta finalmente utilizada para la visualización y el análisis de grafos, tanto pequeños como de gran tamaño. Como se ha mencionado anteriormente, se trata de un programa de código abierto, es decir, que su código fuente puede ser modificado o actualizado por sus usuarios, y multiplataforma, ya que se encuentra disponible para los sistemas operativos Windows, MacOS X y Linux.

Este software, como sus propios desarrolladores describen, permite “importar, exportar, manipular, analizar, filtrar, representar, detectar comunidades y exportar grandes grafos y redes” (**Bastian, Heymann y Jacomy**, 2009). Gracias a sus visualizaciones y la gran cantidad de opciones aplicables a estas, Gephi es de gran utilidad para encontrar patrones de comportamiento de usuarios o tendencias de búsqueda o temáticas, entre otras, en las bases o conjuntos de datos que se carguen en él.

Sus principales características son las siguientes:

- **Visualización en tiempo real:** Cualquier cambio que se realiza en los distintos parámetros modificables disponibles, se visualiza automáticamente en el grafo.

- **Diseño:** Ofrece una serie de algoritmos de diseño de gráficos, cuya función es distribuir los nodos y aristas del grafo de distintas formas para permitir interpretaciones y lecturas más claras sobre este.

- **Métricas:** Gephi pone a disposición del usuario una gran cantidad de medidas estadísticas sobre los datos cargados que permiten realizar análisis de distintos tipos sobre redes sociales y grafos.

- **Redes a lo largo del tiempo:** Permite la visualización de la evolución de gráficos temporales.

- **Creación de cartografías:** Las etiquetas, colores y tamaño de los nodos y aristas se pueden modificar según criterios seleccionables para una mejor visualización, pudiendo exportar los resultados en los formatos PDF, SVG y PNG.

- **Filtrado dinámico:** Se pueden aplicar filtros a los grafos, para realizar consultas sobre los datos cargados, en tiempo real, como se ha mencionado en el primer punto “Visualización en tiempo real”.

- **Visualización y edición de tablas:** En el módulo “Laboratorio de datos” se pueden observar los datos importados en forma de tabla, así como modificarlos, crear nuevos y eliminarlos.

- **Importación y exportación de archivos:** Permite importar todos los formatos de grafos mencionados anteriormente en la introducción del apartado Software (ver **4.3. Software**).

- **Extensible:** Las funcionalidades de Gephi se pueden ampliar mediante la instalación de *plugins*.

4.3.3.1. Estructura, funcionamiento y proceso de trabajo

A continuación, y con la ayuda del artículo de **Martin Grandjean** (2015), se explica el funcionamiento del programa mediante el proceso de carga de, como ejemplo, la

Matriz resultante del *script* 2 (ver **4.3.2.2.2. Script 2**). Para ello, se debe proceder a seleccionar "Nuevo proyecto" en la inicialización de Gephi para, seguidamente, en el menú "Archivo", seleccionar "Importar hoja de cálculo…" y cargar el ".csv" correspondiente. En la ventana emergente (**Ilustración 25**) permite elegir (**1**) el separador de columnas correspondiente, el tipo de tabla a elegir y el conjunto de caracteres de los datos, es decir, la codificación para permitir la lectura de la máquina, por defecto UTF-8. Dicha codificación, para la tabla importada, se ha cambiado a ISO 8859-1, correspondiente al alfabeto latino. En esta ventana, también se observa una "Previsualización" (**2**) de la tabla importada para comprobar que todo esté correctamente.

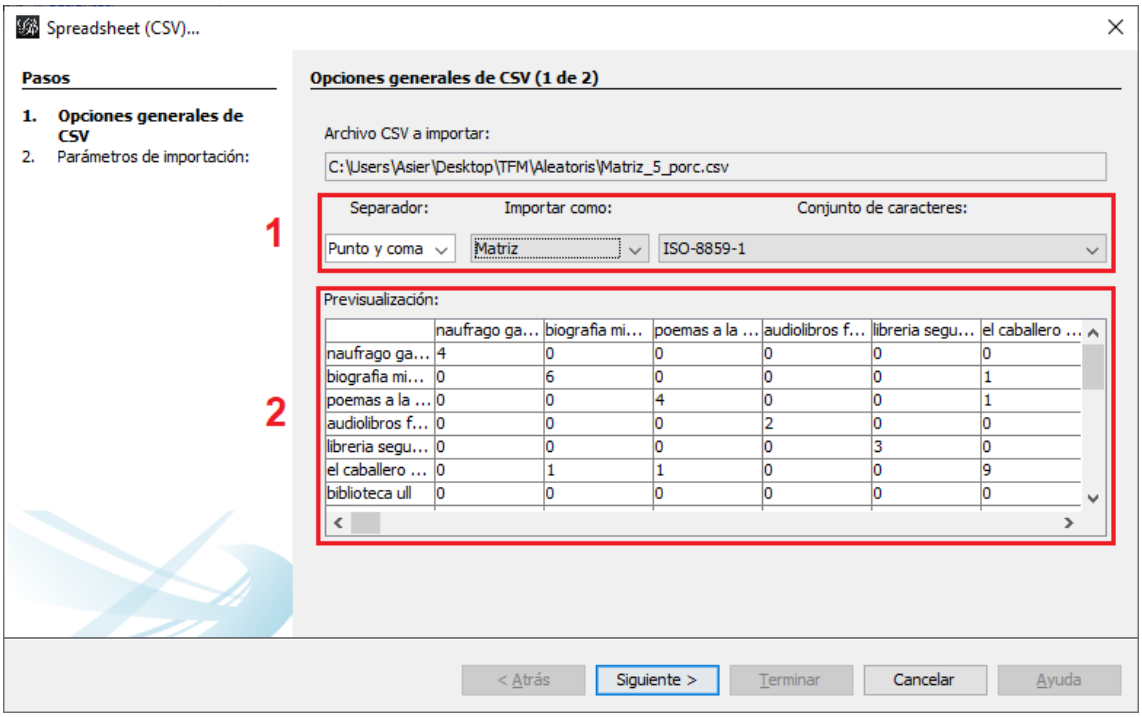


Ilustración 25. Ventana de importación en Gephi. Fuente: elaboración propia

El siguiente paso, los "Parámetros de importación", se deja cómo aparece por defecto, ya que en el caso de esta investigación, no es relevante. Posteriormente, emerge el "Informe de importación" (**Ilustración 26**), en el que se mostraría cualquier tipo de error o alerta (**1**) que se pudiera llegar a dar. En este paso, también se selecciona el tipo de grafo a visualizar (**2**). Dichos tipos se encuentran ya descritos en el estado del arte (ver **3.3. Teoría de grafos**) y a continuación se recuerda su descripción:

- **Dirigido:** Las aristas entre nodos tienen una dirección asignada según los datos.
- **No dirigido:** Las aristas consisten en relaciones simétricas, sin dirección.
- **Mixto:** Como su nombre indica, tiene la capacidad de mostrar tanto relaciones de dirección como de simetría.

Dada la naturaleza de la matriz importada (matriz simétrica), se selecciona el tipo de grafo no dirigido. En la misma ventana, también se informa al usuario del número de nodos y aristas (**3**) existentes en la matriz, en este caso 1.544 y 131.548, respectivamente.

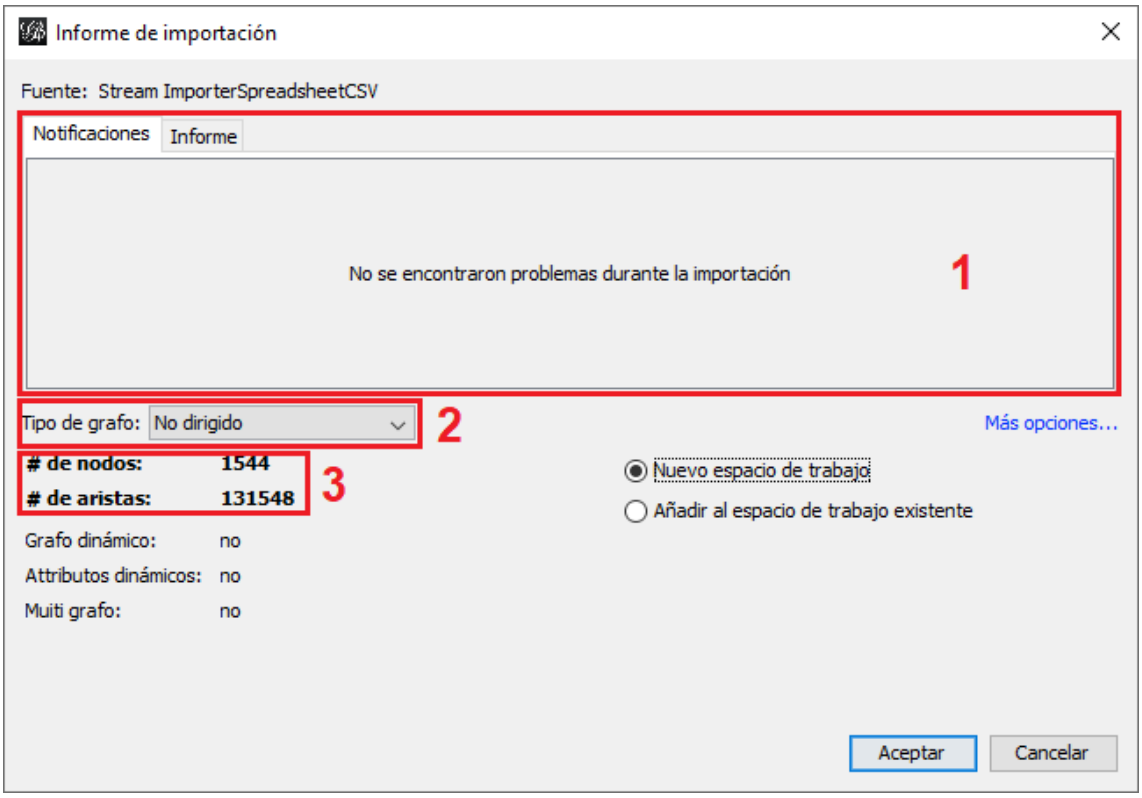


Ilustración 26. Informe de importación en Gephi. Fuente: elaboración propia

Una vez cargada la matriz, se visualiza una masa de nodos y aristas de color negro, ilegible e imposible de interpretar (**Ilustración 27**). Para transformar esta masa en un grafo interpretable y visualmente atractivo, se procedió a trabajar con la interfaz general de Gephi.

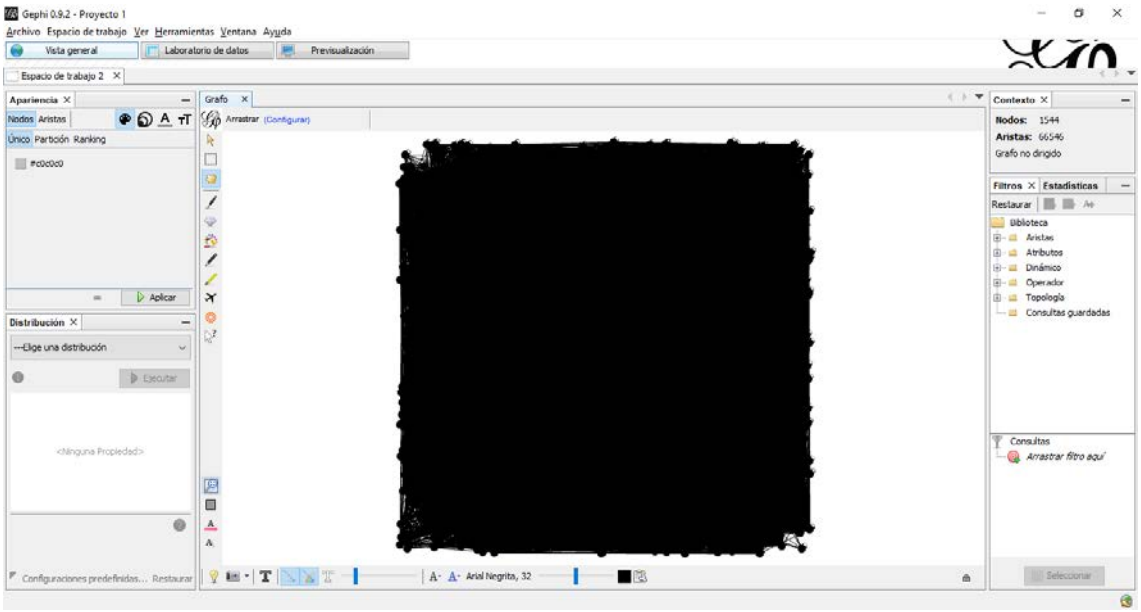


Ilustración 27. Masa de nodos inicial en Gephi. Fuente: elaboración propia

En dicha interfaz se observa, aunque inicialmente aparece el módulo “Vista general”, que se trabaja también con los módulos “Laboratorio de datos” y “Previsualización”. Seguidamente, se explica brevemente cada una de ellas:

- **Vista general (Ilustración 27):** Este módulo sirve para las funciones de visualización y modificación de los grafos. Se compone a su vez de las cinco pestañas siguientes totalmente modificables y adaptables en cuanto a tamaño se refiere:

    1. Apariencia (**Ilustración 28**): Permite seleccionar la coloración y el tamaño de los nodos y aristas según que atributos estadísticos se seleccionen, así como de sus etiquetas.

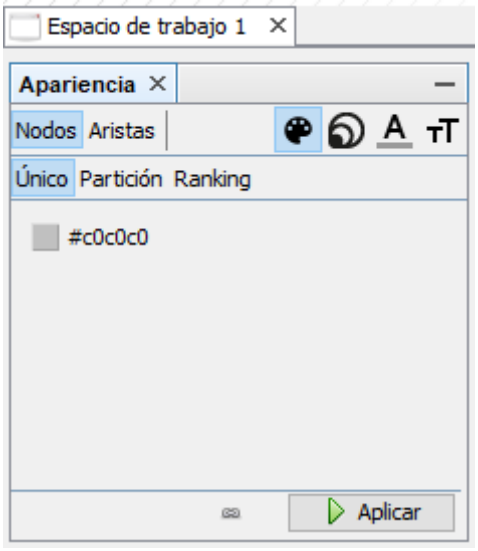


Ilustración 28. Pestaña Apariencia del módulo Vista general en Gephi. Fuente: elaboración propia

Para empezar a modificar la masa anterior (**Ilustración 27**), se procedió a modificar el tamaño de los nodos según el número de conexiones que les corresponde, atributo llamado grado (Apariencia > Nodos > Tamaño > Ranking > ---Escoge un atributo: Grado). Se seleccionó el tamaño mínimo y máximo por defecto, 10 y 100 respectivamente, y se aplicaron los cambios (**Ilustración 29**).

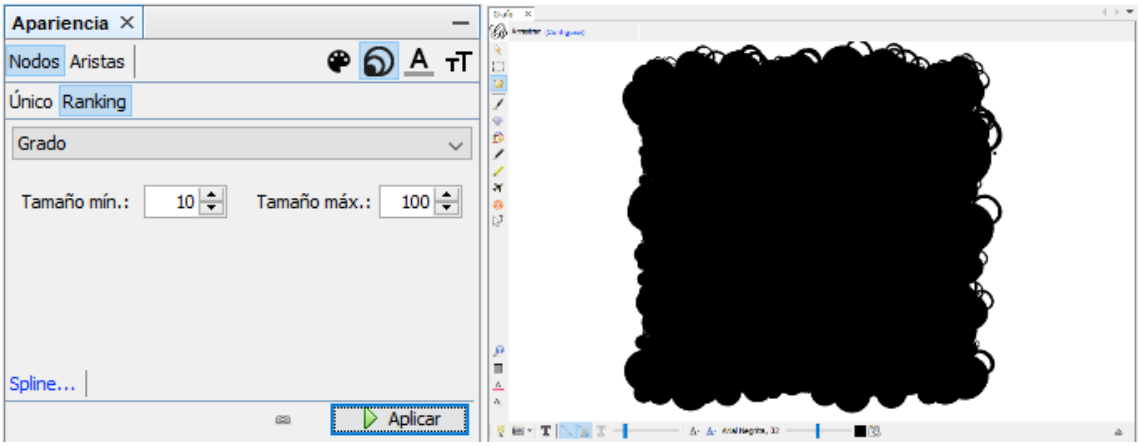


Ilustración 29. Cambio del tamaño de los nodos en Gephi. Fuente: elaboración propia

Como se observa, los nodos de la red han tomado diferentes formas según su Grado.

2. Distribución (**Ilustración 30**): Permite elegir el tipo de distribución de las visualizaciones que más se adapte a los datos introducidos en Gephi y seleccionar los parámetros pertinentes.

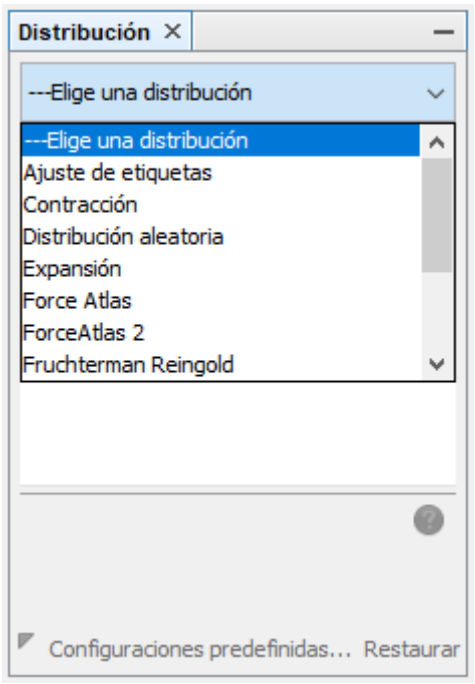


Ilustración 30. Pestaña Distribución del módulo Vista General en Gephi. Fuente: elaboración propia

Ofrece una variedad modesta de estructuras, de las cuales, en la **Ilustración 31** se muestran, como ejemplo, las distribuciones *Force Atlas 2* (**Jacomy**, 2011), *Force Atlas* y *Fruchterman Reingold* (**Fruchterman y Reingold**, 1991) aplicadas al grafo generado.

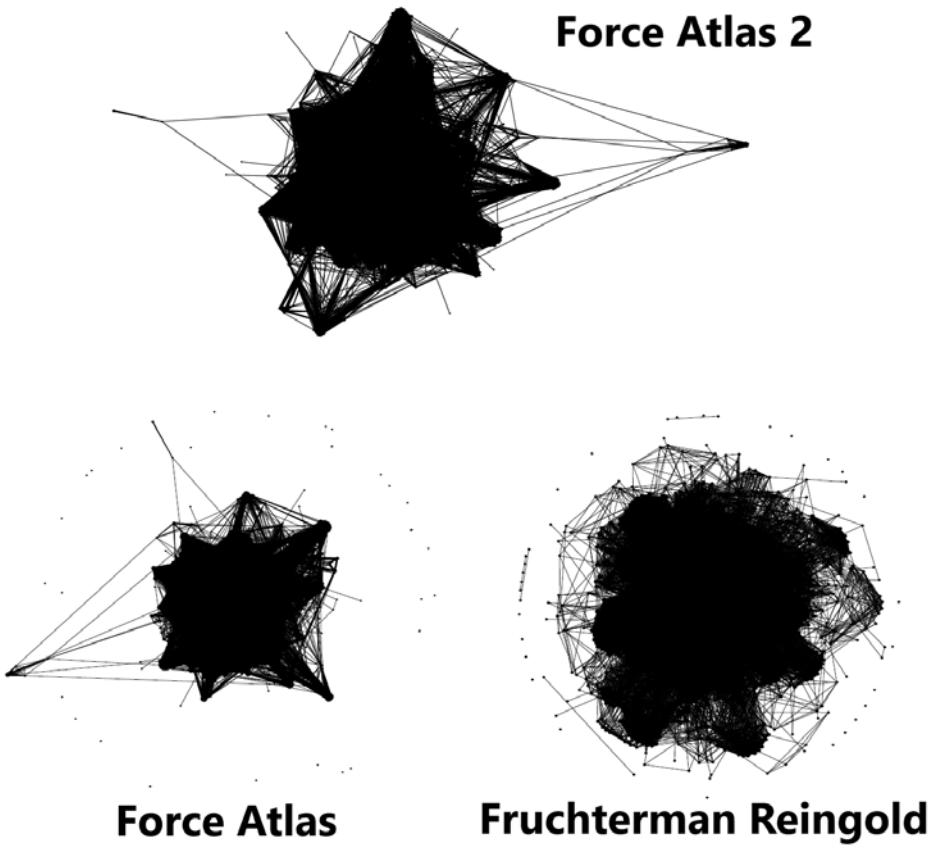


Ilustración 31. Ejemplos de distribuciones en Gephi. Fuente: elaboración propia

Cabe destacar que la ejecución de las distribuciones carga infinitamente, por lo que se debe pausar manualmente en cuanto se observa que el grafo está estable.

3. Grafo (**Ilustración 32**): Es la vista central del programa, donde aparecen y se modifican las visualizaciones a medida que se seleccionan distribuciones, atributos de apariencia, filtros, etc. En ella se puede realizar zoom, mover nodos, añadir o suprimir nodos y aristas, seleccionar colores y visualizar las etiquetas de los nodos entre otras opciones.

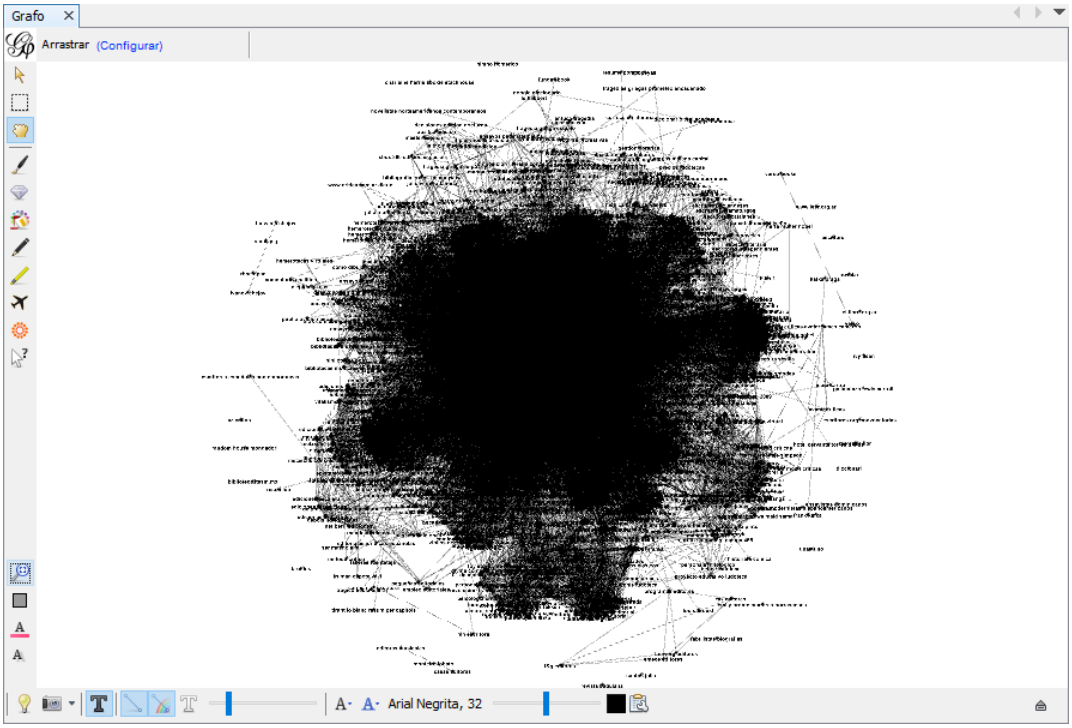


Ilustración 32. Pestaña Grafo el módulo Vista general en Gephi. Fuente: elaboración propia

4. Contexto (**Ilustración 33**): Esta pequeña pestaña muestra la información del grafo en cuanto a nodos, aristas y el tipo de grafo elegido (Dirigido, No dirigido o Mixto). En el caso de la matriz cargada, contiene 1544 nodos, 66546 aristas y se trata de un grafo no dirigido.

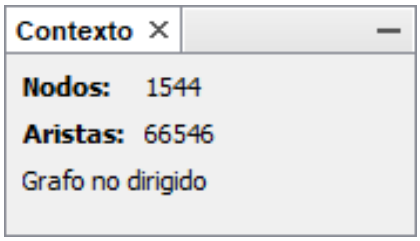


Ilustración 33. Pestaña Contexto en el módulo Vista general en Gephi. Fuente: elaboración propia

5. Filtros / Estadísticas:

   - Filtros (**Ilustración 34**): Se trata de un panel, dividido en dos partes (Biblioteca y Consultas), de tipo “arrastrar y soltar” (*Drag & Drop*) en el que se muestra una biblioteca de filtros navegable desde la que se pueden arrastrar dichos filtros a la

ventana inferior para generar consultas sobre los datos del grafo, que muestran u ocultan nodos o aristas según características como el peso, grado, texto, etc.

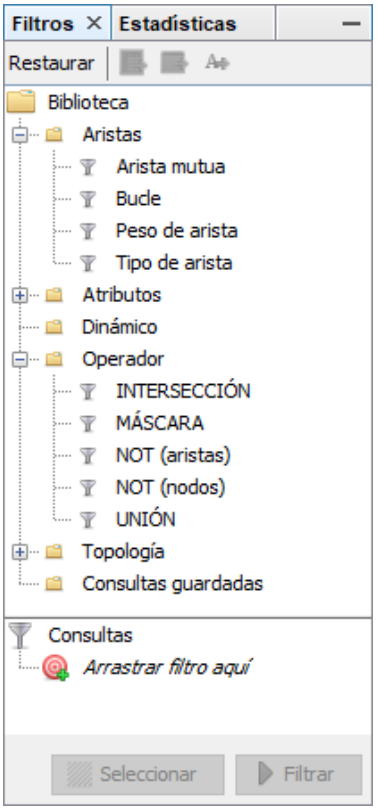


Ilustración 34. Pestañas Filtros del módulo Vista General en Gephi. Fuente: elaboración propia

Entre estos filtros, se pueden encontrar, por ejemplo, "Bucle", para eliminar aristas que generen bucles, "Peso de arista", para ver u ocultar aristas más o menos pesadas, atributos como el rango, la igualdad o particiones o hasta operadores booleanos como intersección, unión y NOT.

- Estadísticas (**Ilustración 35**): La ejecución de estadísticas permite darle color al grafo, desde la pestaña Apariencia (Partición y Ranking) (**Ilustración 28**), para así poder distinguir de forma más cómoda las comunidades o clústeres que formen los datos.

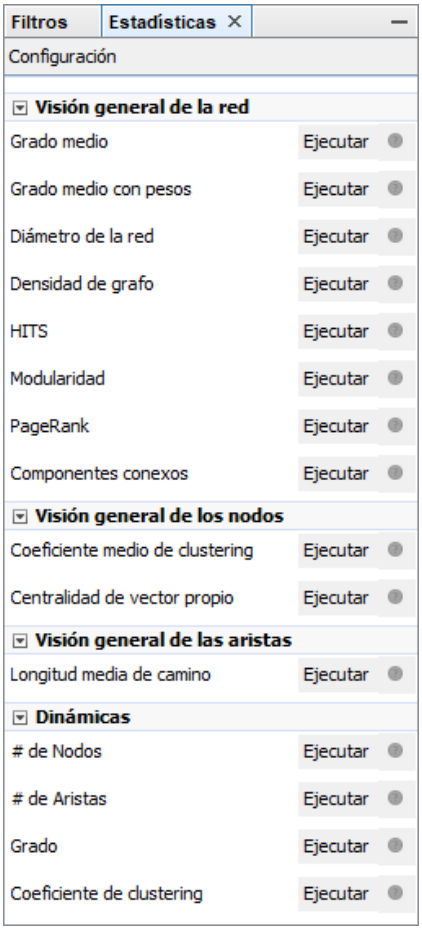


Ilustración 35. Pestaña Estadísticas de la Vista general en Gephi. Fuente: elaboración propia

Al seleccionar cualquier medida estadística, Gephi muestra una ventana nueva con la descripción pertinente de la medida y de los parámetros seleccionables (**Ilustración 36**). Algunas de las medidas disponibles más importantes son la Modularidad, las estadísticas de Grados o las de Coeficiente de clustering.

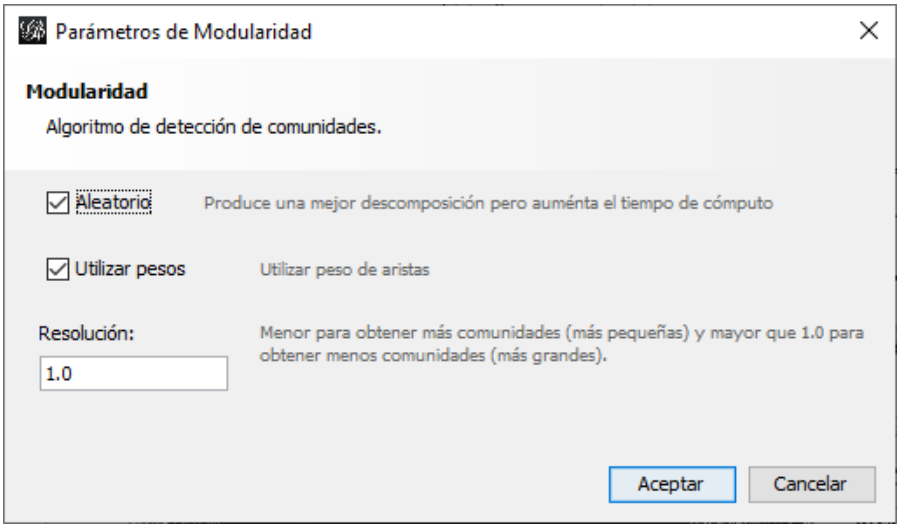
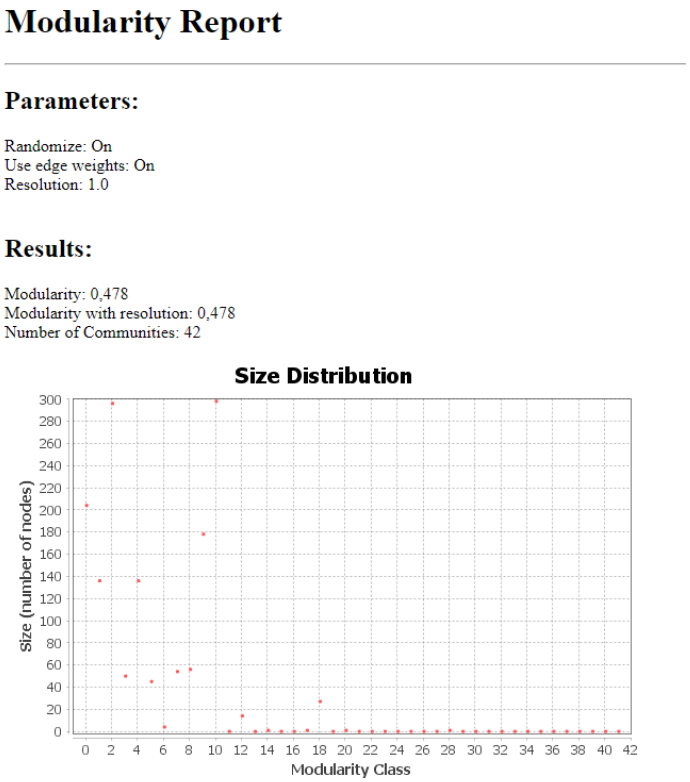


Ilustración 36. Ventana de parámetros de la estadística Modularidad en Gephi. Fuente: elaboración propia

Una vez ejecutada la estadística, se genera un informe (**Ilustración 37**) con, en el caso de la Modularidad, los parámetros de la estadística y sus resultados, como, por ejemplo, el número de comunidades detectadas, en este caso 42.



Modularity Report

Parameters:

Randomize: On
Use edge weights: On
Resolution: 1.0

Results:

Modularity: 0,478
Modularity with resolution: 0,478
Number of Communities: 42

Size Distribution

Size (number of nodes)

Modularity Class

Ilustración 37. Ejemplo de reporte de Modularidad en Gephi. Fuente: elaboración propia

Finalmente, acudiendo al apartado Partición de la pestaña Apariencia (**Ilustración 38**), seleccionando "*Modularity Class*" en el desplegable y aplicando los cambios, se colorea el grafo.

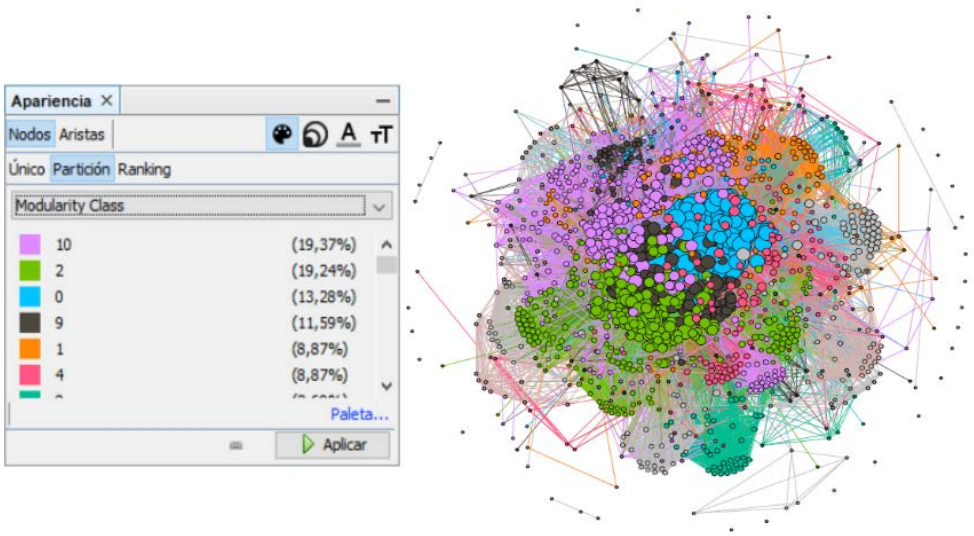


Ilustración 38. Coloreado del grafo en Gephi. Fuente: elaboración propia

- **Laboratorio de datos (Ilustraciones 39 y 40):** El segundo módulo de Gephi muestra toda la información referente a los datos importados al programa. Dicha información es totalmente organizable y modificable, pudiendo también añadir nueva o eliminar información prescindible. Está separado en dos tablas, la tabla de nodos y la tabla de aristas:

  - Tabla de nodos (**Ilustración 39**): en ella se observa la información disponible de cada nodo, en este caso: la ID, la etiqueta, y la estadística seleccionada anteriormente.

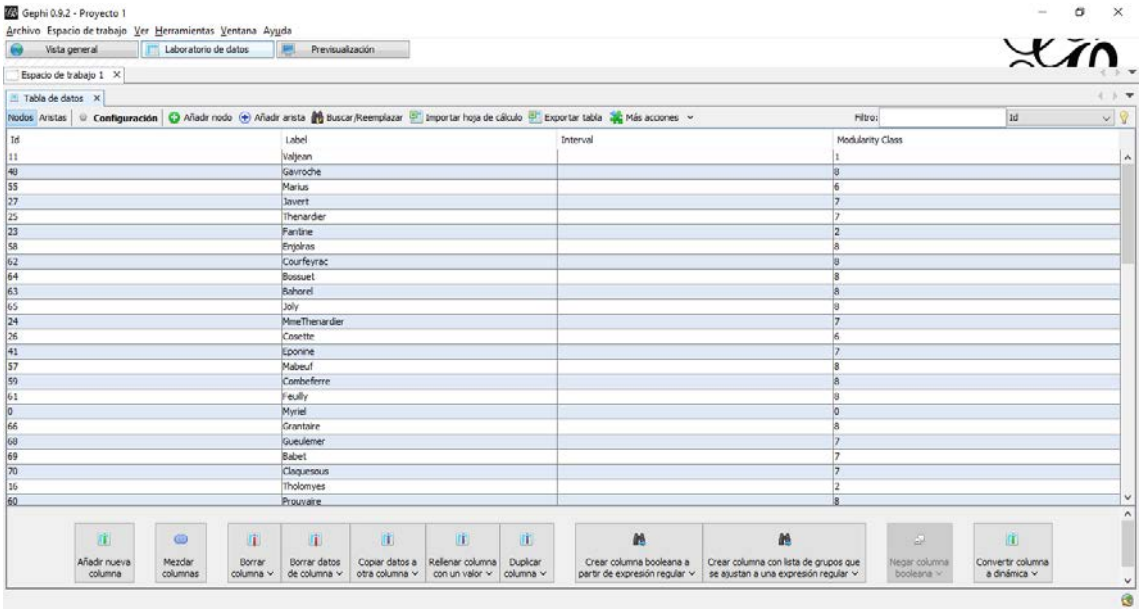


Ilustración 39. Tabla de nodos del módulo Laboratorio de datos en Gephi. Fuente: elaboración propia

  - Tabla de aristas (**Ilustración 40**): En esta tabla se muestra la información de las aristas, siendo esta: su nodo origen y su nodo destino, el tipo de arista, la ID, la etiqueta, el intervalo y su peso.

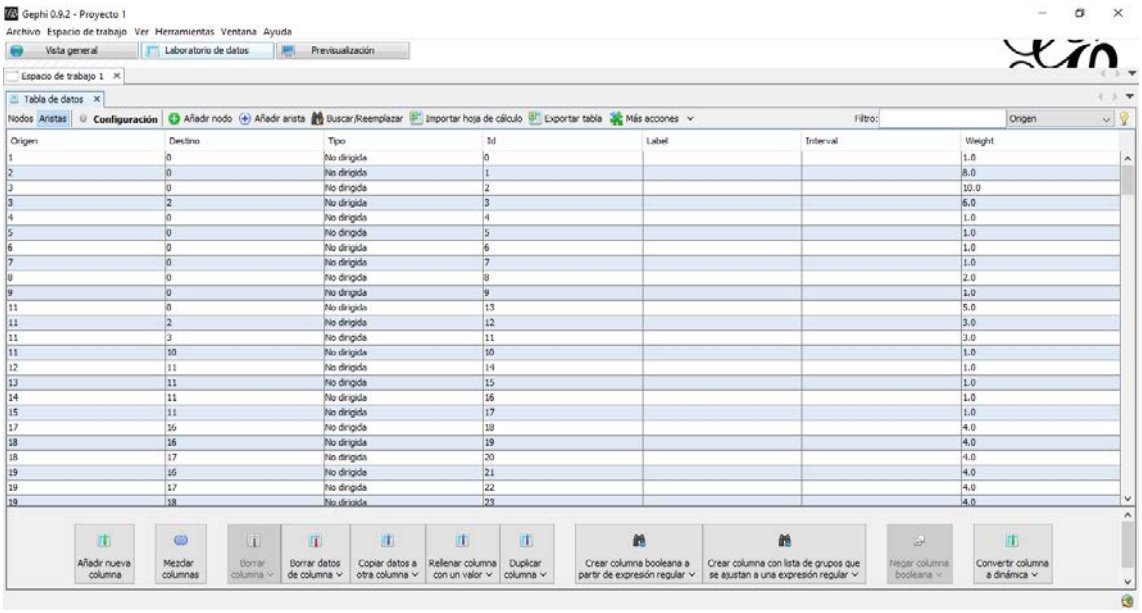


Ilustración 40. Tabla de aristas del modulo Laboratorio de datos en Gephi. Fuente: elaboración propia

Los datos mostrados en estas tablas son completamente ordenables por cada una de las columnas presentes en ellas, por ejemplo, alfabético por etiqueta o de mayor a menor por peso.

- **Previsualización (Ilustración 41):** El último módulo de trabajo del programa, es, a su vez, la última serie de pasos a realizar antes de llevar a cabo la exportación de grafos, ya que en él se modifican las características estéticas definitivas del grafo, por ejemplo, convertir las aristas rectas en curvas.

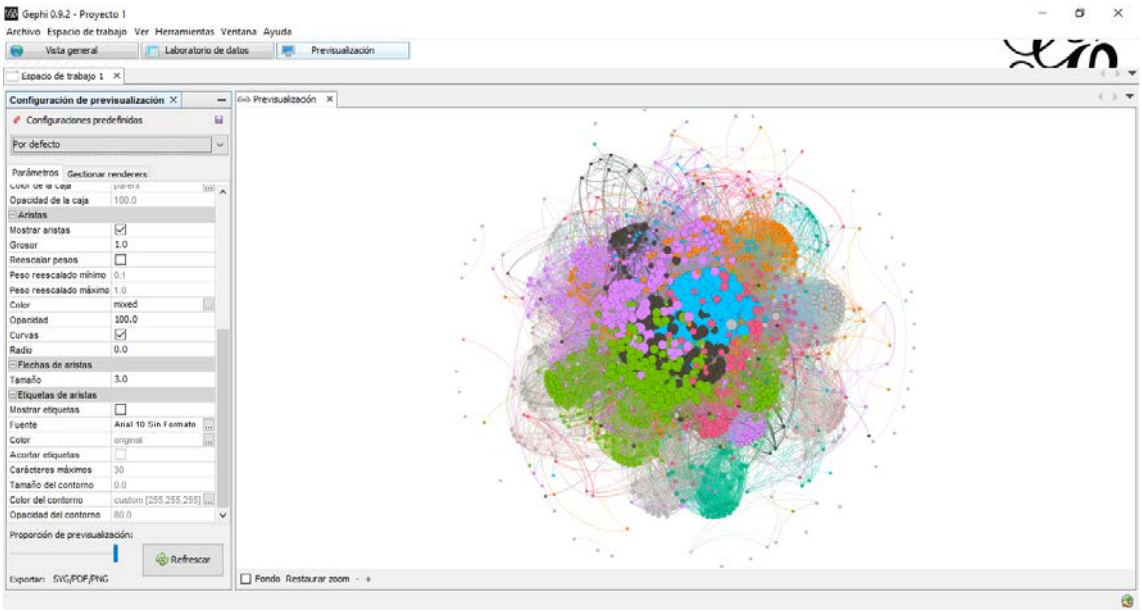


Ilustración 41. Módulo Previsualización en Gephi. Fuente: elaboración propia

4.3.3.2. Elección de la distribución del grafo a utilizar

Los algoritmos de distribución elegidos de entre los ofrecidos por Gephi han sido, como se ha mencionado anteriormente, los siguientes: *Force Atlas*, *Force Atlas 2* y *Fruchterman Reingold*. Realizando una pequeña comparativa entre los tres para elegir uno concreto a utilizar en la investigación, se decidió lo explicado a continuación.

En primer lugar, el hecho de que *Force Atlas 2* fuese como una especie de actualización o nueva versión de *Force Atlas*, propició el descarte de este último con bastante facilidad.

Por otra parte, tanto *Force Atlas* como *Force Atlas 2*, implican una fuerza de repulsión entre nodos demasiado alta, lo que provoca que algunos nodos se expandan a mucha distancia de la parte común del grafo, complicando así el análisis del grafo generado, sobre todo en pantallas como la del **Dispositivo 1** (ver **4.2. Hardware**).

En cuanto a visualización gráfica, *Fruchterman Reingold* resulta mucho más agradable a la vista que las otras dos distribuciones, debido a su simulación del gráfico como un sistema de partículas de masa: “los nodos son partículas de masa y las aristas son resortes entre dichas partículas” (**Heymann**, 2011), que mantiene las comunidades más concentradas alrededor del centro.

Por las razones expuestas se eligió el algoritmo *Fruchterman Reingold* para llevar a cabo las visualizaciones de los grafos.

4.3.3.3. Requisitos mínimos específicos

Según **Bastian, Heymann y Jacomy** (2009), Gephi requiere versión de Java 7 o superior. En este trabajo, con el **Dispositivo 1** (ver **4.2. Hardware**) concretamente, se ha usado Java 8.

También requiere 500 MHz de CPU, 128MB de memoria RAM y la versión 1.2 de OpenGL (Open Graphics Library) para la generación de gráficos 2D o 3D. Otro requisito no tan importante en cuanto a esto último es que la tarjeta gráfica no tenga más de 8 años de antigüedad, siendo la presencia de cualquier tipo de estas un requisito indispensable.

Network size (nodes + edges)	~Memory suggested
~1000	128mo
~10,000	512mo
~100,000	2go
~1M	>8go

Ilustración 42. Relación tamaño-RAM de los requisitos de Gephi. Fuente: elaboración propia

En la **Ilustración 42** se observa la relación entre la talla de la red cargada y la cantidad de memoria RAM exigida.

5. Resultados

Para la creación de los grafos, se escogió el algoritmo *Fruchterman Reingold* (ver **4.3.3.2. Elección de la distribución del grafo a utilizar**) por su forma de distribución de los nodos y aristas más uniforme y agradable de manejar, resultando bastante sencillo visualizar comunidades y relaciones a simple vista una vez aplicadas las medidas estadísticas.

Para la aplicación del algoritmo *Fruchterman Reingold* el programa ofrece tres parámetros modificables: Área, Gravedad y Velocidad. El Área representa, como su nombre indica, el área que va a ocupar el grafo, la Gravedad indica la fuerza de atracción entre nodos (como planetas, cuanta más gravedad más atracción, según **Stox** (2017)) y la Velocidad representa la velocidad de convergencia, cuanta más menor precisión. En el presente trabajo se han elegido dichos parámetros con los valores siguientes respectivamente: 20.000, 10.0 y 10.0.

Las medidas estadísticas mencionadas anteriormente fueron, primero, el tamaño de los nodos según el grado y, segundo, la estadística de Modularidad para colorear el grafo.

Por un lado, el grado de un nodo, recordando lo explicado en el estado del arte (ver **3.3. Teoría de grafos**), representa el número de aristas que convergen en un mismo nodo.

Por otro lado, la Modularidad es un algoritmo creado para la detección de comunidades. En los parámetros de su ejecución se puede elegir una resolución, que cuanto mayor resulta, menos comunidades aparecen y cuanto menor, más lo hacen. En esta investigación, dicha resolución se ha dejado por defecto, en 1.

5.1. Grafo 1: Matriz del 5%

A partir de la extracción de resultados explicada en el subapartado Script 2 de RStudio localizado en la Metodología (ver **4.3.2.2.2. Script 2**), se procedió a testar la importación de los diferentes CSV generados, como se explica en el apartado mencionado, de mayor a menor porcentaje.

El máximo porcentaje que se pudo importar en el **Dispositivo 1** (ver **4.2. Hardware**), resultó ser el 5%, por lo que se procedió a trabajar con dicha matriz. Siguiendo los pasos explicados en el apartado Gephi de la Metodología (ver **4.3.3.1. Estructura, funcionamiento y proceso de trabajo**), se generó el siguiente grafo utilizando la distribución basada en el algoritmo de *Fruchterman Reingold*, de la que vemos la visualización gráfica resultante continuación (**Ilustración 43**).

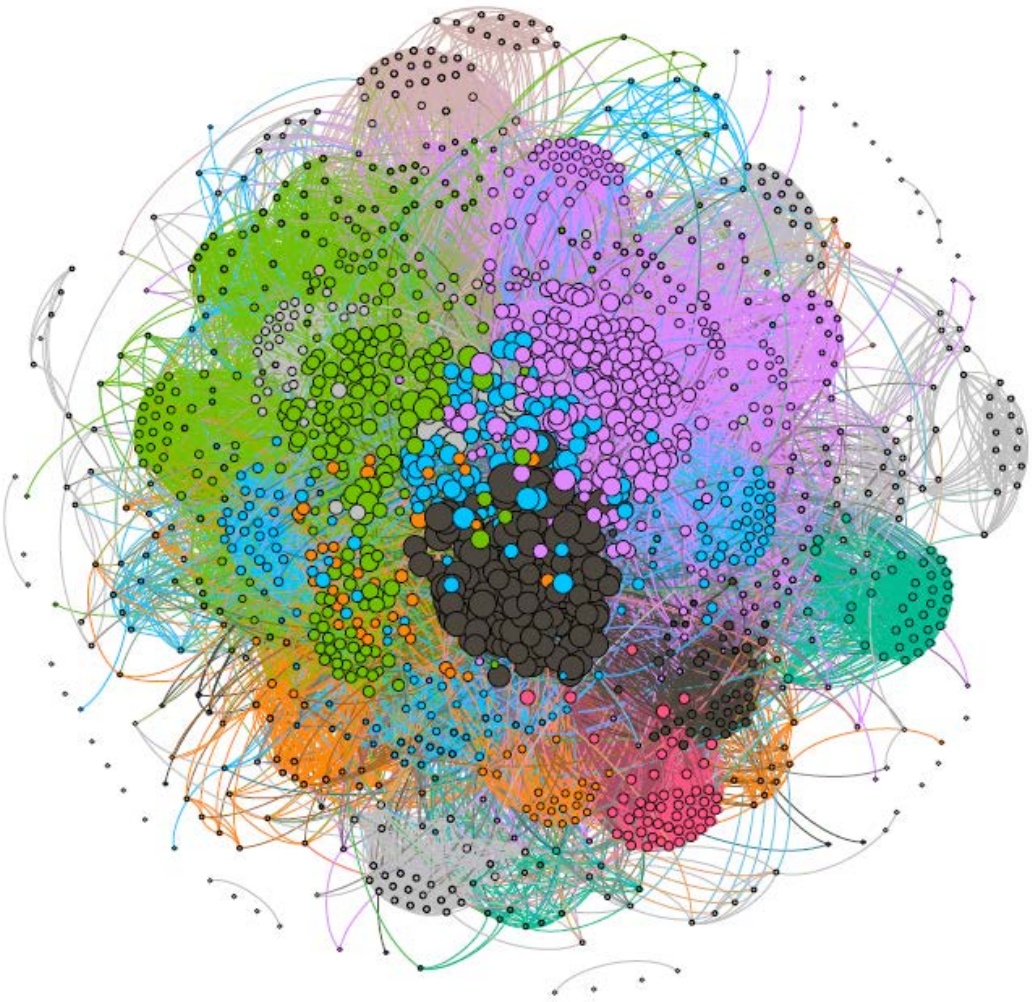
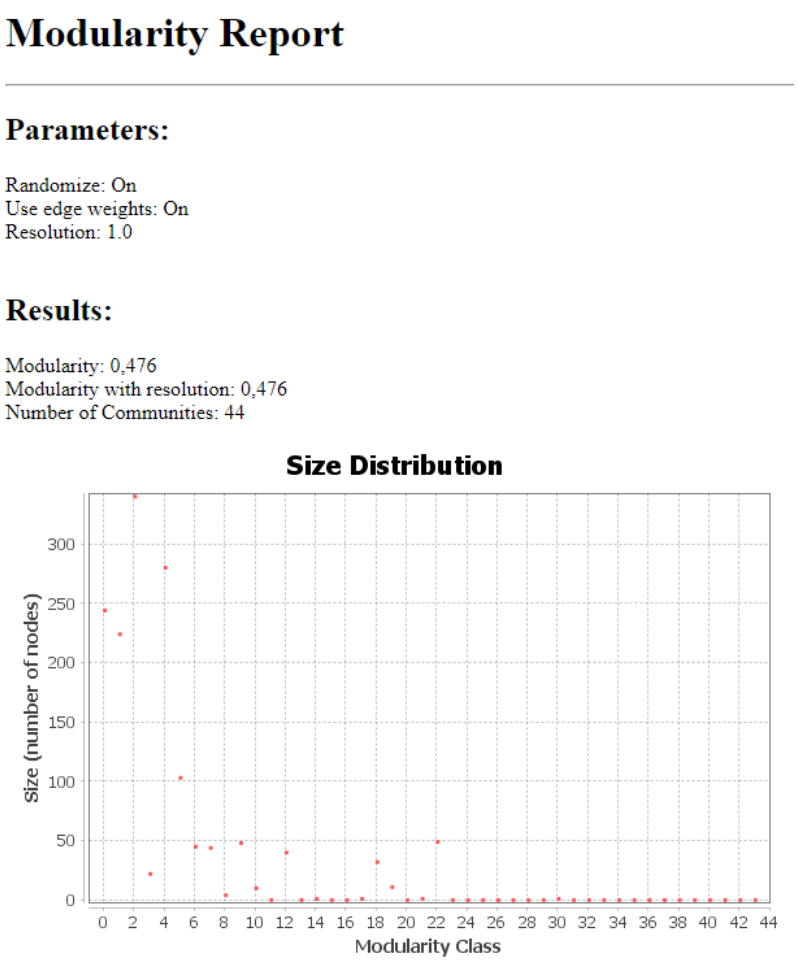


Ilustración 43. Grafo 1: Visualización gráfica. Fuente: elaboración propia

El grafo generado, según la pestaña Contexto de Gephi, se halla compuesto por 1.544 nodos y 66.546 aristas y se trata de un grafo de tipo no dirigido, tipo de grafo que se elegirá en cualquier grafo realizado en este trabajo debida la tipología simétrica de la matriz.

En la visualización anterior se observan bastantes comunidades bien diferenciadas gracias a los colores aportados por la estadística de Modularidad, indicando en su informe un total de 44 de estas (**Ilustración 44**).



Modularity Report

Parameters:

Randomize: On
Use edge weights: On
Resolution: 1.0

Results:

Modularity: 0,476
Modularity with resolution: 0,476
Number of Communities: 44

Ilustración 44. Grafo 1: Reporte de Modularidad. Fuente: elaboración propia

Volviendo al grafo y ampliándolo un poco para observar, primero, los nodos en sí, se observa que cada uno de los nodos tiene una arista de tipo lazo o bucle (**Ilustración 45**).

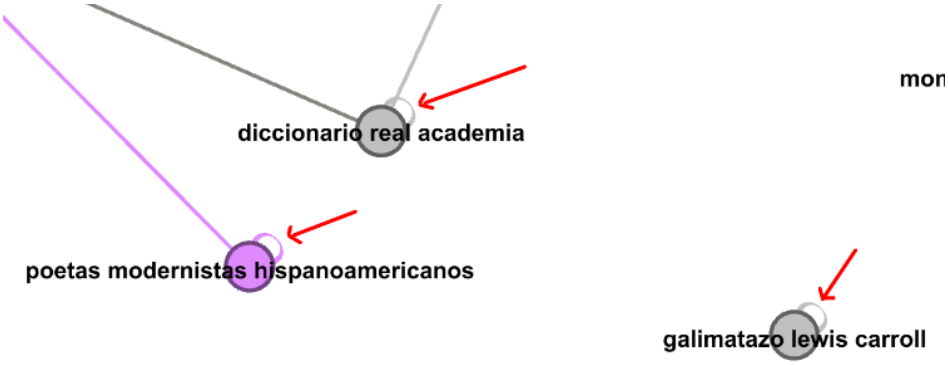


Ilustración 45. Grafo 1: Ejemplo de lazos o bucles. Fuente: elaboración propia

Analizando la matriz, se deduce que este hecho se debe a la diagonal de esta, marcada en amarillo en la **Ilustración 46**, correspondiente al recuento de palabras contenidas en expresiones de búsqueda sobre sí mismas, debido a la simetría de la matriz.

	A	B	C	D	E
1		naufrago gabriel garcia marquez	biografia miguel de cervantes y saavedra	poemas a la muerte	audiolibros frances
2	naufrago gabriel garcia marquez	4	0	0	0
3	biografia miguel de cervantes y saavedra	0	6	0	0
4	poemas a la muerte	0	0	4	0
5	audiolibros frances	0	0	0	2

Ilustración 46. Matriz 5%: Ejemplo diagonal. Fuente: elaboración propia

Dado que esta coincidencia total en muchas expresiones de búsqueda supone un número bastante alto de coincidencia que puede interferir en la visualización del grafo siendo totalmente redundante, se decide eliminar dicha diagonal de la matriz, proceso que se encuentra explicado en el subapartado Script 3 del apartado RStudio en Metodología (ver **4.3.2.2.3. Script 3**).

5.2. Grafo 2: Matriz del 5% sin diagonal

Una vez eliminada la diagonal (ver **4.3.2.2.3. Script 3**) de la matriz del 5% y extraído el CSV correspondiente, se importó el nuevo archivo en Gephi, repitiendo los mismos pasos ya llevados a cabo con anterioridad (**Ilustración 47**).

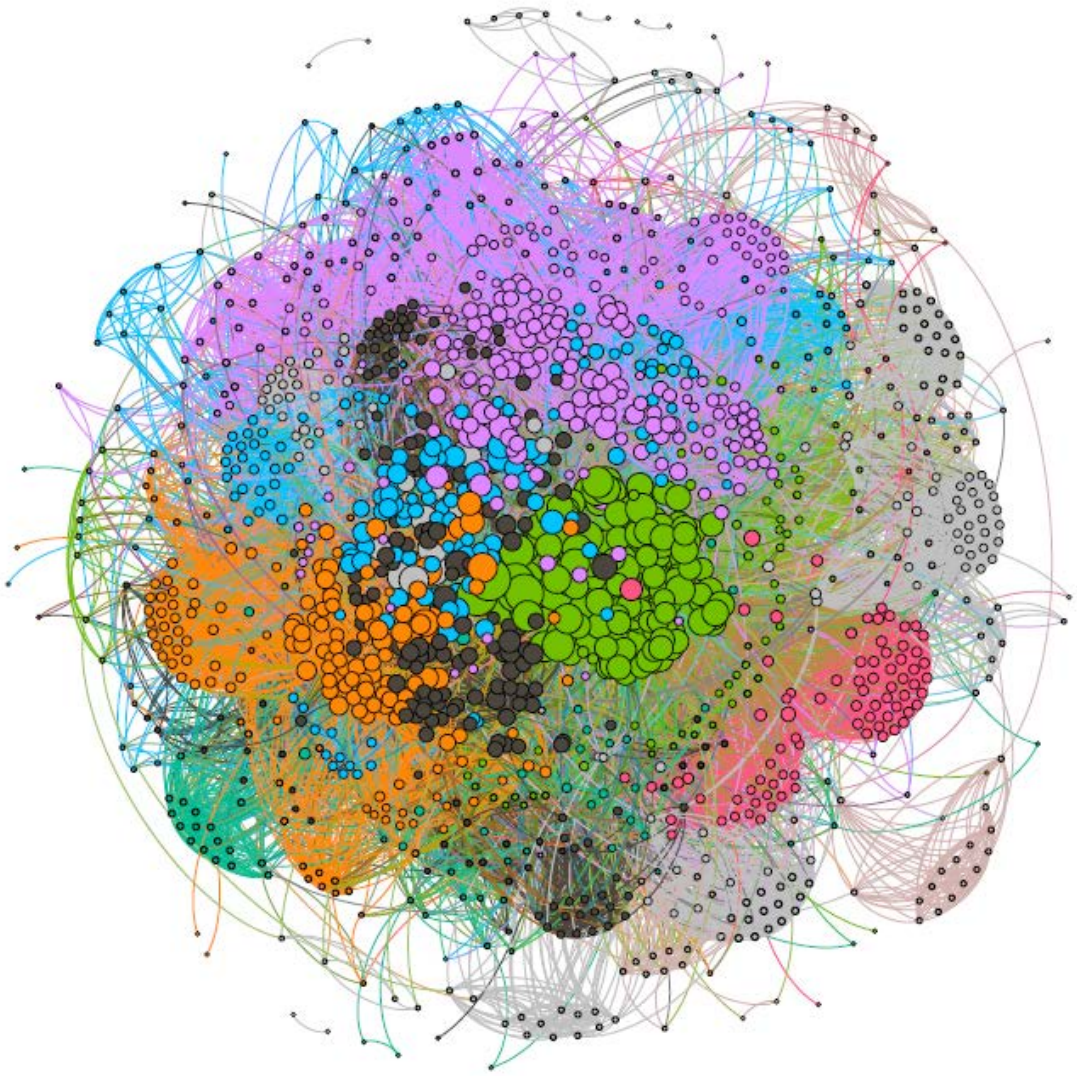


Ilustración 47. Grafo 2: Visualización gráfica. Fuente: elaboración propia

Esta vez, el grafo generado, según la pestaña Contexto, se compone de 1.519 nodos y 65.002 aristas, exactamente 1.544 aristas menos que en el grafo anterior, siendo este último número el total de celdas que componían la diagonal. Por lo que se puede observar, la eliminación de ella efectivamente quitó algunos datos redundantes o ruido del grafo.

Generando el informe de modularidad de este nuevo grafo (**Ilustración 48**), se observó que el número de comunidades se había reducido de 44, en el grafo anterior, a 18.

Modularity Report

Parameters:

Randomize: On
Use edge weights: On
Resolution: 1.0

Results:

Modularity: 0,472
Modularity with resolution: 0,472
Number of Communities: 18

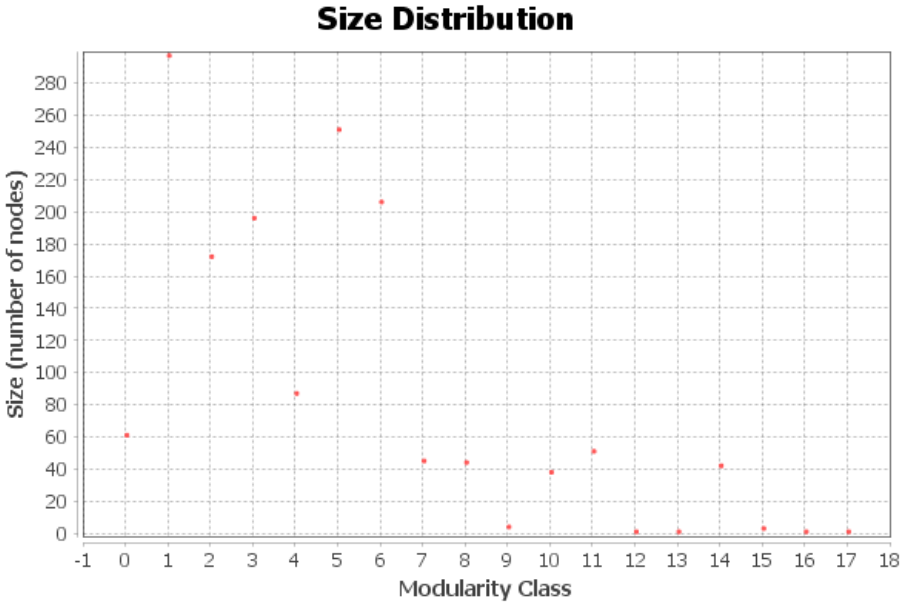


Ilustración 48. Grafo 2: Reporte de Modularidad. Fuente: elaboración propia

Aunque el número de nodos, sumado a la extensión de algunas etiquetas y la mezcla de diversas comunidades lo dificultaban, explorando dichas comunidades en el grafo, se observó a simple vista que, las más grandes, se conforman en su mayoría de expresiones de búsqueda relacionadas por palabras clave como “de”, “del”, “el” o para (**Ilustración 49**) o “a” y “y”, que simplemente actúan como conectores en una expresión de búsqueda y carecen de cualquier significado. Las comunidades influenciadas por estas palabras, se acumulan y entremezclan en su mayoría en el centro del grafo, careciendo este cúmulo de nodos de ningún significado aparente.

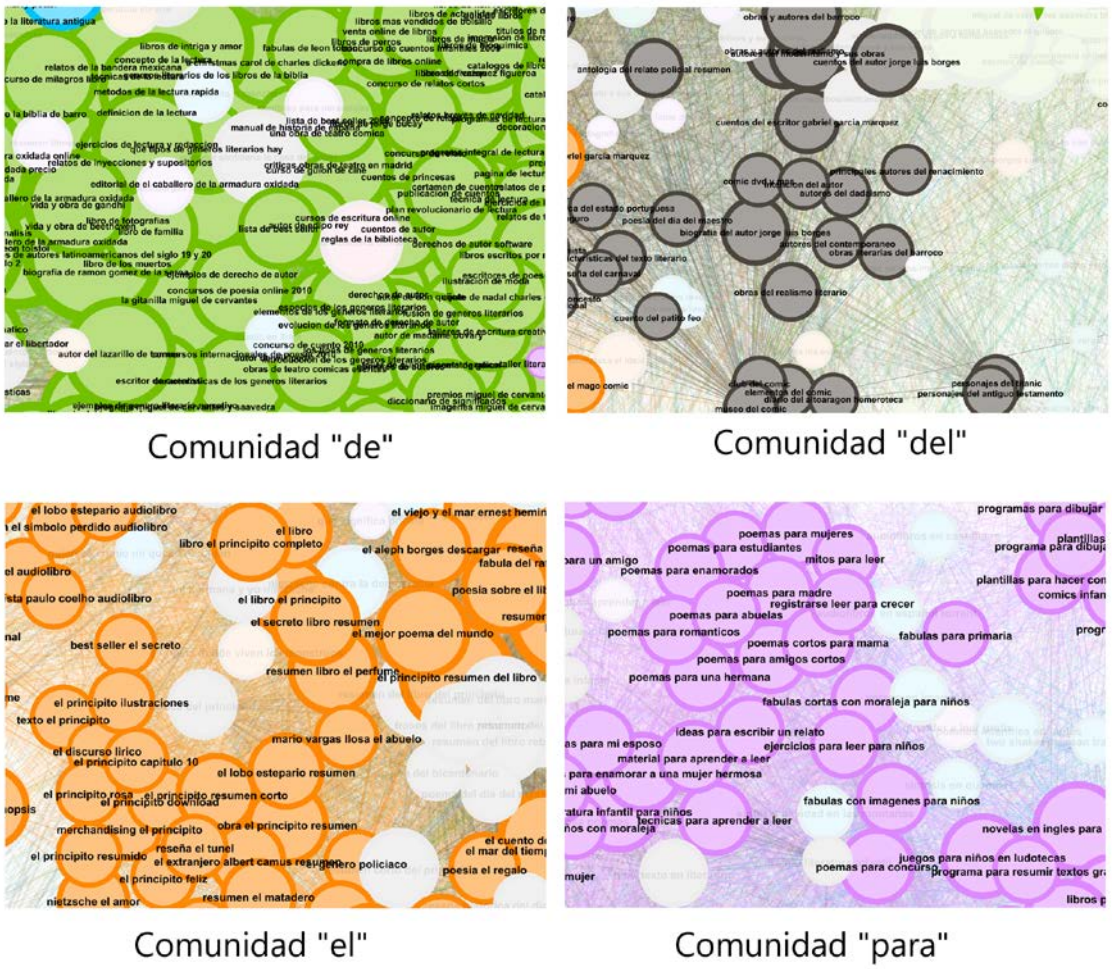


Ilustración 49. Grafo 2: Comunidades "de", "del", "el" y "para". Fuente: elaboración propia

En las comunidades que se distinguen más claramente, ya que se hallan separadas de este cúmulo central, se pueden observar relaciones más evidentes y significativas entre los nodos que las componen, pudiendo detectar fácilmente las palabras clave en común, como es el ejemplo de las comunidades de las palabras clave “biblioteca” o “nietzsche” (**Ilustración 50**) que, en su gran mayoría, no contienen conectores de los mencionados anteriormente, que provocarían su mezcla con el cúmulo central.

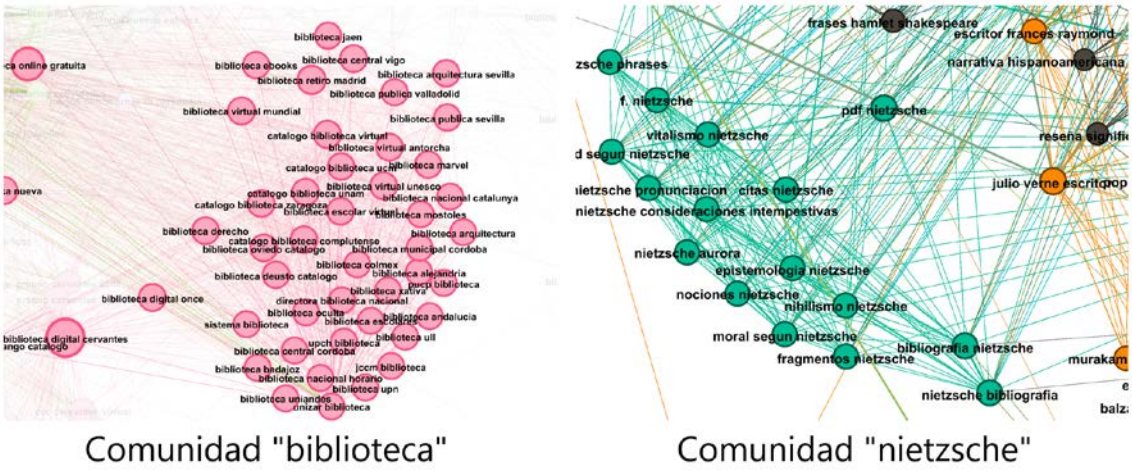


Ilustración 50. Grafo 2: Comunidades "biblioteca" y "nietzsche". Fuente: elaboración propia

Debido a los hechos comentados, se decidió, observando el grafo y el laboratorio de datos, realizar una lista de las palabras clave que más ruido generaban para, modificando la matriz de nuevo, eliminarlas de las expresiones de búsqueda y así, poder visualizar comunidades temáticas de una forma más clara. Este proceso se encuentra explicado en el subapartado Script 4 (ver **4.3.2.2.4. Script 4**) y la lista de palabras clave a eliminar fue la siguiente: “y”, “el”. “la”, “los”, ”las”, “por”, “para”, “a“, “de”, “del” y “en”.

5.3. Grafo 3: Matriz del 5% sin diagonal ni ruido

Ya eliminadas las palabras clave redundantes y rehecha la matriz (ver **4.3.2.2.4. Script 4**), el nuevo CSV resultante se importó en Gephi y, después de aplicarle el mismo proceso que a los anteriores grafos, se observó en la pestaña contexto que este grafo estaba formado por 1.506 nodos y 28.242 aristas.

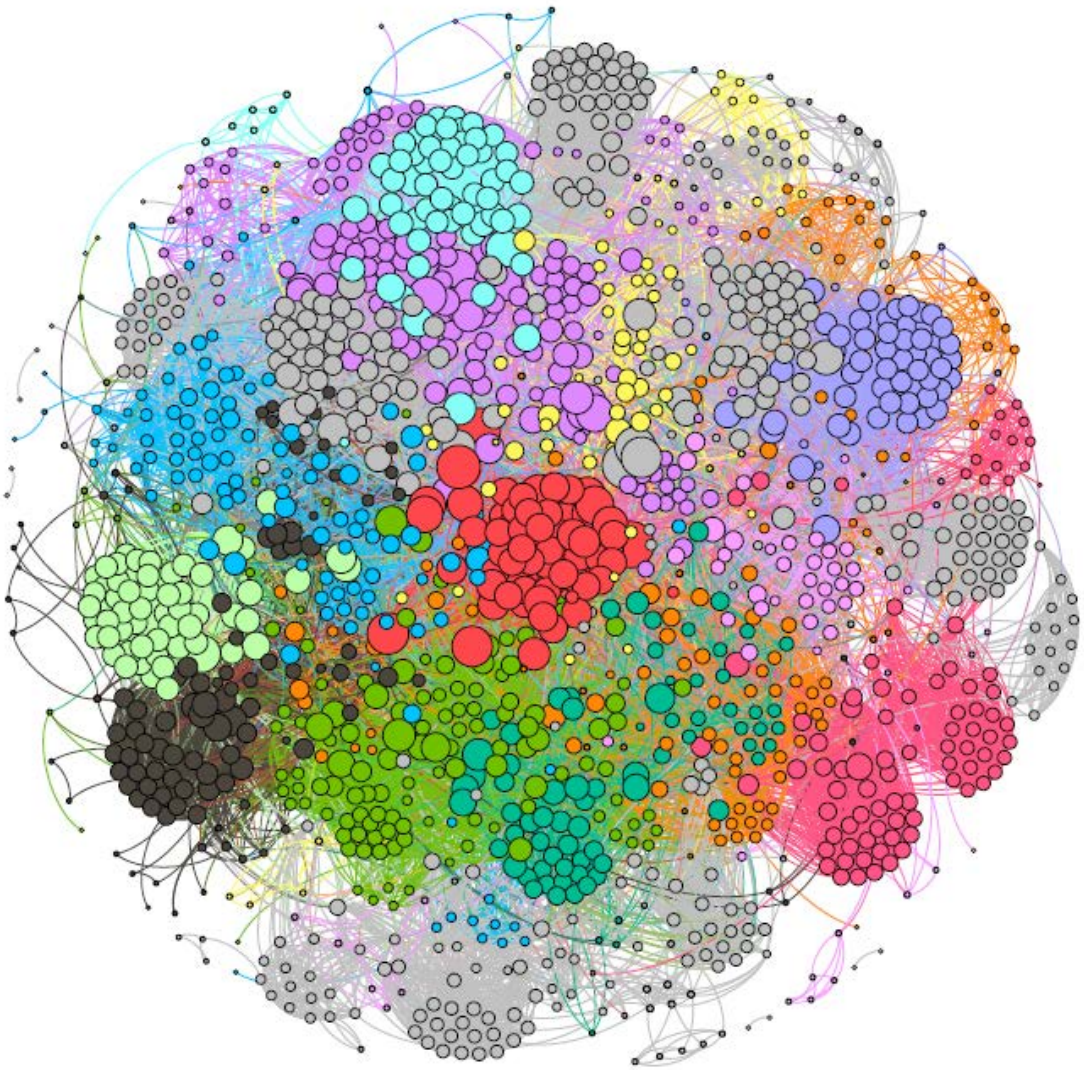


Ilustración 51. Grafo 3: Visualización gráfica. Fuente: elaboración propia

Como se observa en el grafo (**Ilsutración 51**), en esta ocasión aparecieron más comunidades sin color (gris) que en las anteriores ocasiones y, esta vez, algunas de ellas en posiciones importantes dentro del grafo. Según el informe de modularidad generado (**Ilustración 52**), la red consta de 27 comunidades, de las cuales, observando la pestaña Apariencia, 14 no han sido coloreadas.

Modularity Report

Parameters:

Randomize: On
Use edge weights: On
Resolution: 1.0

Results:

Modularity: 0,768
Modularity with resolution: 0,768
Number of Communities: 27

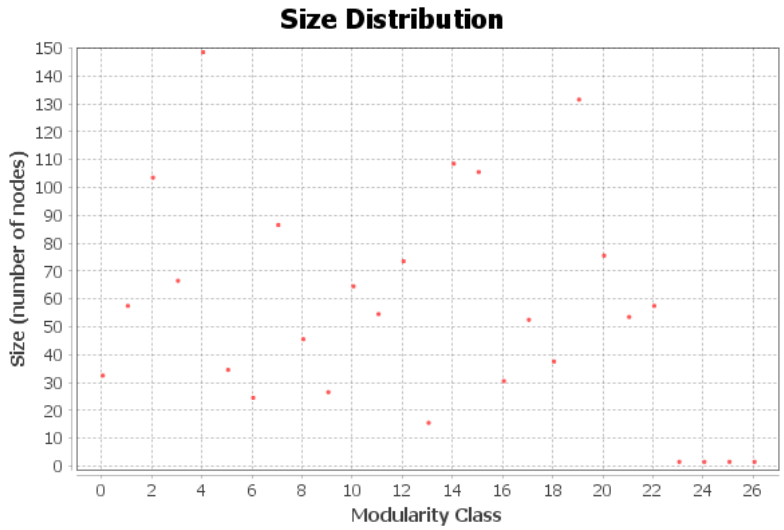


Ilustración 52. Grafo 3: Reporte de Modularidad. Fuente: elaboración propia

Como este hecho puede llegar a confundir en la visualización juntando comunidades diferentes que podrían no tener mucho que ver, se procedió a colorearlas para evitarlo, quedando tal como se ve en la **Ilustración 53**.

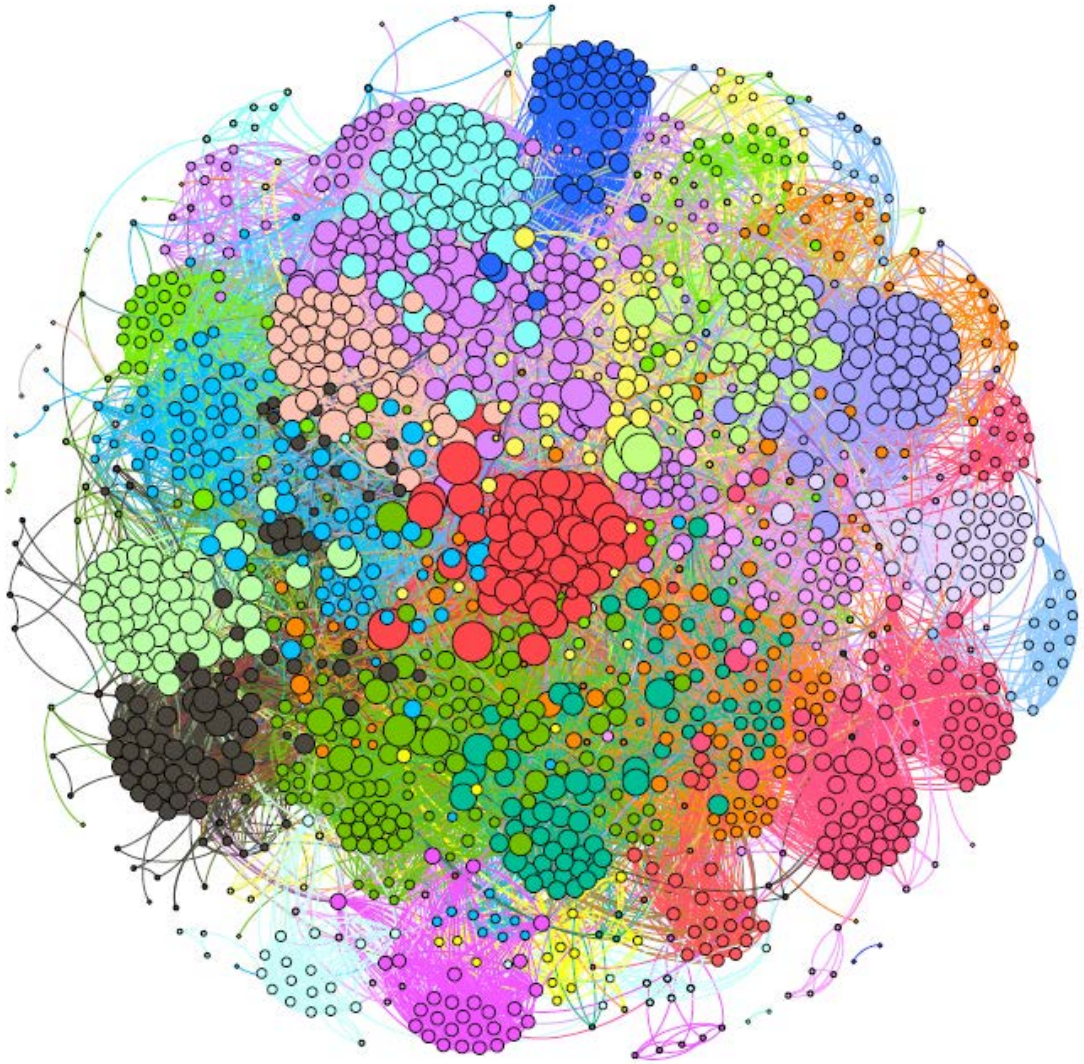


Ilustración 53. Grafo 3: Visualización gráfica coloreada. Fuente: elaboración propia

5.3.1. Comunidades

Es en este momento cuando se procedió al análisis de las comunidades encontradas en este grafo. A continuación se presenta un desglose de ellas (**Ilustración 54**) definidas por su palabra clave más importante (más repetida), explicando las consideradas como más relevante y, indicando de cada una de ellas, los nodos más relevantes de cada una según su grado, de mayor a menor.

Este último paso se llevó a cabo gracias al módulo Laboratorio de datos de Gephi, en el que, después de ejecutar la estadística Grado, se ordenaron las expresiones de

búsqueda por este. Una vez ordenadas, se introdujo en el buscador de este módulo la palabra clave y se analizaron los resultados ofrecidos.

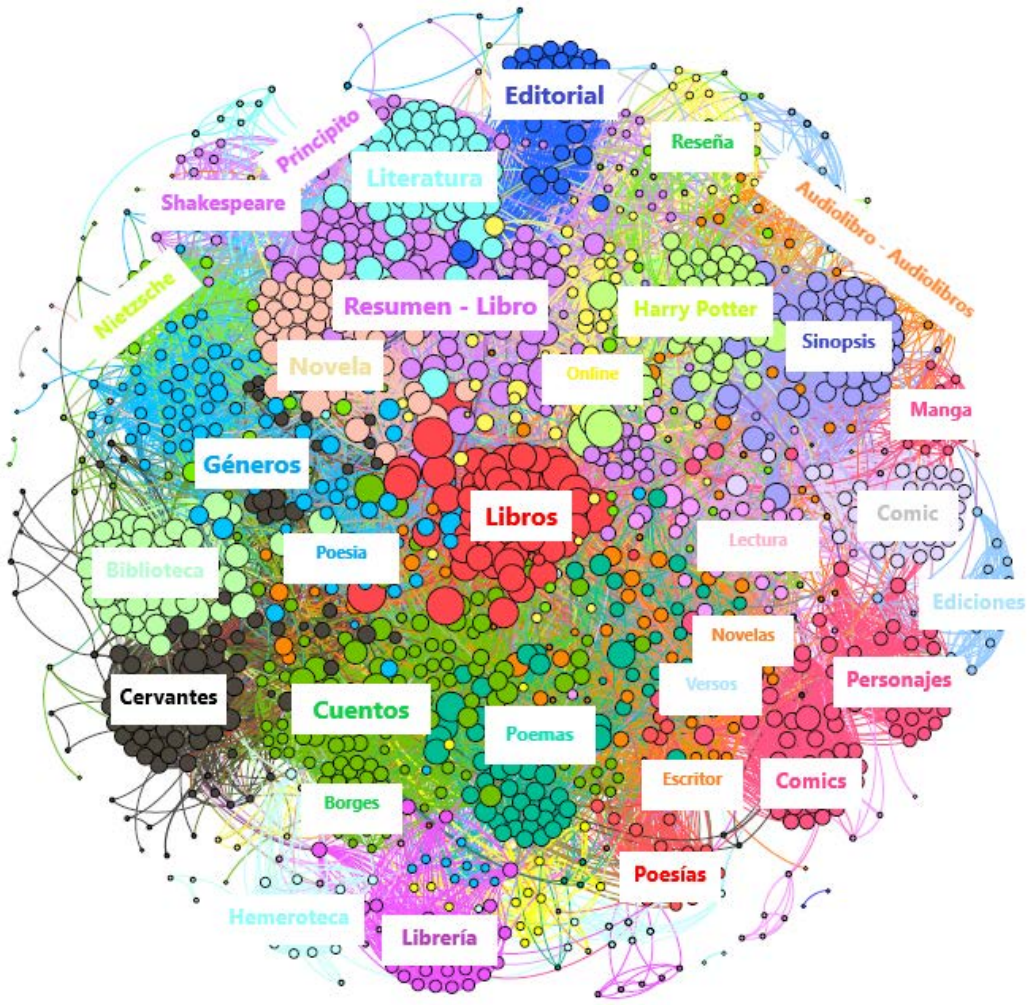


Ilustración 54. Grafo 3: Visualización gráfica con etiquetas de comunidades. Fuente: elaboración propia

5.3.1.1. Comunidad "libros" y Comunidad "libro"/"resumen"

Como se observa en la figura anterior, la comunidad más grande y céntrica, de color rojo, como se podía esperar siendo un conjunto de búsquedas del ámbito de la literatura, corresponde al formado por las expresiones contenedoras de la palabra clave "libros" (**Ilustración 55**).

Esta comunidad está formada por una gran variedad de expresiones de búsqueda, conteniendo palabras clave de casi todas las comunidades encontradas, como se explica a continuación.

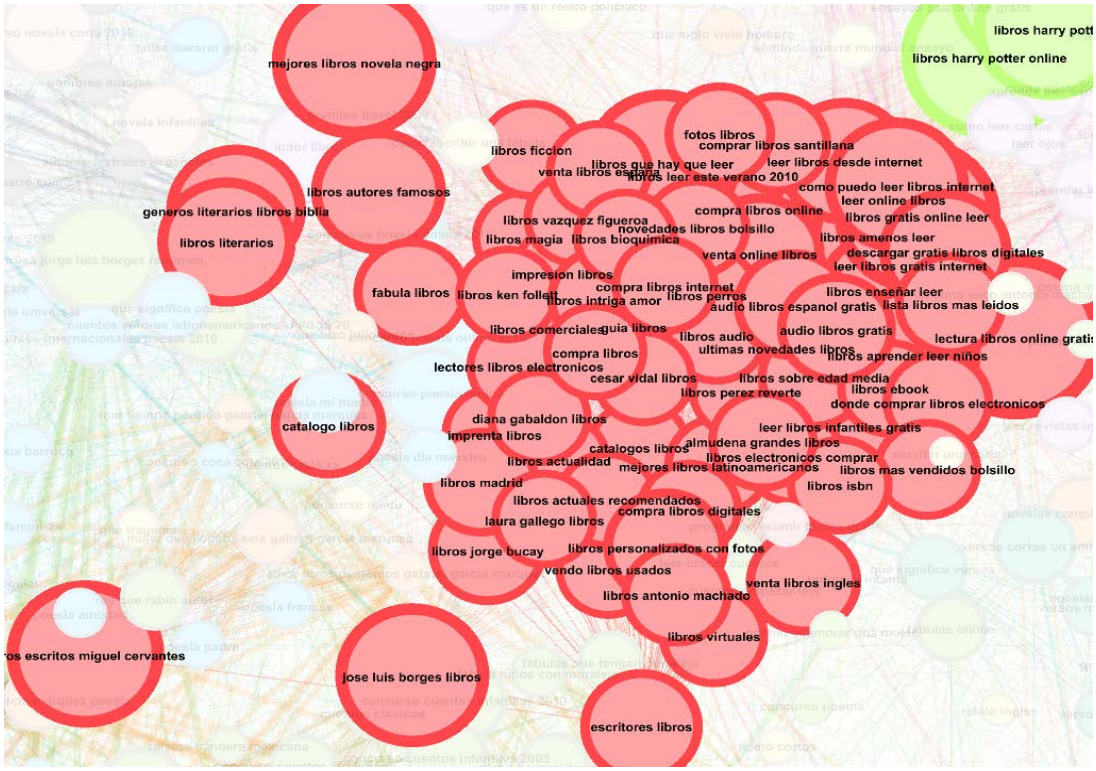


Ilustración 55. Grafo 3: Comunidad "libros". Fuente: elaboración propia

Los nodos más grandes de esta comunidad y por tanto, las expresiones de búsqueda con más conexiones se corresponden con las expresiones “lectura libros online gratis” y “libros harry potter online”. Otras consultas importantes en esta comunidad son: “libros gratis online leer”, “mejores libros novela negra”, “resumen libros”, “leer libros infantiles gratis” y “libros escritos por miguel cervantes”, entre otras. Como se puede ver, las palabras clave que componen estas expresiones a su vez, cada una, forman comunidades enteras de consultas.

Siguiendo con las comunidades más grandes, se observa que otra de ellas es la correspondiente a la palabra clave “libro”, en color lila. Aunque la comunidad anterior se encuentra más concentrada, esta se muestra más esparcida, probablemente por su alta relación con la comunidad de la palabra clave “resumen”, a la que se encuentra prácticamente junta por nodos que contienen cadenas de estas dos palabras “resumen libro” (**Ilustración 56**).

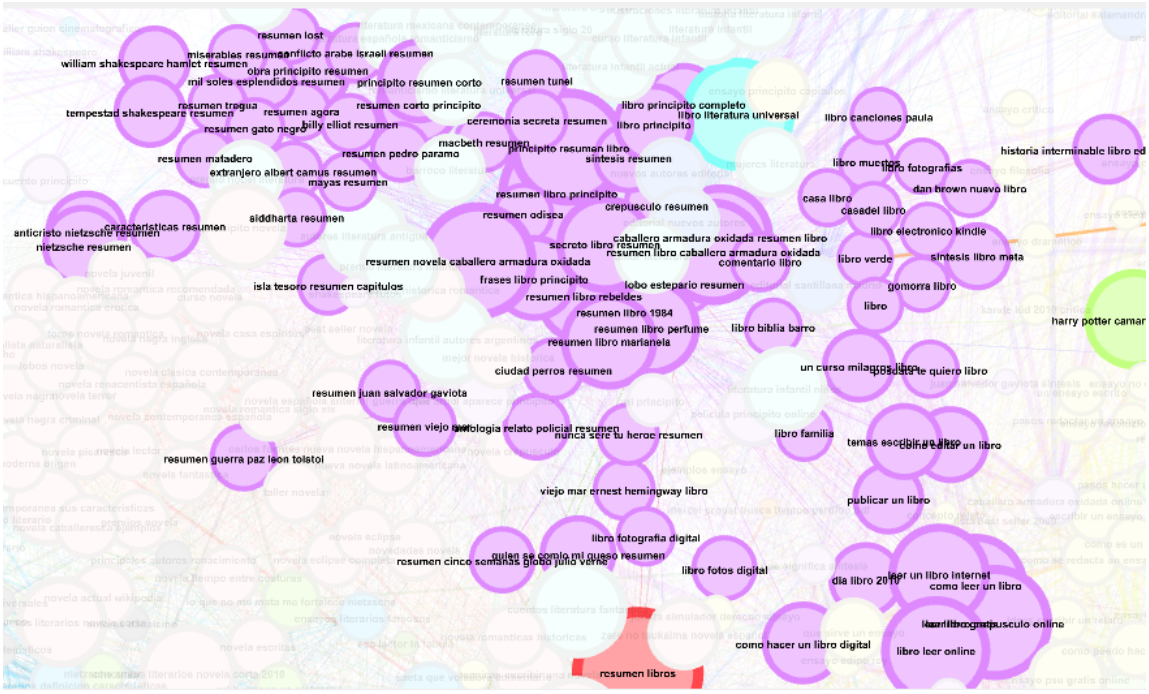


Ilustración 56. Grafo 3: Comunidad "libro"/"resumen". Fuente: elaboración propia

Las expresiones de búsqueda que componen estas dos comunidades suelen caracterizarse por estar formadas por las palabras clave correspondientes a la comunidad sumada títulos de libros o nombres de autores, incluso todo junto. Esto no significa que todos los nodos sean así, ya que, al ser dos de las comunidades más grandes, contienen bastantes aristas a otras diferentes, como por ejemplo la palabra clave “online”, “literatura” o “principito”. Las más importantes detectadas en estos cúmulos han sido:

Para el término “libro”: “resumen libro principito”, “principito resumen libro”, “leer libro crepúsculo online”, “libro literatura universal”, “libro leer online”, “caballero armadura oxidad resumen libro” y “resumen libro caballero armadura oxidada”.

Para el término “resumen”: “resumen libros”, “resumen libro principito”, “principito resumen libro”, “resumen novela caballero armadura oxidada”, “caballero armadura oxidada resumen libro”, “resumen libro caballero armadura oxidada” y “intrusa jorge luis borges resumen”. Consultas más propias del término en sí, se encuentran relacionadas, como se ha mencionado anteriormente, con obras literarias (“perfume”, “1984”, “rebeldes”, “marianela”, “principito”) o con autores (“borges”, “nietzsche”, “shakespeare”). En el grafo también se pueden encontrar otras comunidades muy similares a esta, siendo estas “sinopsis” y “reseña”.

5.3.1.2. Comunidad "novela"

Otra comunidad de las más extensas es la correspondiente al término "novela", en color carne claro (**Ilustración 57**). En esta, las expresiones de búsqueda en su mayoría se componen de la palabra "novela" sumada a algún tipo o género de novela, como por ejemplo: "corta", "histórica", "picaresca", "fantástica", "romántica", "erótica", etc. Este hecho no quita que también se encuentre "novela" en expresiones con títulos de obras, como es el caso de "eclipse", "crepúsculo" o "principito".

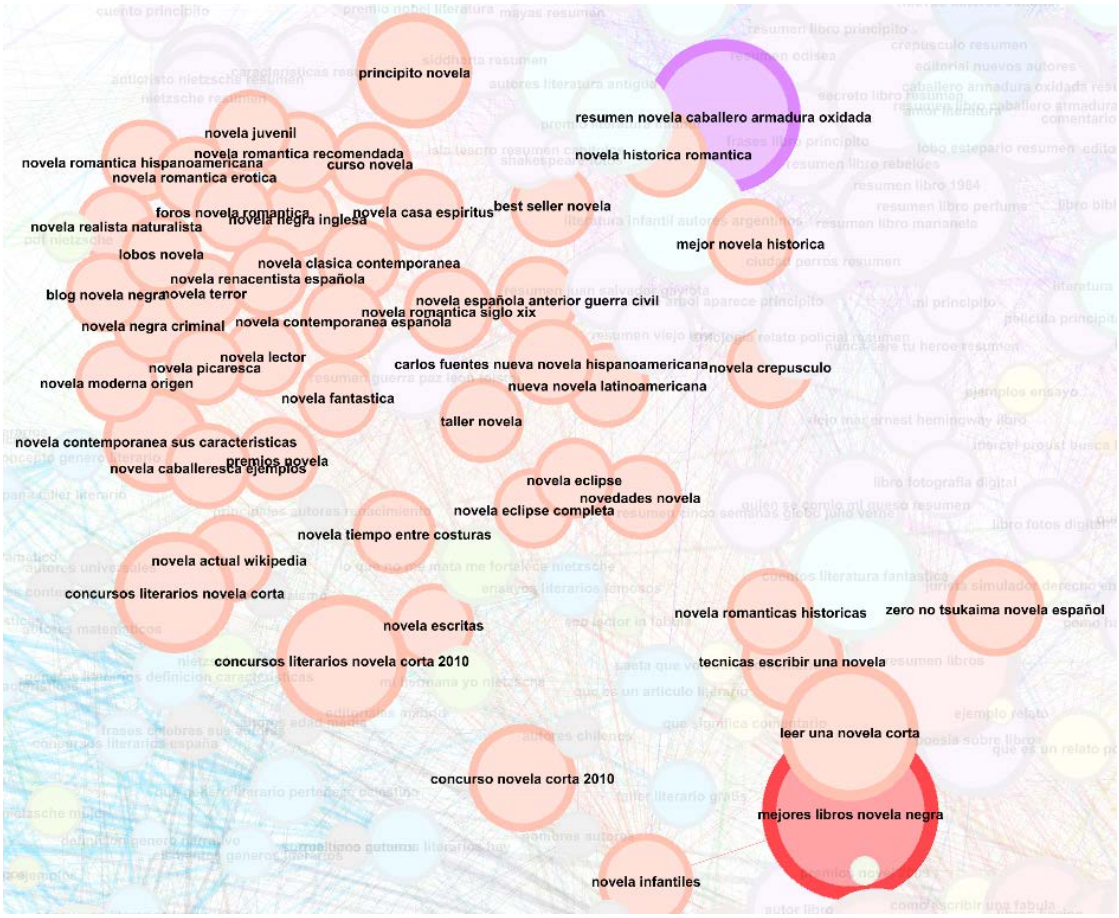


Ilustración 57. Grafo 3: Comunidad "novela". Fuente: elaboración propia

Las expresiones más importantes de esta comunidad son: "mejores libros novela negra", "resumen novela caballero armadura oxidada" o "leer una novela corta".

En el grafo también encontramos la comunidad "novelas", prácticamente igual a la anterior.

5.3.1.3. Comunidad "Literatura"

En esta comunidad (**Ilustración 58**) se encontró una situación algo similar a la anterior. En ella, en color turquesa, observamos que las expresiones se forman juntando también la palabra clave "literatura" con tipologías o géneros, como "juvenil", "postmodernista", "barroca" o "escrita" entre otros, pero, también se observan expresiones con palabras clave referentes a países, como por ejemplo: "inglesa", "chilena", "colombiana", "árabe", "mexicana", etc.

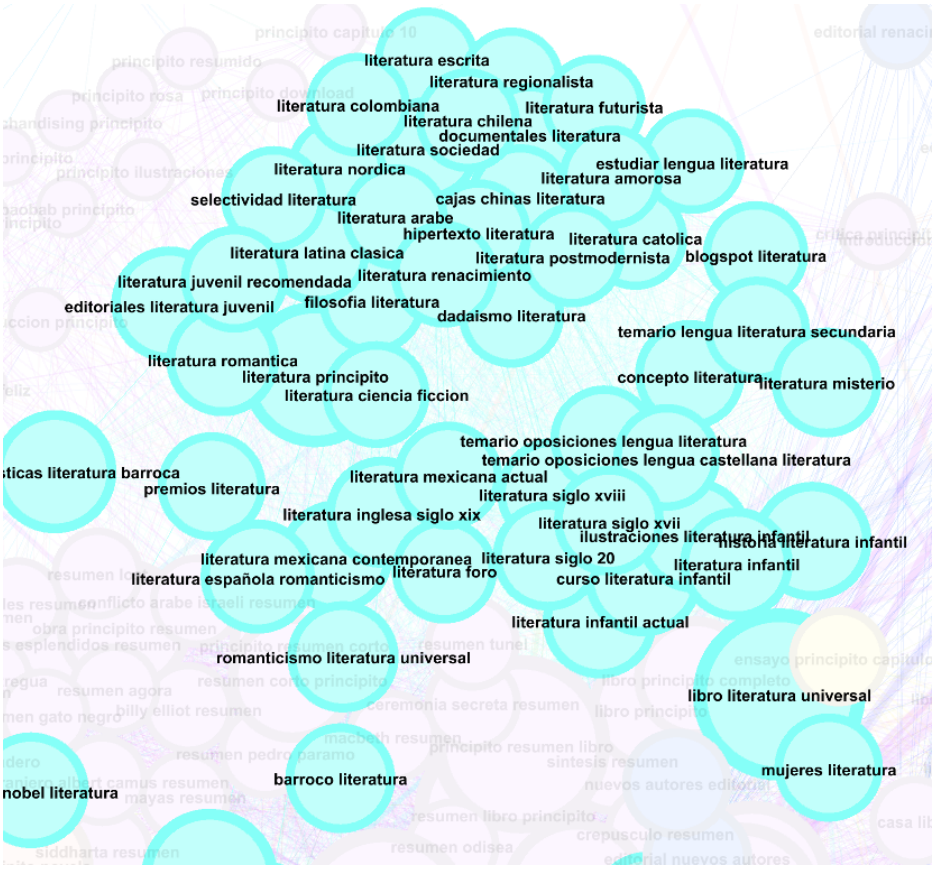


Ilustración 58. Grafo 3: Comunidad "literatura". Fuente: elaboración propia

Otras formas de consulta destacadas en esta comunidad son las formadas por la palabra "literatura" sumada a palabras clave de temática académica: "estudiar", "selectividad" y "oposiciones", entre otras.

Las expresiones más importantes que encontramos en ella fueron: “libro literatura universal”, “cuentos literatura fantástica”, “literatura infantil autores argentinos”, “literatura principito”, “autores literatura antigua” y “literatura infantil niños”.

5.3.1.4. Comunidad “generos”/”genero”/”literarios”/”literario”

Esta comunidad (**Ilustración 59**), coloreada en azul claro, está casi al completo formada por expresiones de búsqueda que implican las palabras clave “género” o “géneros” y “literario” o “literarios” en conjunto.

La característica descrita implica también la aparición de subgrafos, dentro de esta comunidad, que implican alguna de las palabras clave anteriores de forma individual, como por ejemplo “literario” con “texto”.

Ilustración 59. Grafo 3: Comunidad "generos"/"genero"/"literario"/"literarios"

Por estos hechos, esta comunidad en un principio muestra una apariencia bastante sólida, pero con un examen más profundo se observan las dispersiones mencionadas, resultando casi imposible poder visualizarla al completo seleccionando nodos. Las consultas más relevantes encontradas fueron:

Para el término “generos”: “generos literarios libros biblia”, “obras generos literarios”, “que tipos de generos literarios hay” y “generos literarios sus caracteristicas”. Se observó que ninguna consulta con este término se encuentra separada del término “literarios”, pero por su parte, esta última si que se pudo encontrar de forma independiente a “generos” en: “libros literarios”, “concursos literarios novela corta 2010” o “concursos literarios novela corta”, por ejemplo.

Para el término “genero”: “genero literario poesia”, “obras literarias genero épico”, “que genero literario pertenece celestina” o “genero literario sus características”, “ejemplos genero literario narrativo”. El término “literario”, al igual que en el caso anterior con “literarios”, aparece en consultas más variadas de forma independiente a “genero”: “que es un artículo literario”, “obras realismo literario”, “taller literario gratis” y “romanticismo literario caracteristicas”.

5.3.1.5. Comunidad “harry potter”

Comunidad correspondiente a las obras de la saga Harry Potter (**Ilustración 60**), de color verde, dentro de la cual encontramos la palabra clave “harry potter”, indivisible, sumada a los subtítulos de sus libros, a los términos “película” y “descarga” e incluso a subtítulos de sus videojuegos.

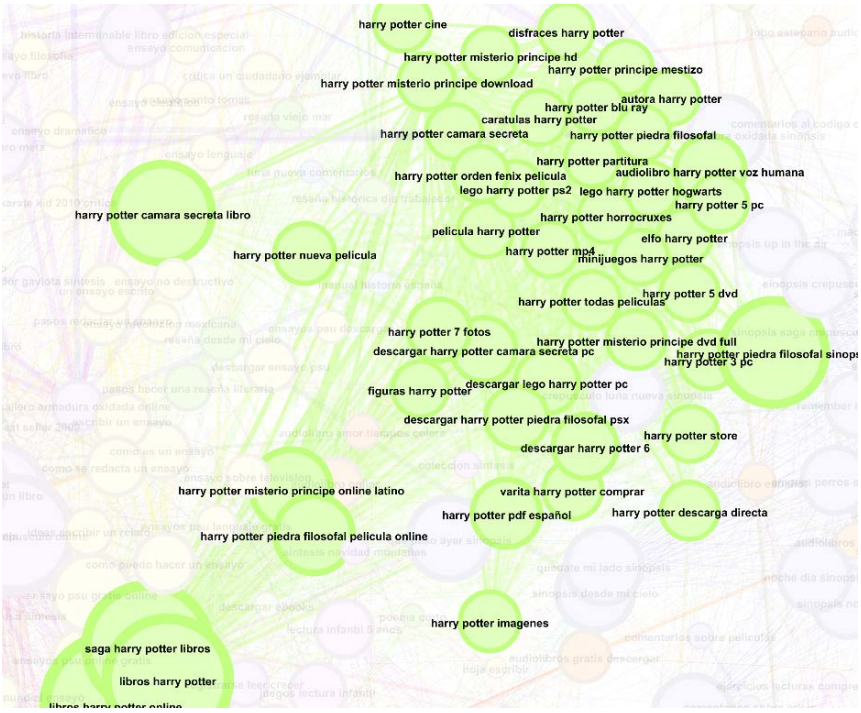


Ilustración 60. Grafo 3: Comunidad "harry potter". Fuente: elaboración propia

Las expresiones de búsqueda más relevantes de esta comunidad fueron: “libros Harry potter online”, “saga harry potter libros”, “libros harry potter”, “Harry potter piedra filosofal sinopsis”, “harry potter camara secreta libro” o “Harry potter misterio príncipe online latino”.

5.3.1.6. Comunidades periféricas destacables

A partir de este punto, se analizaron las comunidades más periféricas más destacables, siendo estas las que se sitúan alrededor del centro del grafo pero lo suficientemente separadas como para formar comunidades bastante compactas.

5.3.1.6.1. Comunidades “biblioteca”, “editorial”, “librería” y “hemeroteca”

La comunidad “biblioteca” (**Ilustración 61**) consiste en una comunidad formada por la palabra clave “biblioteca”, coloreada en verde claro. En ella se encuentra en casi la totalidad de las expresiones la palabra clave “biblioteca” sumada a localizaciones geográficas (“valladolid”, “madrid”, “aragon”) o universidades (“deusto”, “complutense”). Las consultas más importantes de esta comunidad han sido: “biblioteca digital cervantes”, “biblioteca online gratuita”, biblioteca online”, “biblioteca autores cristianos bac”, “biblioteca luis angel Arango catalogo”, “mi biblioteca”, “biblioteca infantil”, “bases de datos una biblioteca”, “biblioteca retiro madrid”, etc.

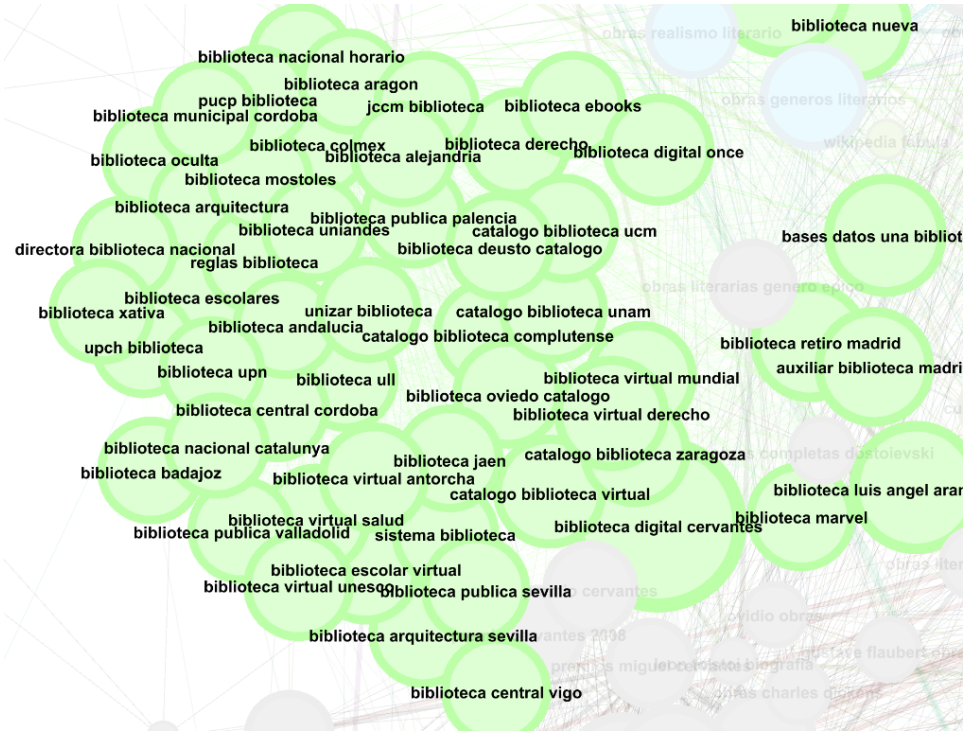


Ilustración 61. Grafo 3: Comunidad "biblioteca". Fuente: elaboración propia

Similar a estas, en cuanto a características, son las comunidades de las palabras clave “editorial” y “librería”, e incluso, la de la palabra clave “hemeroteca” (**Ilustración 62**).

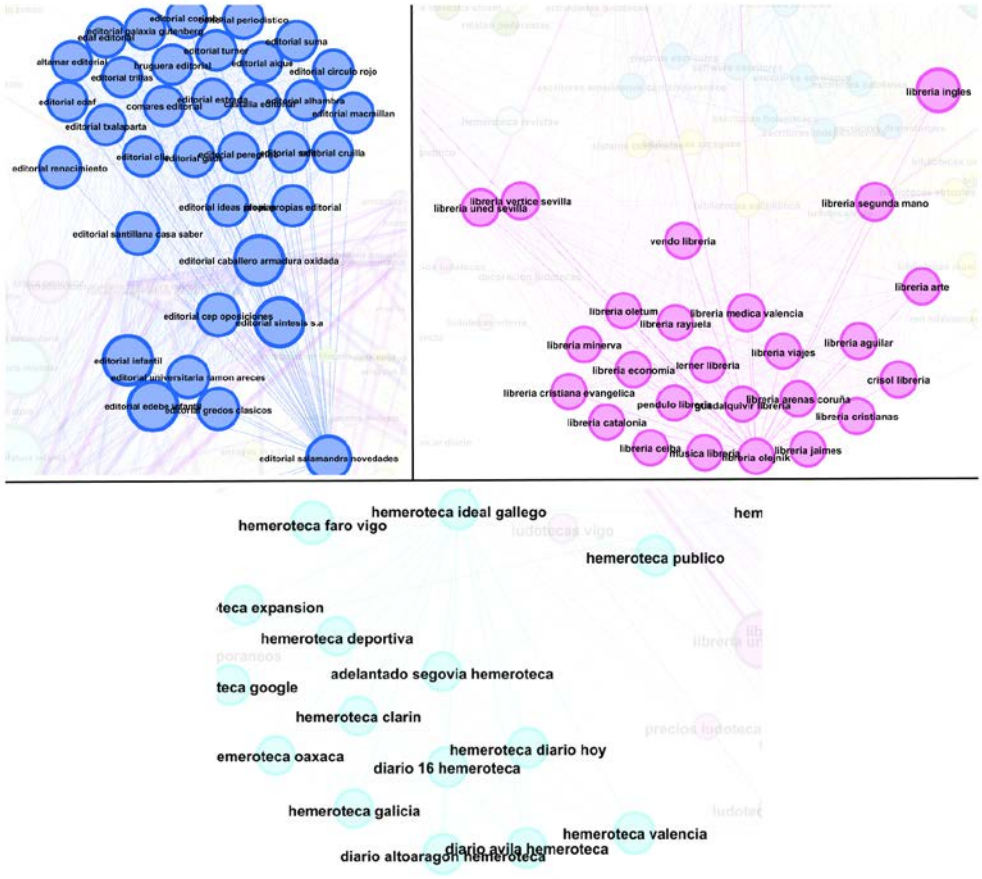


Ilustración 62. Grafo 3: Comunidades "editorial", "librería" y "hemeroteca"

La comunidad “editorial”, en color azul, se compone en su mayoría de conjuntos con nombres de editoriales (“nuevos autores”, “edebe”, “síntesis”, “santillana”, “salamandra”, “gredos”) y las consultas más importantes fueron: “editoriales literatura juvenil”, “editorial nuevos autores”, “nuevos autores editorial”, “editorial caballero armadura oxidada”, “editorial infantil”, “editorial edebe infantil”, etc.

La comunidad “libreria”, en color fucsia, se compone, al igual que la anterior, en su mayoría de conjuntos con nombres de librerías (“uned”, “vertice”, “crisol”, “aguilar”) y las consultas más relevantes fueron: “librería uned madrid”, “uned librería virtual”, “librería ingles”, “librería segunda mano”, “librería viejo”, “librería vertice sevilla”, etc.

La comunidad “hemeroteca”, en color azul claro, básicamente se compone de nombres de hemerotecas o de periódicos, destacando las consultas: “hemeroteca digital abc”, “hemeroteca faro de vigo”, “hemeroteca valencia”, “hemeroteca revistas”, “diario altoaragon hemeroteca” o “hemeroteca publico”, por ejemplo.

5.3.1.6.2. Comunidades “nietzsche”, “shakespeare”, “cervantes” y “borges”

Estas cuatro comunidades (**Ilustración 63**) comparten casi la misma estructura de expresiones de búsqueda, siendo en su mayoría conjuntos entre los nombres de los autores y los títulos de sus obras.

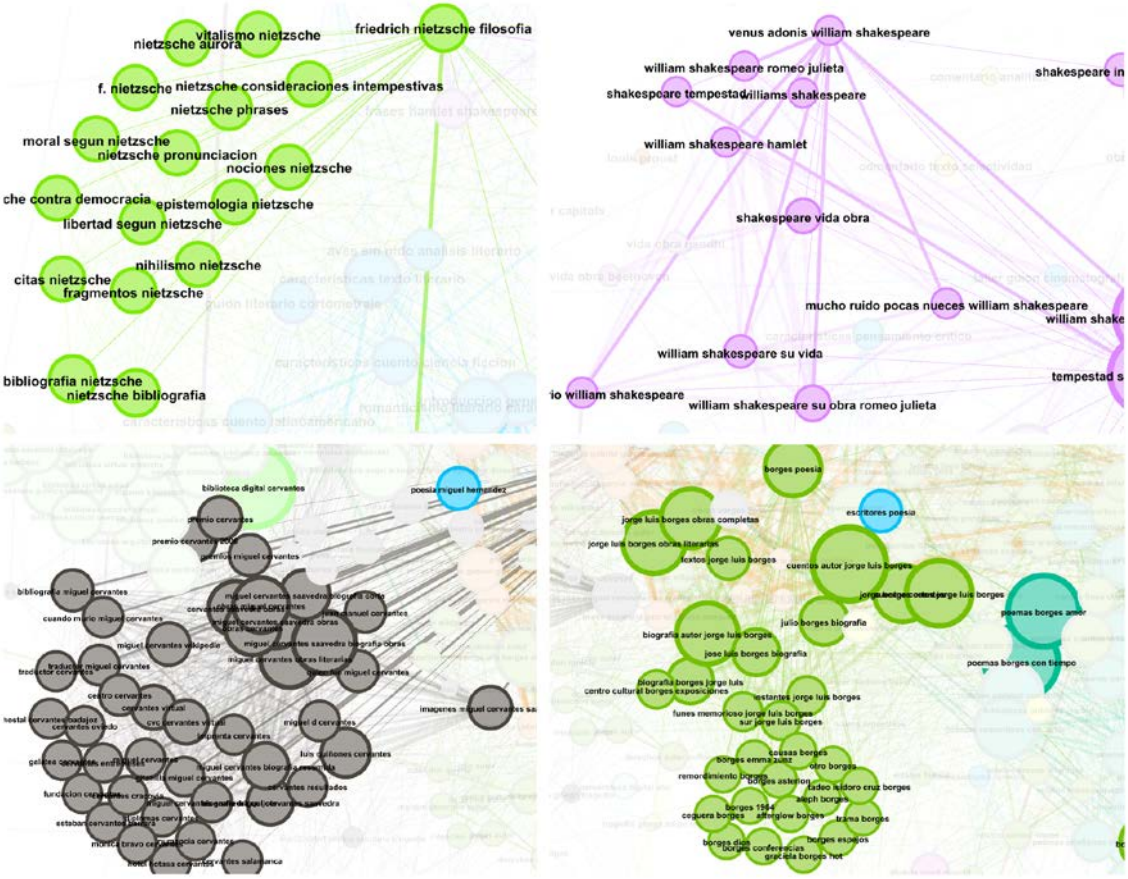


Ilustración 63. Grafo 3: Comunidades "nietzsche", "shakespeare", "cervantes" y "borges". Fuente: elaboración propia

La comunidad “nietzsche”, en color verde, se compone por expresiones sobre el autor Friedrich Nietzsche y su pensamiento y obras referentes a él (“vitalismo”, “nihilismo”, “amor”, “moral”) y sus obras (“consideraciones intempestivas”, “aurora”). Las consultas más relevantes en cuanto a Nietzsche son “anticristo nietzsche resumen” y “nietzsche resumen”, que pertenecen a la comunidad “resumen”, siendo el enlace entre estas dos. En cuanto a las consultas más relevantes dentro del conjunto en sí, se encontraron: “friedrich nietzsche obras”, “lo que no me mata me fortalece

nietzsche", "mi hermana yo nietzsche", "nietzsche amor" y "nietzsche mujer", entre otras.

La comunidad "Shakespeare", en color morado, pertenece al autor William Shakespeare y se compone casi en su totalidad, como se ha dicho, de conjuntos formados por su nombre o apellido y el título de sus obras ("tempestad, "hamlet", "romeo julieta"). Las consultas más relevantes sobre este autor, al igual que con Nietzsche, suponen el nodo con la comunidad de "resumen", perteneciendo a esta, las cuales son "tempestad shakespeare resumen" y "william shakespeare Hamlet resumen". Las consultas más importantes en el conjunto en sí, fueron: "william shakespeare su obra romeo Julieta", "shakespeare fotos", "shakespeare vida obra", "frases hamlet shakespeare", etc.

La comunidad "cervantes", en color gris, correspondiente al autor Miguel de Cervantes se compone de una manera similar a las anteriores, consistiendo sus dos consultas más relevantes ("libros escritos miguel cervantes" y "biblioteca digital cervantes") en los nodos con las comunidades "libros" y "biblioteca", perteneciendo a ellas respectivamente. Las expresiones más relevantes de su comunidad fueron: "miguel cervantes Saavedra biografia obras", "miguel cervantes obras literarias" y "miguel cervantes Saavedra obras", entre otras.

Por último, la comunidad "Borges", en verde también, perteneciente al autor Jorge Luis Borges, sigue casi la misma estructura que las anteriores ("libros", "resumen"), aunque añadiendo nodos en común con la comunidad "poemas" y "cuentos", debido al contenido de su obra. La consulta más importante sobre Borges corresponde a "jose luis borges libros" ya que une su comunidad con la de "libros". Caso curioso el de esta búsqueda, ya que contiene una errata, siendo el nombre del autor Jorge y no José. Las expresiones más importantes fueron: "aleph borges descargar", "textos jorge luis borges", "instantes jorge luis borges" y "sur jorge luis Borges", por ejemplo.

Cabe destacar que, debido al hecho de ser sobre autores estas comunidades, una palabra clave compartida entre las cuatro ha sido "biografia", aunque sin llegar a definir una comunidad propia destacable.

5.3.1.6.3. Comunidades "cuentos", "poemas", "poesia", "poesias" y "versos"

Las comunidades contenidas en este apartado (**Ilustración 64**) contienen estructuras similares, dado que suelen ir todas las palabras clave que las definen acompañadas con palabras clave de autores y de tipologías temáticas.

Ilustración 64. Grafo 3: Comunidades "cuentos", "poemas", "poesia", "poesias" y "versos". Fuente: elaboración propia

En el caso de la comunidad “cuentos”, en color verde, un tanto dispersa, las expresiones de búsqueda se forman junto a autores como “borges”, “garcia marquez”, “vargas llosa” y “flaubert”. También se forman con palabras clave referentes a su tipología o características, como “graciosos”, “literatura fantástica”, “cortos” y “infantiles”. Las consultas más relevantes de esta comunidad resultaron: “cuentos autor jorge luis borges”, “cuentos literatura fantastica”, “cuentos cortos jorge luis borges” y “cuentos escritor gabriel garcia marquez”.

La comunidad “poemas”, en color turquesa oscuro arriba, al igual que la anterior, contiene expresiones junto con autores como “borges”, “vargas llosa”, “mario benedetti” y “pablo neruda”. Pero esta comunidad se caracteriza más por los conjuntos de la palabra clave “poemas” con tipologías o géneros, como por ejemplo: “amistad”, “romanticos”, “infantiles”, “romanticismo”, “barroco”, “mujer” o “muerte”. Las expresiones de búsqueda más importantes de este conjunto fueron las siguientes: “poemas Borges con tiempo”, “poemas Borges amor”, “poemas sobre lectura”, “poemas amistad con autor” y “poemas romanticos con autor”, entre otros.

El conjunto “poesias”, en color rojo, en cuanto a contenido se estructura prácticamente igual que el anterior, teniendo en cuenta que la palabra “poesias” contiene etimológicamente a “poemas”. Las expresiones de búsqueda más relevantes de esta comunidad son: y “borges poesias”, “poesias versos”, “poesias niños cortas”, “poesias agua niños” y “poesias dia padre”. De la comunidad “poesia”, en color azul, poco más de lo anterior se puede decir, solo destacando que, sus expresiones más notables fueron: ““poesia sobre libro”, “concurso poesia online 2010”, “concursos poesia online 2010”, “concurso poesia online” y “borges poesia”.

Para finalizar, la comunidad “versos”, en color turquesa oscuro abajo, se compone en su mayoría de palabras clave relacionadas con el tipo o temática de versos (“romanticos”, “cortos graciosos”, “funebres”, “emos”, “amor”) o para dedicarlos a personas (“amigo”, “suegra”, “esposa”). Los nodos más importantes unen a este conjunto con los conjuntos “poemas” y “poesias” (“poemas versos romanticos”, “poemas 40 versos” y “poesias versos”) y las consultas más importantes en sí fueron: “versos cortos un amigo”, “versos un amigo especial” y “que significa versos”.

5.3.1.6.4. Comunidades “comics”, “personajes”, “comic” y “manga”

En este último grupo de comunidades (**Ilustración 65**) se encuentran las relacionadas con los cómics, el cómic japonés llamado manga, y sus personajes, este último hecho coincide también con que algunos títulos de los que aparecen son también el nombre de sus personajes principales, como “Deadpool” o “Naruto”.

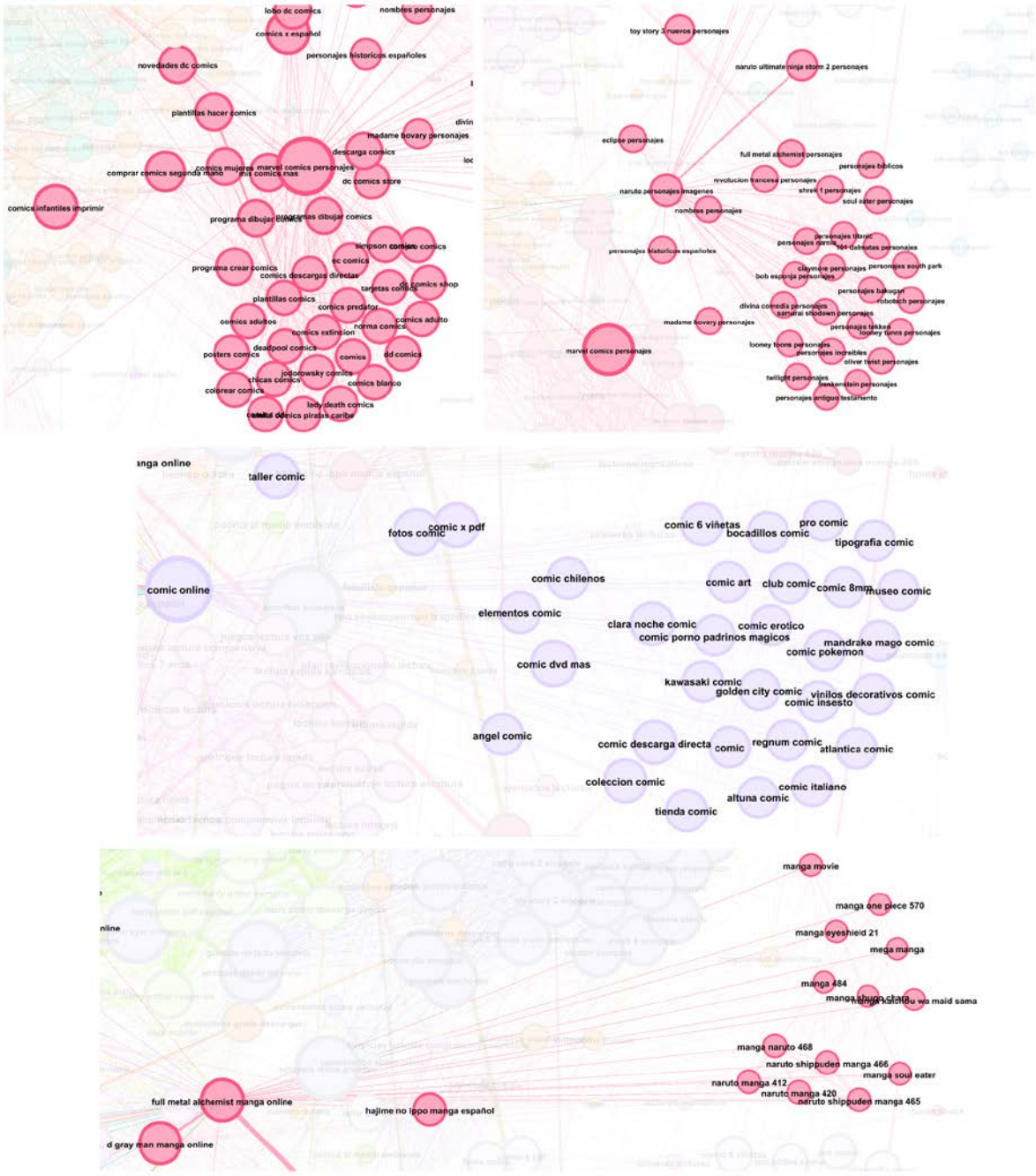


Ilustración 65. Grafo 3: Comunidades "comics", "personajes", "comic" y "manga". Fuente: elaboración propia

Para empezar, la comunidad “comics”, en color fucsia arriba a la izquierda, se compone de consultas constituidas por palabras clave sobre tiendas, empresas o marcas de cómics (“dc”, “marvel”, “norma”) o de series bastante conocidas (“deadpool”, “simpson”). En esta comunidad se encuentran también bastantes expresiones con las palabras “comprar” y “tienda”. Los nodos más importantes de este conjunto son los siguientes: “marvel comics personajes”, “dc comics online”, “comics

infantiles imprimir” y “comics x español”. La comunidad “comic” no se diferencia mucho en cuanto a tipología, solo destacar el detalle que en estas dos comunidades se encuentran la mayoría de búsquedas de índole pornográfica en el grafo.

En la comunidad “personajes”, en color fucsia arriba a la derecha, todas las expresiones de búsqueda están formadas por la palabra clave “personajes” sumada a alguna serie, saga, videojuego, película, anime, libro o manga. Como ejemplos de estas consultas se muestran también las más importantes de la comunidad, siendo las siguientes: “marvel comics personajes”, “naruto personajes imagenes”, “naruto ultimate ninja storm 2 personajes”, “toy story 3” o “shrek 1 personajes”. En esta comunidad también se encuentra alguna búsqueda sobre personajes históricos.

5.3.1.7. Otras comunidades

En este subapartado se van a mencionar comunidades no tan destacadas como las anteriores, o que no llegan a tener una comunidad propia definida, pero que tienen su importancia en el grafo.

La más importante de estas, es la comunidad de la palabra clave “online”, ya que se trata de un término que aparece prácticamente en todas las comunidades presentes en el grafo, por lo que resulta muy dispersa.

En la periferia del grafo, encontramos también pequeñas comunidades como la del término “principito” referente a dicha obra, “audiolibros” referente a dicho formato de libros, “ludoteca” referente a los centros de este tipo.

Otras comunidades de la tipología de la descrita anteriormente de la palabra clave “online”, es la comunidad formada por las consultas con el término “descargar” y las comunidades con los términos “autor” y “escritor.

5.3.2. Nodos más importantes

En este apartado se listan los 20 nodos más importantes que se encontraron en este grafo. Estos nodos contienen las consultas generales que más aristas reciben y, en su contenido, se pueden observar palabras clave de diversas de las comunidades más importantes, ya que llegan a funcionar como enlace entre ellas. Dichos nodos son:

- “lectura libros online gratis”
- “libros harry potter online”
- “libros gratis online leer”
- “mejores libros novela negra”
- “resumen libros”
- “leer libros infantiles gratis”
- “libros escritos miguel cervantes”
- “leer online libros”
- “libros que hay que leer”
- “saga harry potter libros”
- “libros harry potter”
- “jorge luis borges libros”
- “libros leer este verano 2010”
- “libros aprender leer niños”
- “resumen libro principito”
- “principito resumen libro”
- “resumen novela caballero armadura oxidada”
- “como puedo leer libros internet”
- “biblioteca digital cervantes”
- “leer libro crepúsculo online”

6. Conclusiones

6.1. Conclusiones del objetivo principal

Encontrar una metodología de trabajo efectiva sobre conjuntos masivos de datos consistentes en expresiones de búsqueda o palabras clave extraídas de motores de búsqueda, con independencia de su temática. Dado el escaso estado del arte encontrado en cuanto al análisis de redes sociales de palabras clave, se estima necesaria una iniciativa de este tipo.

Por lo que se ha podido observar durante el proceso de investigación y, a partir de los resultados obtenidos, se puede decir que la metodología expuesta en el presente trabajo ha resultado exitosa y podrá ser replicada en el futuro a escala más grande para el manejo de conjuntos masivos de consultas extraídas de motores de búsqueda.

Dicha metodología ha permitido extraer información de relevancia para el posicionamiento SEO en el ámbito de la literatura, como podría ser información sobre los libros o sagas más buscados, que según la muestra estudiada en la presente investigación, entre los años 2004 y 2016, serían las sagas *Harry Potter* de J. K. Rowling y *Crepúsculo* de Stephenie Meyer o el libro *El caballero de la armadura oxidada* de Robert Fisher. También se pudo observar información sobre autores más buscados, en este caso Jorge Luis Borges y Miguel de Cervantes, entre otros.

6.2. Conclusiones de los objetivos específicos

6.2.1. Conclusiones del objetivo específico #1

Analizar de forma básica, de las seleccionadas en un primer momento tras diversas búsquedas de comparativas en cuanto a requisitos y funcionalidades, la mejor herramienta para trabajar en el campo del presente trabajo, concretamente, en la sección de la generación y análisis de grafos, siendo estas: Pajek, SocNetV, Gephi, Neo4j y UCINET.

A partir del análisis explicado en el apartado “”Software” y su subapartado ”Justificación” de la elección sobre las herramientas seleccionadas en principio, tomando en base criterios de funcionamiento correcto, interfaz agradable e intuitiva y compatibilidad de formatos, se ha llegado a la conclusión de que el mejor programa

para la generación y análisis de grafos en la presente investigación ha sido Gephi, por lo que se recomienda su uso en futuros trabajos sobre el ámbito.

6.2.2. Conclusiones del objetivo específico #2

Clasificar las expresiones de la muestra a analizar según los tipos de intención de búsqueda del usuario descritos en la introducción, a recordar:

Informacionales: el usuario busca información.

Navegacionales: el usuario intenta llegar a una página.

Transaccionales: el usuario busca comprar.

En el grafo 3, se ha podido observar que hay múltiples ejemplos de cada uno de estos tipos de expresiones.

La inmensa mayoría de expresiones de búsqueda contenidas en el dataset se corresponden con la categoría de expresiones informacionales, como son búsquedas sobre información de libros, autores, sinopsis, resúmenes, etc.

Cabe destacar, como expresiones navegacionales, las expresiones, por ejemplo, contenidas en la comunidad de la palabra clave “biblioteca”, con las cuáles se parece estar buscando la página web de la biblioteca correspondiente.

En cuanto a búsquedas transaccionales, también se encuentran diversas expresiones. Concretamente, por ejemplo, en la comunidad formada por la palabra clave “comics”, en la que se ven repetidos varias veces los términos “comprar” y “tienda”.

6.2.3. Conclusiones del objetivo específico #3

Examinar la tipología de las palabras (verbos, sustantivos, nombres propios, etc.) que conforman las expresiones analizadas y en qué formas se utilizan con más frecuencia dentro de la muestra.

La gran mayoría de términos contenidos en el grafo se trata de sustantivos, hecho evidente dado que, por definición, los sustantivos definen seres, entidades u objetos.

Se encontraron verbos para indicar acciones tales como “leer”, “descargar”, “comprar”, “estudiar” y “publicar”, entre muchos otros. Cabe destacar que siempre se encontraron en forma de verbos infinitivos, por lo que parece ser una característica bastante extensa en cuanto a las búsquedas en Internet. Junto a verbos, se encontraron también adverbios interrogativos como “como” y pronombres interrogativos como “que” para realizar consultas en forma de preguntas.

En cuanto a nombres propios, se encontraron gran cantidad de ellos y en variedad de formas, es decir, completos, solo apellidos o un solo apellido. Ejemplos de estos son los autores Jorge Luis Borges, Miguel de Cervantes, Friedrich Nietzsche, William Shakespeare, Mario Benedetti, César Vidal, Antonio Machado, Laura Gallego, Ken Follett y Diana Gabaldon o el dibujante Alejandro Jodorowsky, entre muchos otros.

Se encontraron también múltiples adjetivos, como por ejemplo, en las comunidades referentes a tipologías de obras literarias, acompañando a las palabras clave correspondientes para indicar géneros literarios o especificaciones sobre dichas tipologías o sobre características de los textos buscados.

En cuanto a preposiciones, conjunciones y artículos, se eliminaron bastantes dado que, su poco significado propio, provocaba interferencias entre nodos, causando ruido.

6.2.4. Conclusiones del objetivo específico #4

Encontrar algún tipo de relación, o vínculo común, entre los clústeres más grandes aparecidos del análisis de la muestra seleccionada.

Las relaciones principales, dado que hay múltiples, entre las comunidades más importantes, se podrían resumir con la siguiente lista, ya que son representadas en su mayoría por algún nodo de gran tamaño sobre el que convergen dos o más comunidades. Contando con el hecho de que la comunidad más grande es la de “libros”, se va a enfocar la lista en las relaciones directas, es decir, caminos de 1 o 2 aristas de distancia, de esta con el resto de comunidades grandes:

- Comunidad “libros”: tiene fuertes relaciones con:
    - La comunidad “novela”: a través del nodo “mejores libros novela negra”.
    - La comunidad “generos”/”genero”/”literarios”/”literario”: a través del nodo.

- La comunidad “libro”/"resumen”: a través de los nodos “resumen libros” (subcomunidad “resumen”), “leer libros desde internet” y “libro leer online” (subcomunidad “libro”).
- La comunidad “literatura”: a través de los nodos “libros aprender niños” y “literatura infantil niños”.
- Con la comunidad “harry potter”: a través de los nodos “libros harry potter online”, “saga harry potter libros” y “libros harry potter online”.

Como se puede observar en la estructura de las consultas mencionadas, resulta prácticamente imposible, con los medios disponibles, llevar a cabo una clasificación concreta de las relaciones entre las comunidades, dada la gran cantidad y variedad de estructuras que se encontraron en las expresiones de búsqueda.

7. Futuras investigaciones

7.1. Propuesta #1

Puesto que la metodología utilizada en este trabajo ha dado lugar a resultados satisfactorios en el conjunto de datos seleccionado, sería interesante ampliarla y pulirla aplicándola a nuevos conjuntos de datos.

Para ello se podría probar a utilizar esta metodología sobre conjuntos de datos que difieran del conjunto utilizado en distintos aspectos como por ejemplo el tamaño, la temática o incluso el idioma. De este modo sería posible definir cada vez mejor el alcance, los puntos fuertes y los puntos débiles de la metodología, así como proponer mejoras que permitan en un futuro poder llegar a aplicarla sobre cualquier conjunto de datos, independientemente de sus características.

7.2. Propuesta #2

Una propuesta de investigación a partir de los resultados obtenidos en el presente trabajo sería la siguiente: investigar qué tipo de páginas web (posicionamiento SEO) responden a las expresiones o conjuntos más utilizadas de los clústeres más extensos. Es decir, introducir dichas expresiones de búsqueda en el motor Google y, con algún programa especializado en la extracción de información de webs, analizar las páginas mejor posicionadas en la página SERP generada.

Un programa de este tipo podría ser, por ejemplo, la suite GATE (*General Architecture for Text Engineering*), bastante extendida y conocida en el campo del Procesamiento de lenguajes naturales (PLN).

7.3. Propuesta #3

En caso de extraer conjuntos que no encajen en los tipos de expresiones de búsqueda explicados en el presente trabajo (ver **1. Introducción**; ver **2.2.2. Objetivo específico**

#2; ver **6.2.2. Conclusiones del objetivo específico #2**), definir posibles nuevas categorías para la descripción de estos.

8. Bibliografía y referencias

ALONSO, R. (5 septiembre, 2020). Guía SEO en Google para principiantes [Artículo en web]. Recuperado 11 de septiembre de 2020, de https://miposicionamientoweb.es/guia-seo-para-principiantes/#-que-es-el-seo-o-posicionamiento-en-buscadores

ALONSO MORENO, Á, GONZÁLEZ HERNÁNDEZ, O. M. Y LAVERA ULLOA, I. (2012). *Krowface: Interfaz de simulación de redes sociales* (Trabajo de curso, Universidad Complutense de Madrid). Recuperado 22 de agosto de 2020, de https://eprints.ucm.es/16105/

BAEZA-YATES, R. Y RIBEIRO-NETO, B. (1999). *Modern Information Retrieval*. Reading (Massachussets): Addison-Wesley.

BASTIAN M., HEYMANN, S. Y JACOMY, M. (2009). Gephi: an open source software for exploring and manipulating networks. En International AAAI Conference on Weblogs and Social Media. Recuperado de https://gephi.org/

BATAGELJ, V. Y MRVAR, A. (1996). Pajek – Program for Large Network Analysis (5.09) [Software]. Recuperado de http://mrvar.fdv.uni-lj.si/pajek/

BENCKENDORFF, P. (2009). Themes and Trends in Australian and New Zealand Tourism Research: A Social Network Analysis of Citations in Two Leading Journals (1994-2007). *Journal of Hospitality and Tourism Management*, 16(1), pp 1-15. doi: 10.1375/jhtm.16.1.1

BOCCARDO BOSONI, G. Y RUIZ BRUZZONE, F. (2019). *RStudio para Estadística Descriptiva en Ciencias Sociales*. Santiago: Departamento de Sociología, Facultad de Ciencias Sociales, Universidad de Chile. Recuperado 29 de agosto de 2020, de https://bookdown.org/gboccardo/manual-ED-UCH/

BORDIGNON, F. R. A. Y TOLOSA CHACÓN, G. H. (2007). Recuperación de información: un área de investigación en crecimiento. *Ciencias de la información*, 38(1-2), pp. 13-24.

BORGATTI, S. P., EVERETT, M. G. Y FREEMAN, L. C. (2002). Ucinet for Windows: Software for Social Network Analysis (6.714) [Software]. Harvard, MA:

Analytic Technologies. Recuperado de https://sites.google.com/site/ucinetsoftware/home

BRITISH LIBRARY Y JOINT INFORMATION SYSTEMS COMMITTEE [JISC]. (2008). Informe Ciber. Comportamiento informacional del investigador del futuro. *Anales de documentación*, 11, pp. 235-258. Recuperado de https://revistas.um.es/analesdoc/article/view/24921

CARRERAS LARIO, R. (2012). *Cómo clasifica Google los resultados de las búsquedas: factores de posicionamiento orgánico* [Tesis doctoral, Universidad Complutense de Madrid]. Recuperado 10 de Julio de 2020, de https://eprints.ucm.es/17450/

CHAFFEY, D. Y ELLIS-CHADWICK, F. (2014). *Marketing digital: estrategia, implementación y práctica.* México: Pearson Educación.

CHEN, X., CHEN, J., WU, D., XIE, Y. Y LI, J. (2016). Mapping the Research Trends by Co-word Analysis Based on Keywords from Funded Project. *Procedia Computer Science*, 91, pp. 547-555. doi: 10.1016/j.procs.2016.07.140

CHIRINOS, N. (2010). *Teoría de grafos. Conceptos básicos* [Diapositivas de PowerPoint]. Recuperado 12 de septiembre de 2020, de https://www.slideshare.net/naborchirinos/conceptos-teoria-de-grafos-5778778

COMBARIZA, G. (2003). Una introducción a la teoría de grafos. En *Memorias XIV Encuentro de Geometría y II encuentro de Aritmética* (pp. 565-591). Bogotá, Colombia: Universidad Pedagógica Nacional. Recuperado de http://funes.uniandes.edu.co/6102/

CROFT, W. B. (1987). Approaches to intelligent information retrieval. *Information Processing & Management*, 23(4), pp. 249-254. doi: 10.1016/0306-4573(87)90016-1

DAURIA, F. (2014). *Una introducción a la teoría de grafos* [Diapositivas de PowerPoint]. Recuperado 12 de septiembre de 2020, de https://www.slideshare.net/federicodau/introduccin-a-lateoradegrafos2014paraimprimir-2

DEAN, B. (22 enero, 2020). Google's 200 Ranking Factors: The Complete List (2020) [Artículo en web]. Recuperado 11 de septiembre de 2020, de https://backlinko.com/google-ranking-factors

FRUCHTERMAN, T. M. Y REINGOLD, E. M. (1991). Graph Drawing by Force-Directed Placement. *Software: Practice and Experience*¸ 21(11).

GRANDJEAN, M. (14 octubre, 2015). GEPHI – Introduction to network analysis and visualization [Artículo en web]. Recuperado 2 de agosto de 2020, de http://www.martingrandjean.ch/gephi-introduction/

INJANTE, R. Y MAURICIO, D. (2020). Método para recomendar factores de posicionamiento personalizados en el motor de búsqueda de Google. *Revista española de Documentación Científica*, vol. 43(1), e253. doi: 10.3989/redc.2020.1.1628

JACOMY, M. (2011). ForceAtlas2, the new version of our home-brew Layout [Artículo en web]. Recuperado 12 de septiembre de 2020, de https://gephi.wordpress.com/2011/06/06/forceatlas2-the-new-version-of-our-home-brew-layout/

KALAMARAS, D. (2015). Social Network Visualizer (SocNetV). Social network analysis and visualization software (2.5) [Software]. Recuperado de https://socnetv.org/

KHO, J., CHO, K. Y CHO, Y. (2013). A study on recent research trend in management of technology using keywords network analysis. *Journal of Intelligence and Information Systems*, 19(2), pp. 101-123. doi: 10.13088/jiis.2013.19.2.101

KOTLER, P. Y ARMSTRONG, G. (2013). *Fundamentos de marketing*. México: Pearson Educación.

KORFHAGE, R. R. (1997). *Information Storage and Retrieval*. Nueva York: Wiley Computer Publishing.

LEDFORD, J. L. (2008). *SEO Search Engine Optimization Bible*. Indianapolis: Wiley Publishing, Inc.

LYONS, J. (1991). *Natural language and universal grammar: essays in linguistic theory*. Nueva York: Cambridge University Press.

MEDINA HERNÁNDEZ, A. I. (2012, mayo 23). *Competencias de Información: Definir una necesidad de información* [Diapositivas de PowerPoint]. Recuperado 9 de julio de 2020, de https://www.slideshare.net/crevirtualnuco/definir-una-necesidad-de-informacin-13046747

MEMBIELA-POLLÁN, M. E. Y PEDREIRA-FERNÁNDEZ, N. (2019). Herramientas de Marketing digital y competencias: una aproximación al estado de la cuestión. *Atlantic Review of Economics*, 3(3). Recuperado de http://www.aroec.org/ojs/index.php/ARoEc/article/view/99

MICELI, J. E. (2008). Los problemas de validez en el análisis de redes sociales: algunas reflexiones integradoras. *Redes. Revista Hispana para el Análisis de Redes Sociales*, 14(1), pp. 1-45. doi: 10.5565/rev/redes.117

MIGUEL, S. E., CAPRILE, L. Y JORQUERA VIDAL, I. (2008). Análisis de co-términos y de redes sociales para la generación de mapas temáticos. *El Profesional de la Información*, 17(6), pp. 637-646. doi: 10.3145/epi.2008.nov.06

MILLET LUACES, R., BEYRIS BRINGUEZ, M. I. Y ROSALES ALMAGUER, M. A. (2012). Colonia de hormigas aplicada a la teoría de grafos. *Acta Latinoamericana de Matemática Educativa*, pp. 545-552. Recuperado de

MOLINA, J. L. (2001). El análisis de redes sociales. Aplicaciones al estudio de la cultura en las organizaciones. *Athenea Digital*, 0. Recuperado de https://atheneadigital.net/article/view/n0-molina/15-html-es#.X2EGm_Q6mtM.link

MONTES, I. (s. f.). *Tema 5: Introducción a la Teoría de Grafos* [Diapositivas de PowerPoint]. Recuperado 12 de septiembre de 2020, de http://ocw.uniovi.es/pluginfile.php/6029/mod_resource/content/0/Tema%201%20-%20Imprimir.pdf

MORATO, J., SÁNCHEZ-CUADRADO, S., MORENO, V. Y MOREIRO, J. A. (2012). Evolución de los factores de posicionamiento web y adaptación de las herramientas de optimización. *Revista Española de Documentación Científica*, 36(3), e018. doi: 10.3989/redc.2013.3.956

MORENO PILA, D. (2017). *Análisis y desarrollo de un modelo para el diagnóstico del posicionamiento SEO* [Trabajo de final de grado, Universidad de Cantabria]. Recuperado de http://hdl.handle.net/10902/13412

NEO4J ENGINEERING. (2010). Neo4j Desktop (1.3.4) [Software]. San Mateo, CA: Neo4j, Inc. Recuperado de https://neo4j.com/download

OLARU, V. (2019). *Análisis del comportamiento de búsqueda sobre información de autores literarios en Google, por parte de usuarios de España y Estados Unidos* [Trabajo de final de máster, Universitat Politècnica de València]. Recuperado de http://hdl.handle.net/10251/128928

OLIVIER PERALTA, E. (20 abril, 2020). Google Keyword Planner: qué es y cómo usarlo mejor que nadie [Artículo en web]. Recuperado 19 de agosto de 2020, de https://es.semrush.com/blog/google-keyword-planner-que-es/

R-TOOLS TECHNOLOGY INC. (2000-2020). RStudio Help – Requisitos del sistema. Recuperado 28 de agosto de 2020, de https://www.r-studio.com/es/Unformat_Help/systemrequirements.html

R CORE TEAM. (2017). R: A language and environment for statistical computing (4.0.2) [Software]. Vienna, Austria: R Foundation for Statistical Computing. Recuperado de https://www.R-project.org

ROMERO, D. (27 junio, 2015). SERPs: ¿qué son y cómo funconan? [Artículo en web]. Recuperado 12 de septiembre de 2020, de https://www.inboundcycle.com/blog-de-inbound-marketing/que-son-las-serps

RSTUDIO TEAM. (2020). RStudio: Integrated Development for R (1.3.1056) [Software]. Boston, MA: RStudio, PBC. Recuperado de https://rstudio.com/

RUSSELL, J. M., MADERA JARAMILLO, M. J. Y AINSWORTH, S. (2009). El análisis de redes en el estudio de la colaboración científica. *Redes. Revista Hispana para el Análisis de Redes Sociales*, 17, pp. 39-47. Recuperado de http://www.redalyc.org/articulo.oa?id=93112847002

SABATÉ GARRIGA, F., BERBEGAL MIRABENT, J. CONSOLACIÓN, C. Y CAÑABATE CARMONA, A. (2009). La utilización de estrategias SEO en el sector de la venta de libros. *Intangible Capital*, 5(3), pp. 321-346. doi: 10.3926/ic.2009.v5n3.p321-346

SERRANO-COBOS, J. (2019). *Hábitos de recuperación de información en motores de búsqueda sobre lectura, libro y bibliotecas en España (2004-2016)* (Tesis doctoral, Universidad de Zaragoza). Recuperado 5 de febrero de 2020, de https://zaguan.unizar.es/record/83967/files/TESIS-2019-142.pdf

SILVA, E. E. DA, SANTOS, A. A. DOS, PINHEIRO DA SILVEIRA, M. A. Y REIS MOURAO, P. J. (2020). Financial efficiency, actors and interactions: A study of the player flow between clubs and the Sao Paulo semifinalists teams in 2017. *InternexT: Revista Eletronica de Negocios Internacionais da ESPM*, 15(1), p. 88. Recuperado de https://go.gale.com/ps/anonymous?id=GALE%7CA619740375&sid=googleScholar&v=2.1&it=r&linkaccess=abs&issn=19804865&p=IFME&sw=w

SPANOU, K. Y BEKIARI, A. (2020). Analyzing Social Network of Destructive Behaviours in Universities. *International Journal of Sociology and Education*, 9(1), pp. 60-92. doi: 10.17583/rise.2020.4642

STOX, P. (16 marzo, 2017). Easy visualizations of PageRank and Page Groups with Gephi [Artículo en web]. Recuperado 13 de septiembre de 2020, de https://searchengineland.com/easy-visualizations-pagerank-page-groups-gephi-265716#.WMt9Rau60Z4.twitter

TOFFLER, A. (1973). *El "Shock" del Futuro*. Barcelona: Plaza & Janés.

VILARES FERRO, J. (2005). *Aplicaciones del procesamiento del lenguaje natural en la recuperación de información en español* (Tesis doctoral, Universidade da Coruña). Recuperado 9 de julio de 2020, de http://hdl.handle.net/10045/1258

VILLACAÑAS VELASCO, V. (2014). *Análisis y Comparativas de Herramientas de Búsqueda y Predicción de Información en Redes Sociales* (Trabajo de final de grado, Universidad de Alcalá). Recuperado 10 de septiembre de 2020, de http://hdl.handle.net/10017/20782

VIÑAS, M. (16 diciembre, 2015). ¿Es Google siempre el mejor motor de búsqueda? Una comparativa con Bing, DuckDuckGo y Twitter [Artículo en web]. Recuperado 11 de septiembre de 2020, de https://www.totemguard.com/aulatotem/2015/12/es-google-siempre-el-mejor-motor-de-busqueda-una-comparativa-con-bing-duckduckgo-y-twitter/

WITTGENSTEIN, L. (1988). *Investigaciones filosóficas*. Barcelona: Editorial Crítica.