TFM

Marçal Mora-Cantallops[1], Roberto Santamaría Ayuso[1], Eva Muñoz Herráiz[1], and salvador.sanchez[1]

[1]Affiliation not available

October 7, 2018

(OK)

1. Introducción y contexto

1.1. Introducción

La llegada de la llamada Web 2.0 convirtió Internet en una plataforma que no tan solo permite sino que potencia la creación de contenido por parte de sus usuarios. Los consumidores de información pasaron a ser también productores (un comportamiento a menudo reconocido con el término inglés *prosumers)* (Toffler, 1980) y, con ello, llegó la explosión de los medios de información no profesionales, tales como blogs y foros, cambiando la forma en la que las personas se comunican, ya sea para conversaciones privadas o para discusiones abiertas. Los foros en línea, en particular, forman una estructura que facilita las interacciones intensivas entre sus participantes (Holtz et al., 2012).

Estas comunidades basadas en la Web se han convertido en sitios importantes en los que la gente busca y comparte su experiencia, habla y trabaja con personas de intereses similares o discute y debate sobre temas de lo más diverso. Estas comunidades suelen diferir, en estructura y forma, de otras redes que se forman en Internet, como las desarrolladas en plataformas sociales (Facebook, Twitter, etc.). Y es que las comunidades que forman los foros presentan una estructura más reglada, con usuarios de distinto prestigio (desde los administradores hasta los invitados), con temas y subtemas, con interacciones públicas (en los hilos) y privadas (mensajes directos), así como una organización más clara en su construcción y flujo de creación de temas y respuestas a los mismos.

La naturaleza dinámica de los foros hace que, además, crezcan y se moldeen a lo largo del tiempo; para entender las relaciones entre usuarios y temas no es suficiente con analizar un punto concreto en el tiempo,

una fotografía. En la mayor parte de casos es necesario visualizar la película; en foros esto se traduce en la identificación de usuarios y temas más populares o relevantes en cada intervalo temporal.

Múltiples estudios han aprovechado información extraída de foros online como material de análisis. Abdulla (2007), por ejemplo, analizó el contenido de tres de los foros online más populares en lengua árabe sobre los ataques del 11-S y llegó a la conclusión que la mayoría de sus usuarios rechazaban el Islamismo como justificación para el atentado. Por su lado, Copes and Williams (2007) estudiaron la intersección entre el comportamiento anormal y la subcultura a través de un foro dedicado a los *straightedge*, jóvenes que deciden abstenerse de beber alcohol, fumar tabaco y consumir drogas de por vida. La violencia doméstica fue el objeto de estudio de de Vries and Valadez (2008), que concluyeron que "el anonimato percibido en los foros de Internet parece ofrecer un sitio seguro a los supervivientes para construir las narrativas de sus experiencias con las relaciones abusivas de forma colaborativa". Galasińska (2010) se fijó en los discursos de los migrantes (y no migrantes) polacos para explicar cómo la transformación del país estaba haciendo que muchos de los emigrados pensaran en volver a Polonia. Las discusiones en foros sobre la detención del dictador chileno Augusto Pinochet centraron el trabajo de Tanner (2001), mientras Holtz and Wagner (2009) investigaron el discurso racista y antisemita de varios foros de extrema derecha. La medicina y la salud también han aprovechado el contenido online de los foros para investigar, como por ejemplo en el caso de Sneijder and te Molder (2004), que se fijaron en los usuarios de un foro de veganismo para estudiar cómo orientaban la relación entre la elección de dieta, salud y responsabilidad. No obstante, la extracción de material de estas comunidades para investigación todavía puede ser considerado como un campo emergente (Skitka and Sargis, 2006).

Algunos estudios existentes presentan y proponen distintas formas de analizar las comunidades de foros en función de su tipo; por ejemplo, algunos se centran en las comunidades de ayuda (como podría ser la resolución de problemas informáticos) (Zhang et al., 2007), otros en las redes de terrorismo en la "dark web" (para detectar individuos relevantes) (Zhang et al., 2009) y otros en comunidades educativas (participación de los estudiantes) (Suraj and Roshni, 2015). Pero la falta de un marco común de trabajo hace básicamente imposible la comparación de resultados entre ellas, la caracterización de las similitudes y diferencias entre las diversas comunidades existentes.

Por otro lado, las dificultades técnicas, que dificultan el trabajo de análisis de foros que sobrepasan cierto umbral de temas, usuarios o posts. Algunos algoritmos presentan problemas en su aplicación a nivel computacional (la búsqueda de camarillas o cliques, por ejemplo, o el aplanado de la red bipartida) pero otros

también lo hacen a nivel conceptual. Especialmente relevante es este hecho en los algoritmos de detección de comunidades, a los que se suele dar tanta importancia en estos entornos, y que raramente proporcionan resultados estables en su ejecución (Kwak et al., 2009).

En resumen, el análisis en profundidad de la literatura revela los dos principales problemas a los que se enfrenta el campo:

- Ausencia de una metodología unificada que permita comparar estudios de distintas comunidades.
- Dificultades técnicas que presentan las redes masivas en aspectos como la extracción pero especialmente en lo referente al procesamiento y análisis de la información.

De la necesidad de dar respuesta a estos dos puntos surge el presente trabajo.

1.2. Foros online

Un foro online se puede definir como un sitio de discusión en línea en el que sus usuarios pueden conversar de forma asíncrona mediante mensajes escritos. Se diferencia del chat y otras formas de conversación online por su estructura (que suele ser jerárquica en forma de árbol), por su extensión (que, por regla general, es mayor) y por la conservación del historial de las conversaciones. La jerarquía no se encuentra solamente en la estructura de los árboles de discusión, sino que los propios usuarios suelen estar identificados por su nivel dentro del sitio, desde administradores o fundadores hasta nuevos usuarios o, incluso, invitados.

Como cualquier sistema de comunicación, los foros tienen su propia nomenclatura para referirse a sus elementos. Sus conversaciones, por ejemplo, son habitualmente conocidas como "hilos" (*threads*) y cada uno de sus mensajes es una "entrada" (*post*). En general, además, cada comunidad suele definir sus propias reglas de comunicación, que van desde la forma de escribir hasta los comportamientos no deseados y penalizados por los administradores del sitio. Habitualmente, distintos temas se tratan en distintas secciones temáticas; los usuarios pueden iniciar una discusión - un hilo - en cualquiera de esas secciones o sub-secciones. El resto de usuarios pueden responder al primer mensaje o a respuestas de otros usuarios. No hay una regla única, pero mientras que en general los foros están accesibles públicamente para cualquier usuario de Internet,

para iniciar un hilo suele ser necesario estar registrado. Eso no quita, sin embargo, que existan multitud de comunidades privadas en las que no se puede participar sin previo registro (y, a veces, incluso verificación).

A diferencia de otras plataformas sociales, los foros son tan antiguos como Internet; sus orígenes se encuentran en los sistemas de tableros de discusión (*Bulletin Board System*) que se popularizaron en las primeras redes de la década de los 80. Analizarlos es interesante desde el punto de vista de las comunidades virtuales y de interés; si bien algunos sitios pueden estar soportados por gente que se conoce o por organizaciones oficiales, la mayoría son usados por individuos cuyo único nexo es la temática. Esa concentración de usuarios e intereses provoca que, en los mismos, la discusión esté muy centrada y, además, une en un sitio relativamente pequeño a un gran número de integrantes del grupo de interés. Así, los foros facilitan el análisis de los discursos propios de dichas comunidades.

Holtz et al. (2012) listan algunas de las motivaciones que se esconden tras la elección de los foros como objeto de estudio:

- Material abundante, en todas las dimensiones. En primer lugar, por la facilidad de encontrar foros dedicados a prácticamente cualquier temática (los estudios sobre integrismo y radicalismo están especialmente extendidos), y en segundo, por la cantidad de entradas y usuarios en muchos de ellos, que pueden sobrepasar los millones. Además, el material de estudio ya se encuentra digitalizado, lo que facilita el procesado posterior.
- Datos "naturales" con pocas restricciones sociales, dada su estructura jerárquica. Así, su estructura definida simplifica el proceso de selección e interpretación de la información. Para algunos autores (Moloney et al., 2003), los foros son una especie de "grupo de discusión virtual" sin moderación, en los que los miembros de la comunidad discuten sobre temas sin la interferencia de un investigador (y, por lo tanto, sin que éste modifique su expresión y comportamiento). Así, esta información puede ser considerada más natural y espontánea. A la vez, la suma de actores en conjunto es más que la suma de los actores de forma individual, ya que las intervenciones de cada uno de ellos provoca cada vez más y más detalladas intervenciones (Stephenson et al., 1991).
- La sinceridad, dado el anonimato parcial de los contribuidores, que hace que las respuestas y contribuciones sean más claras y, a menudo, menos "políticamente correctas". Así, es más fácil encontrar

opiniones radicales, extremistas o con cargas ideológicas importantes, preocupándose menos por la aceptación del público general (Glaser et al., 2002) (aunque hay que tener en cuenta que sí pueden buscar la aceptación del grupo de forma interna).

- Datos públicos, que añaden reproducibilidad y transparencia al análisis.

No obstante, el estudio de los foros también tiene sus inconvenientes, tales como:

- Esa anonimidad que potencia la recogida de datos limita a su vez el análisis al eliminar la posibilidad de trazar los datos demográficos de los usuarios; a menudo, además, los pocos datos que pueden estar disponibles (edad, sexo, ocupación) son difícilmente verificables.
- La sinceridad de la que se hablaba en las ventajas tiene el inconveniente de la tendencia a que las opiniones sean más extremas o más ofensivas en la red de lo que serían en la realidad. Ya no se trata de ser o no ser "políticamente correcto", sino de buscar el extremo contrario para ser lo más "políticamente incorrecto" posible (Williams et al., 2002). Algunos estudios evidencian, no obstante, que los usuarios de los foros online tienden a dar sus opiniones reales en la mayoría de temas, aunque a veces sean más agresivos u ofensivos de lo normal (Glaser et al., 2002).
- Privacidad. El aspecto ético debe ser tenido en cuenta y, aunque la conversación en Internet se pueda considerar pública, hay ciertos ámbitos en los que se debería respetar la privacidad de las conversaciones. Hay múltiples discusiones abiertas al respecto, especialmente en medicina, pero el consenso parece estar en que algunos foros están pensados para crear una esfera privada o de confianza (por ejemplo, foros de autoayuda para enfermos o víctimas) y no deberían ser tratados sin consentimiento explícito de sus usuarios, mientras que otros de índole política, religiosa o de grupos de interés hacen una función pública. De hecho, muchos de ellos aprovechan esa "publicidad" para atraer a nuevos miembros. En cualquier caso, debe considerarse caso a caso la necesidad de la protección de la información extraída.
- Representatividad. Si analizamos el aspecto estadístico de la muestra; ¿es ese foro o comunidad online realmente representativo de la población? En algunos casos específicos es altamente probable que no toda la población a estudiar tenga la misma facilidad de acceso al foro y esto hay que tenerlo en cuenta a la hora de establecer conclusiones generalizables.

Pero, quizás la pregunta más importante en relación a los foros sea "¿por qué funcionan los foros?". La respuesta es bastante general ya que puede ser aplicada a todos los medios de componente social, incluso

(y más) a los más recientes (Facebook, Twitter, Instagram, LinkedIn, etc.). Y es que todos estos medios tienen especial capacidad para mantener y amplificar los llamados enlaces débiles (*weak ties*) (Granovetter, 1977). Los enlaces débiles se pueden definir como conexiones sociales entre personas que tienen poca (o nula) intensidad emocional; se podría decir que están de acuerdo en las condiciones mínimas para relacionarse pero que se parecen poco y consumen poca energía en mantener esa relación. Aun así, estos enlaces son de alto potencial en la transmisión de información , como estudió Granovetter. El resumen es que el bajo contenido emocional de esos enlaces es lo que posibilita, precisamente, que esas personas tengan opiniones muy distintas en múltiples temas pero que puedan exponerlos sin entrar en conflicto. Además, la naturaleza de baja frecuencia y asincronía provoca que la información transmitida suela incorporar elementos nuevos cada vez y con distintos puntos de vistas. Parece que todo son ventajas, pero no. El gran problema es económico: si la probabilidad de que un enlace débil sea útil es prácticamente cero pero el coste de mantener ese enlace es (aunque pequeño) distinto de cero, lo natural es que exista una limitación natural para su acumulación. Hill and Dunbar (2003) argumentan que ese límite cognitivo se sitúa en torno a las 150 conexiones, aunque la varianza de esa media sea muy elevada (hay gente con muchas menos conexiones que otras). La cifra ha sido discutida multitud de veces y es muy probable que sea más elevada (sólo hay que hacer el ejercicio mental), pero lo que importa no es la cifra sino la idea.

Al volver a las redes sociales y sus plataformas, no obstante, es fácil darse cuenta de su efecto en ese coste mental: ayudan a minimizarlo. Es decir, los enlaces débiles son escalables en las plataformas sociales online. La comparación es igualmente simple: es mucho más fácil actualizar el perfil en Twitter o en un foro que contactar a todos sus miembros para contarles el último cambio. Así, ese límite de conexiones se vería ampliado en potencial y alcance hasta cifras inalcanzables sin ese apoyo.

1.3. Análisis de Redes Sociales

El análisis de redes sociales (SNA, por sus siglas del inglés *Social Network Analysis*) es una disciplina que estudia e investiga el reconocimiento, descripción, análisis y visualización de redes complejas en las que las relaciones entre sus elementos tienen un componente social. No se trata de una disciplina reciente: sus orígenes se pueden trazar hasta los estudios de Moreno (1934), introductor de los sociogramas. Pero la disciplina ha experimentado un gran auge tanto en el número de investigaciones como en el de aplicaciones desde finales de los años 1990 y, especialmente, desde principios del siglo XXI. Este auge se debe a dos fuerzas principales

(Barabási, 2016):

- **La eclosión de las herramientas disponibles para el mapeado de redes**. En el pasado resultaba imposible "dibujar" una red que constase de una cantidad elevada de relaciones. Analizar la red de amistades de una persona, por ejemplo, requería realizar encuestas personales y, aún así, las redes obtenidas eran limitadas y difíciles de verificar. Hoy día es posible cartografiar Internet y las relaciones entre páginas Web o utilizar plataformas sociales como Facebook yTwitter para obtener estos mapas de forma masiva, detallada y precisa. La revolución tecnológica de los últimos años no tan sólo ha traído las bondades de la red de redes para el análisis, sino que ha venido acompañada de la existencia de una capacidad de computación más barata y potente, facilitando el acceso, la recolección y el análisis de datos relacionados con las redes sociales.
- **La universalidad de las características de estas redes**. Aunque a simple vista pueda parecer que cada red (sea social o no) es única y presenta muchas diferencias respecto al resto, en realidad uno de los grandes descubrimientos sobre los que se basa el análisis de redes es que, en el fondo, son mucho más parecidas de lo que se intuiría. Y es que la arquitectura de las redes que se encuentran en la sociedad, ciencia, naturaleza y tecnología son similares porque son gobernadas por unos principios comunes que permiten utilizar y aplicar el mismo conjunto de herramientas matemáticas para explorarlos.

En los siguientes apartados se presentan desde algunos conceptos básicos sobre redes hasta métricas más concretas que serán de utilidad y aplicación en este trabajo.

1.3.1 Conceptos básicos

Formalmente, una red es un conjunto de relaciones (Kadushin, 2012). Una red está formada, a su vez, por un conjunto de entidadeso actores (nodos o vértices desde el punto de vista matemático) y una descripción de las relaciones entre éstos. Una red social estaría formada, por lo tanto, por un conjunto de entidades y de relaciones de carácter social.

En el contexto del análisis de redes sociales, un actor es una unidad social discreta, ya sea individual, colectiva o corporativa. Así, un actor puede ser una persona, un país o un departamento de una empresa. Los actores se relacionan los unos con los otros mediante enlaces (también llamadas aristas) que establecen una relación

entre un par de actores. Estas relaciones pueden ser de muchas naturalezas distintas (Wasserman and Faust, 1994), pero algunas de las más habituales son:

- Evaluación de otra persona, como amistad, atracción o respeto.
- Transferencias de recursos, como relaciones empresariales o préstamos.
- Asociación o afiliación, cuando dos actores asisten al mismo evento o son socios de un mismo club.
- Interacción, desde conocer a hablar o enviar mensajes.
- Movimientos entre sitios o estatus, como es el caso de la migración o trabajadores y cambios de empresa.
- Conexión física, como puentes o carreteras en un mapa.
- Relaciones formales, como autoridad.
- Relaciones biológicas, como la familia.

Así, dos actores A y B pueden ser, por ejemplo, dos personas, y la relación entre ellas puede ser tan simple como ser hermanos. La relación de hermanos no tiene una dirección concreta (si A es hermano de B, B es también hermano de A) y, por lo tanto, se clasifica como no dirigida. Otras relaciones, en cambio, sí se ven afectadas por la dirección y se consideran dirigidas. Que a A le guste B no implica, en ningún caso, que B corresponda a A. No obstante, si la relación dirigida es recíproca (es decir, A y B se gustan mutuamente) se habla de una relación simétrica. La mutualidad es más difícil de observar de lo que parece y, por ese motivo, la mayor parte de relaciones de una red dirigida suelen ser anti-simétricas (como padre e hijo o jefe y empleado) (Kadushin, 2012). Las relaciones, a su vez, pueden tener "valor" o no tenerlo, llamado habitualmente como peso. Existen relaciones puramente binarias, que son o no son (como ser amigo o ser hermano) y otras que se pueden valorar por su intensidad (por ejemplo, el número de veces que se llama a los amigos o el valor de los intercambios comerciales entre países).

Dos actores y la (posible) relación entre ellos conforman la unidad mínima del análisis de redes sociales: la díada. El análisis de las díadas de una red puede centrarse en las propiedades de las relaciones entre parejas, en la reciprocidad y en los tipos de relación, pero en general es poco interesante como unidad, al contener poca información. La mayoría de métodos del análisis de redes sociales se centran y se basan, en cambio, en la que se podría llamar la unidad mínima de interés: el conjunto de relaciones entre tres actores, la tríada, estudiada por Simmel en 1908 y citada en Noteboom (2006), de la que se hablará más adelante.

¿Qué hace que dos nodos decidan relacionarse entre ellos? Las relaciones sociales son complejas por naturaleza

pero existen dos principios que las caracterizan:

- **Propincuidad**: la propincuidad se puede definir de forma general como estar en el mismo sitio en el mismo momento, aunque parte de la idea que dos nodos son más propensos a unirse (en igualdad de condiciones) si son geográficamente cercanos. ¿En qué se traduce esto? En que es más probable que, por ejemplo, dos personas sean amigas si viven cerca (Feld and Carter, 1998) o que dos países tengan relaciones comerciales si comparten frontera. Esa proximidad no tiene porqué ser únicamente física; intereses comunes como aficiones, costumbres o gustos musicales pueden ser igualmente importantes para crear relaciones (por ejemplo, los dueños de mascotas tienden a relacionarse con otras personas con mascota o los padres y madres con hijos establecen relaciones entre ellos en las puertas de los colegios).
- **Homofília**: la homofília (Lazarsfeld and Merton, 1954) (amor de lo mismo) es uno de los conceptos más importantes de las redes sociales y se resume en el popular "pájaros de la misma pluma vuelan juntos". Más formalmente, dos individuos que compartan características en una proporción más elevada que la esperada en la población de la que forman parte es más probable que estén conectados (Verbrugge, 1977). También se puede plantear del revés: dos personas conectadas es más probable que compartan características o atributos. La homofília es un círculo que se retroalimenta constantemente: dos individuos conectan porque son parecidos y, a la vez, se vuelven más parecidos porque conectan. Lo mismo sucede con grupos, países, organizaciones o cualquier unidad social.

1.3.2. Niveles de análisis

La distinción más básica en el análisis de redes sociales es, probablemente, aquella entre los diseños de investigación sociocéntricos y egocéntricos (Perry et al., 2018). Las redes egocéntricas son aquellas que se centran en el estudio de las relaciones de un único nodo o individuo (el *ego*) y su entorno social inmediato. Estas redes describen la red definida por el individuo estudiado (por ejemplo, su red de contactos) e intentan responder a cuestiones de investigación relacionadas con sus conexiones y las características tanto de sus contactos (conocidos como *alters*) como las relaciones que tienen entre ellos. El estudio de redes egocéntricas se basa en el principio que los individuos (*egos*) existen en un contexto social (de *alters*) que les afecta y que, por lo tanto, puede hacer más o menos probable que encuentren un trabajo, por poner un ejemplo.

Los estudios sociocéntricos o sociométricos, en cambio, se interesan por una población. En este tipo de redes, los nodos se unen unos con otros y forman caminos que, considerados en su conjunto, terminan construyendo la estructura de la red. Esta estructura es el objeto principal de estudio, junto a las posiciones o roles que ocupa cada nodo en ella. De hecho, en un estudio sociométrico se suelen considerar tres subniveles de estudio: el nivel individual, para estudiar las posiciones de los actores dentro de la red, como su popularidad o influencia; el nivel de los subgrupos, para identificar si existen comunidades y qué las caracteriza; y el nivel de la red completa, para comparar la densidad de sus enlaces o su estructura general, por ejemplo.

Para capturar la estructura de una red social se usa el diseño sociométrico, que consiste básicamente en determinar los límites de una población para seleccionar todos los nodos que se encuentran en ella y luego proceder a anotar todos y cada uno de los enlaces entre nodos de la población. Cuando los citados límites están claramente determinados (por ejemplo, alumnos en una clase o trabajadores de una empresa) se habla de redes sociocéntricas o cerradas. En caso de no tener límites claros (como podría ser la difusión de una información o las influencias en una plataforma social) recibe el nombre de red abierta .

1.3.3. Representación de redes sociales

Una de las formas más habituales de conceptualizar las redes sociales de forma matemática es mediante grafos (Borgatti and Johnson 2018). Es importante aclarar, no obstante, que aquí el concepto de grafo no es el de diagrama, sino el de abstracción matemática (Harary, 1969). Un grafo $G\left(V, E\right)$ consiste de un conjunto de nodos o vértices V y una colección de relaciones o aristas E, que conectan pares de vértices. Así, si un par de nodos u y v están conectados en un grafo G, entonces es posible escribir que $(u, v) \in E\left(G\right)$. Si dos nodos están unidos por una arista, entonces esos dos nodos son adyacentes o vecinos. Como las relaciones pueden ser dirigidas o no dirigidas, los grafos también lo pueden ser igualmente. Las aristas en un grafo dirigido suelen llamarse arcos y la ordenación de las parejas de vértices indica su dirección. La información de un grafo G también se puede expresar en forma matricial, en la referida habitualmente como matriz de adyacencia (Wasserman and Faust, 1994) puesto que cada posición x_{ij} de la matriz indica si los nodos i y j son adyacentes. La matriz de adyacencia de una red no dirigida es simétrica.

Aunque a nivel matemático no sea un requerimiento, en redes sociales se suele analizar grafos cuyas aristas representan siempre el mismo tipo de relación, basándose en el principio que cada relación define una estructura distinta. Es intuitivo, al menos, que las relaciones de amistad y las de gustar en un mismo conjunto de actores pueden presentar algún parecido pero no serán, en ningún caso, idénticas (por ejemplo, es habitual tener muchos amigos pero se suele sentir atracción hacia menos). Si se tiene un grafo con más de una relación entre los mismos vértices se habla de multigrafo; lo más común es, no obstante, analizar cada una de las dimensiones por separado.

La mayoría de las aplicaciones del análisis de redes sociales se centran en redes que contienen actores del mismo tipo o nivel, en las que las relaciones son directas y entre iguales. Estas redes suelen ser conocidas como unimodales. Una red de amigos es un claro ejemplo de ello: los nodos son todos "personas" y las relaciones son de amistad, que son relaciones directas y claras entre "personas". A veces, no obstante, por construcción o por restricción no es posible obtener estas relaciones de forma directa (Borgatti and Johnson, 2018). En estos casos, puede ser posible inferir o predecir los enlaces a partir de afiliaciones, membresías o asistencia a unos mismos eventos. Las relaciones bipartitas tienen interés para la sociología por su dualidad (Breiger, 1974). Lo que significa, al fin y al cabo, es que las ideas, actitudes y conexiones sociales de las personas se forman a través de su pertenencia a distintos grupos y que estos grupos son, a su vez, moldeados por sus miembros. La idea es que dos personas que atienden a las mismas reuniones es probable que se relacionen (aunque no se pueda asegurar con la misma certeza que en las redes unimodales). Así, en estos casos, los nodos tendrían dos tipos distintos: personas y eventos, por ejemplo. Este tipo de redes se conocen como bimodales y los grafos que forman son bipartitos.

1.3.3.1. Grafos bipartitos

Formalmente, un grafo es bipartito si sus nodos pueden ser divididos en dos conjuntos N_1 y N_2 tal que cada arista del grafo es un par de nodos en el que uno de los nodos pertenece a N_1 y el otro a N_2. Es decir, en un grafo bipartito los nodos se pueden dividir en dos grupos y todos los enlaces son entre nodos de grupos distintos; no hay enlaces internos en un mismo grupo. Las redes sociales generadas a través de foros, como muchas otras, son en realidad un caso especial de red en la que los nodos pertenecen a dos tipos distintos. Hay nodos que son usuarios y hay nodos que son hilos en los que comentan, mientras que los enlaces entre usuarios e hilos están determinados por las relaciones que se establecen entre unos y otros. En

esta configuración, sin embargo, no hay relaciones entre nodos del mismo tipo.

Aunque los grafos bipartitos se pueden estudiar directamente (y eso presente algunas ventajas teóricas) (Opsahl, 2013), lo más habitual es proyectarlos (pasarlos de modo 2 o bipartido a modo 1, ver figura 1). Este proceso se conoce como aplanado y genera dos redes nuevas a partir de la original. Una relaciona los usuarios entre sí (hay relación si atienden a un mismo evento, por ejemplo) y otro relaciona a los eventos entre sí (si tienen usuarios en común). Usando ambas proyecciones se pierde relativamente poca información (Everett and Borgatti, 2013) y, a cambio, se gana la posibilidad de aplicar todas las técnicas y algoritmos que solo están definidos para el caso unidimensional.

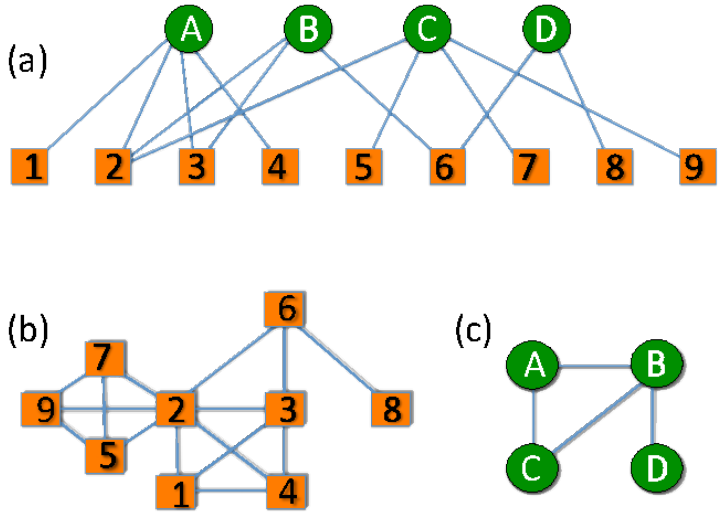


Figure 1: Un grafo bimodal y sus dos proyecciones unimodales. Fuente: http://inspirehep.net/record/1353428/plots

La figura 1 muestra la forma de proyección más sencilla. En (a), los nodos A, B, C y D podrían representar eventos y los nodos numéricos, personas. Nótese como los dos niveles se conectan entre ellos pero no lo hacen internamente. Así, la proyección desde el punto de vista de las personas sería la correspondiente a (b), donde cada enlace indica que los dos nodos conectados coincidieron al menos una vez en un evento. También es posible proyectar los eventos (c), que podría interpretarse como "eventos a los que han asistido las mismas personas". Así, B y D están conectados porque la persona 6 ha asistido a ambos eventos, pero

no hay conexión entre A y D porque no hay nadie que haya coincidido en ambos. La proyección de personas daría pues información sobre las agrupaciones entre ellas (por ejemplo, para detectar grupos de amigos), mientras que la de eventos haría lo propio con las tipologías o características de los eventos.

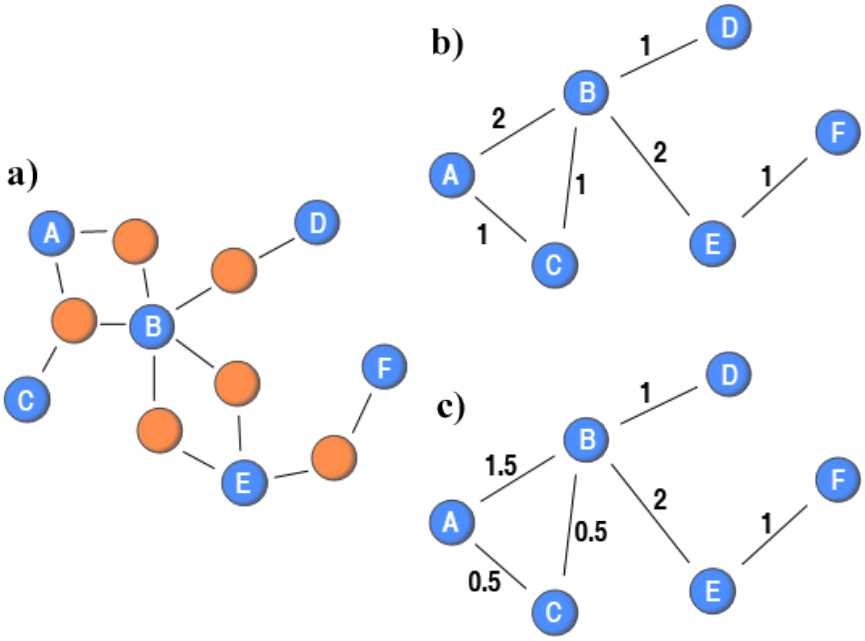


Figure 2: Proyección de una red bipartida añadiendo valor a los enlaces. Fuente: https://toreopsahl.com/tnet/two-mode-networks/projection/

Si bien la proyección tradicional no incorpora pesos en sus aristas, hacerlo evita que se pierda información adicional. En el caso de la Figura 2, se muestra (a) una red bipartida que se proyecta sobre sus nodos A-F en (b). Cada arista en (b) representa, como en la figura anterior, una conexión entre dos personas que han asistido al menos a un mismo evento, pero aquí cada uno de los pesos de (b) se puede identificar, además, como el número total de eventos en los que han coincidido. A y B, por ejemplo, han estado juntos en dos eventos, mientras que A y C sólo lo han hecho en uno. Es presumible que la relación entre A y B tiene más potencial de ser fuerte que entre A y C. Este procedimiento fue extendido por Newman (2001b) para las redes de colaboración científicas (co-autorías de artículos) porque entendía que no debía valorarse de la misma forma una colaboración entre pocos autores (más fuerte) que una entre muchos (más débil). Newman introdujo un descuento en función del tamaño que daba respuesta a esta premisa, siguiendo la fórmula $w_{ij} = \sum_p \frac{1}{N_p - 1}$, donde N_p es el número total de autores para el artículo p. Si se imagina que (a) es, ahora, la red de colaboración y co-autoría entre seis investigadores, es posible entender que B y D han colaborado en un artículo ellos solos (por lo tanto, el peso es 1) mientras que la colaboración entre A y C

es, en realidad, combinada con B, así que el peso de esa relación es menor (0,5).

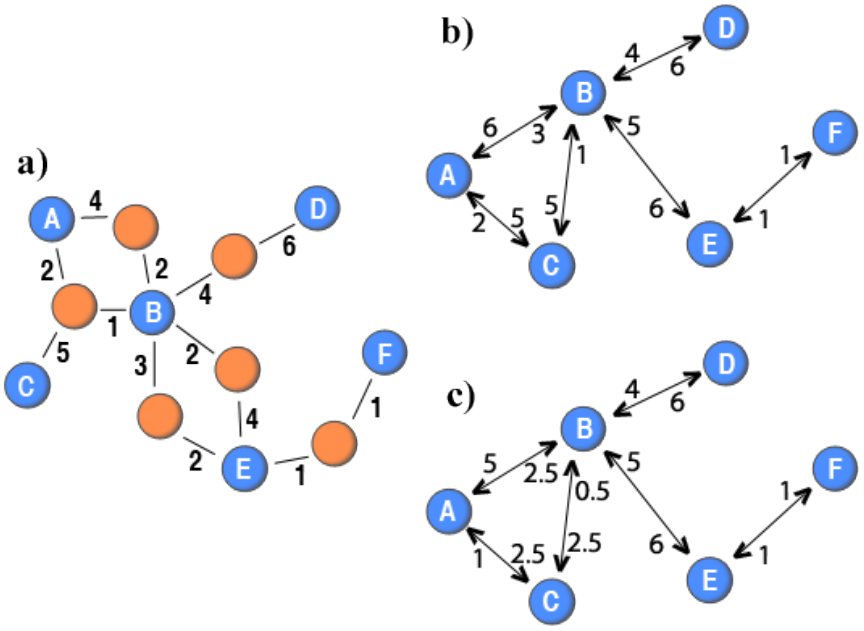


Figure 3: Proyección de una red bipartida valorada. Fuente: https://toreopsahl.com/tnet/two-mode-networks/projection/

Las propuestas de las figuras 1 y 2 solo han considerado redes bipartitas binarias, que tenían o no relación. Sin embargo, también existen grafos bimodales en los que el peso es importante. En el caso de los foros online, de hecho, si se establece una relación bipartita entre hilos (o temas) y los usuarios que escriben en ellos, tiene sentido considerar el número de mensajes escritos por cada usuario para representar la intensidad de su interacción (Figura 3a). Su proyección unimodal (b) sigue la misma lógica que la versión anterior; aquí, no obstante, aunque la relación sea recíproca (si A ha escrito en el mismo foro que B, B ha escrito en el mismo foro que A) no es totalmente simétrica, puesto que A puede haber escrito 6 mensajes y B sólo 3. Así, la dirección indica el número de mensajes que uno de los actores ha escrito en un foro en el que el otro también participa. Volviendo a (b), A ha escrito 2 mensajes en un foro compartido con C, mientras que C lo ha hecho en 5 ocasiones. Es posible generalizar la aplicación del método de Newman (2001a) descontando en función de esa misma intensidad (multiplicada por el peso) (c), que básicamente se reduce a la observación que no tiene la misma relevancia una aportación en un hilo con miles de usuarios que un mensaje en un tema con pocos actores (en cuyo caso la "fuerza" de la interacción es mucho más notable).

1.3.4. Métricas de red

En los siguientes apartados se exponen las principales métricas de análisis en redes sociales.

1.3.4.1. Densidad

Si se considera el número y la proporción de aristas en el total de un grafo se puede observar que el número máximo de aristas posibles está limitado por el número de nodos. En un grafo con g nodos, y excluyendo enlaces reflexivos de un nodo consigo mismo, el máximo de aristas sería $\binom{g}{2} = \frac{g(g-1)}{2}$. Si se tiene en cuenta el número de aristas que hay realmente en el grafo, se llega a la definición de densidad: la proporción de aristas que existe en el grafo respecto a todas las posibles. Si las aristas presentes son L, entonces la densidad (Δ) se define como $\Delta = \frac{L}{\frac{g(g-1)}{2}} = \frac{2L}{g(g-1)}$. Un grafo sin ninguna arista tiene densidad 0 y un grafo completo, con todas las aristas posibles, tiene densidad 1. El resto de posibilidades se mueven en el rango de 0 a 1.

1.3.4.2. Caminos y distancias

En una red, un nodo es adyacente a otro si está conectado con él. Pero esta no es la única forma en la que dos nodos pueden estar conectados, ya que también pueden estarlo de forma "indirecta". Por ejemplo, si A conoce a B y B conoce a C, entonces A podría llegar a acceder a C a través de B. Entender las distancias y los caminos existentes entre nodos es crítico para comprender su estructura y geometría.

Las secuencias de nodos conectados se pueden clasificar de tres formas (Wasserman and Faust, 1994):

- Paseo (*walk*): un paseo es una secuencia de nodos y aristas que empieza y termina en un nodo y en el que cada arista es incidente con los nodos que le preceden y suceden en la secuencia. En un paseo se puede visitar un nodo más de una vez y también se puede empezar y terminar en el mismo nodo.
- Sendero (*trail*): un sendero tiene una restricción adicional y es que no puede pasar dos veces por la misma arista (pero sí repetir nodos).
- Camino (*path*): un camino es todavía más restringido, puesto que no permite que se pase dos veces por

el mismo nodo (y tampoco por la misma arista). Así, un camino es una secuencia de nodos y aristas que se conectan y no se repiten.

La longitud de un paseo, sendero o camino es el número de aristas que aparecen en la secuencia. Una vez conocidas estas métricas básicas, es posible definir el camino geodésico, la distancia y la alcanzabilidad.

- El camino geodésico es el camino más corto (de longitud mínima) entre dos vértices. Puede ser múltiple.
- La distancia entre dos vértices es la longitud del camino geodésico.
- Un nodo es alcanzable a nivel n por otro si existe una ruta entre ellos de distancia n o menos. Dicho de otra forma, dos nodos son alcanzables a nivel n ni su distancia geodésica es menor que n.

1.3.4.3. Conectividad y componentes

Una vez definidas las medidas de caminos y distancias, otra de las propiedades importantes que aparece es la conectividad. Un grafo está conectado si existe al menos un camino entre cualquier par de nodos en el grafo. Dicho de otra forma, una red está conectada si todos los nodos son alcanzables desde el resto de nodos. Si no se cumple esta propiedad, la red está desconectada y cada subconjunto de nodos (maximal) conectado que no está conectado con otros subconjuntos es un componente. Un subconjunto de nodos es maximal si no es posible añadir ningún otro nodo manteniendo la propiedad que lo define. Nótese que si el grafo es conexo, solo tiene un componente. Si está, por ejemplo, partido en dos, se dice que tiene dos componentes y que no es conexo.

De la definición de conectividad y componentes salen dos roles relevantes para los nodos y aristas: los puntos de corte y los puentes. Un punto de corte no es más que aquel nodo que, de eliminarlo, dividiría el grafo en más componentes de los que tiene con el nodo incluido. Un puente es el equivalente en aristas: sería aquella arista que, de eliminarla, partiría el grafo en más componentes. Cuando un grafo está conectado se puede definir la conectividad de grado y la conectividad de línea o arista, que se define como el número de nodos (o aristas) que deben eliminarse para desconectar el grafo.

1.3.5. Métricas de nodo

1.3.5.1. Grado

El grado de un nodo es el número de aristas que inciden en él. Otra forma de definirlo es a partir de los nodos, puesto que el número de aristas incidentes coincide con el número de nodos adyacentes o vecinos que tiene un nodo. El grado de un nodo puede ir desde $g - 1$ (para el caso que el nodo se conecta con todos los demás nodos del grafo) hasta 0 (sin conexiones, en cuyo caso es un nodo aislado). Para el caso dirigido se suele distinguir entre el grado entrante y el saliente; el primero es el número de conexiones que llegan al nodo y el segundo el número de conexiones que salen de él.

1.3.5.2. Centralidad y prestigio

Una de las principales motivaciones para el análisis de redes sociales es la de cuantificar el poder y la influencia de los individuos en función de sus relaciones. Quizás por esta motivación general el concepto de centralidad y de prestigio es uno de los más discutidos en la disciplina, puesto que el mismo significado de "poder" o "influencia" en una red es muy variable.

La centralidad permite medir, desde una perspectiva específica, quién es más importante en una red o quién tiene más posibilidades de ejercer influencia en la misma. En una organización, por ejemplo, el más influyente en la transmisión de información no tiene porqué corresponderse con el jefe de más alto rango; de hecho, normalmente no lo es (Krackhardt and Hanson, 2006). Pero hay poco consenso en torno a lo que significa realmente la centralidad (Borgatti, 2005). La interpretación de una centralidad (de alguno de sus múltiples tipos) en una red de intercambio de información debería ser interpretada de forma distinta a una red familiar, por ejemplo. Al final, de lo que depende es de la propia concepción de poder o influencia en el contexto de estudio y no tanto del propio resultado de la medida.

Cada comunidad tiene sus "celebridades" o nodos con un poder significativamente más elevado que la media de usuarios. No suelen ser demasiados y sus indicadores tienden a ser órdenes de magnitud más elevados. En líneas generales, la centralidad se puede medir con cuatro métricas distintas:

- **Centralidad de grado**. Probablemente la más simple, corresponde al grado del nodo o, lo que es lo mismo, al número de conexiones que tiene ese nodo. Es lo que en Twitter sería el número de *followers* o los suscriptores en YouTube. La centralidad de grado tiene un punto de centralidad "en bruto" o potencial. Tener muchos seguidores es una medida del potencial que tiene la transferencia de un mensaje, sí, pero podría tratarse de vínculos muy débiles y, por lo tanto, sería posible encontrar individuos con menos relaciones en comparación pero mucho más robustas que presentaran mayor influencia. La equivalencia coloquial de un individuo con alta centralidad de grado sería la de una "celebridad" o un "famoso".
- **Centralidad de proximidad**. Si se tiene en cuenta que la capacidad de un individuo para recibir y para transmitir información depende principalmente de la distancia entre el mismo y el resto de la red, se puede inferir que la cercanía entre nodos es también un indicador de centralidad. Así, la distancia define también el rol de un nodo en la red; los nodos más cercanos en media son también aquellos que invierten más tiempo y recursos en relacionarse con los demás. Son, para que se entienda, los individuos más "cotillas", los que están en todos los frentes.
- **Centralidad de intermediación**. La centralidad de intermediación se basa en un principio mucho más estratégico: los nodos que se sitúan en los caminos críticos de la red tienen un poder que aquellos nodos más "del montón" no tienen. La centralidad de intermediación no tan solo identifica el poder relativo, sino que también indica qué nodos están en la frontera de las comunidades; no es difícil de imaginar dos grupos de amigos separados y que no se conocen entre ellos pero con una persona que pertenece a ambos. Cualquier contacto o conexión futura entre ambos grupos depende de esa única persona en común, el *broker* o intermediario.
- **Centralidad de vector propio (eigenvector)**. Hasta ahora, las centralidades que se han visto son posicionales, estructurales de la red; la centralidad de vector propio, en cambio, pretende medir la influencia en las relaciones. Dicho de otro modo, no es lo mismo tener decenas de seguidores de baja importancia que tener unos pocos seguidores de muy alta relevancia. Es más o menos la idea tras el algoritmo de *PageRank* de Google (Page et al. 1999), que no solo valora los enlaces incidentes sino la importancia relativa de los mismos. Los individuos con alta centralidad de vector propio son las eminencias grises, influyentes personajes poderosos que actúan desde las sombras como un poder oculto. Es el caso de *El Padrino*: dando órdenes a individuos con suficiente poder podía ejecutar sus planes sin tener que exponerse.

Aunque no es el objetivo de esta contextualización la de entrar más en detalle, es importante entender que

las medidas de centralidad no son más que eso, medidas. En la mayor parte de los casos solo cuentan la mitad del mensaje; la otra mitad puede obtenerse del análisis de las interrelaciones entre esos personajes, que van más allá de las relaciones binarias. Las triadas (grupos de tres nodos) son la mínima expresión de estas relaciones de mayor interés, porque las podríamos considerar la "sociedad" mínima. Desde esas triadas se puede empezar a entender las sociedades mayores (los *cliques* o camarillas) y los grupos o *clusters,* cuyas características resultarán clave para establecer las comunidades.

1.3.5.3. Coeficiente de agrupación

El coeficiente de agrupación pretende indicar la asociación de los vecinos de un nodo entre sí; es una medida de densidad local y también se puede llamar coeficiente de clustering o densidad egocéntrica. En la figura 4 se muestran varios ejemplos, desde el valor máximo, 1, cuando todos los vecinos de un nodo están conectados, hasta el mínimo, 0, cuando los vecinos dependen estrictamente del nodo de estudio para mantener su conexión.

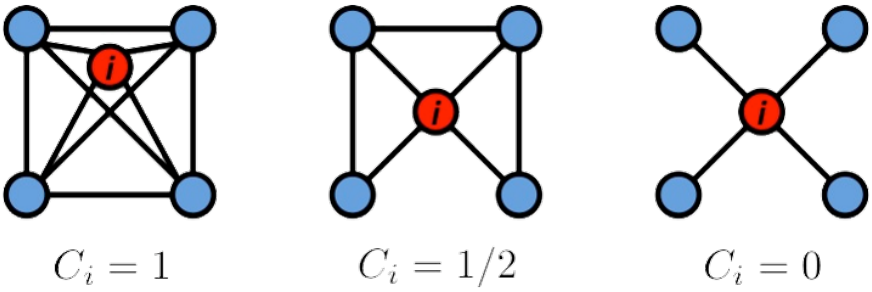


Figure 4: Coeficiente de agrupación. Fuente: http://networksciencebook.com/chapter/2#clustering

1.3.6. Triadas y comunidades

Una triada es, simplemente, un conjunto de tres nodos relacionados de alguna forma. En el caso de grafos dirigidos (en los que las relaciones son direccionales) existen hasta 16 combinaciones posibles en la relación y cada uno de los casos cuenta una historia distinta. Esta estructura fue la primera en estudiarse en redes sociales (Noteboom, 2006) y el trabajo de Georg Simmel (Simmel, 1950) es, probablemente, la pieza más antigua del rompecabezas del SNA.

De las triadas completas se puede pasar a los *cliques* o camarillas, que no son más que su extensión n-dimensional. Por definición, un *clique* es un subgrafo completo y maximal de un grafo, lo que significa que es un conjunto de nodos que cumplen dos condiciones: a) todos los nodos están conectados entre sí (completo) y b) no se puede añadir ningún nodo adicional sin perder la propiedad. Para conectarlo con lo anterior, un *clique* está formado por una serie de triadas completas que se sobreponen. Su significado intuitivo es mucho más fácil de entender: un *clique* en una red social es un grupo de individuos que están conectados fuertemente entre ellos (y de forma más débil con los ajenos al grupo).

Por definición, sin embargo, los *cliques* son muy rígidos y estrictos. Por más densa o conexa que pueda ser una red social en la realidad, difícilmente presentará todas las relaciones posibles entre sus miembros. Solo hace falta imaginar la pérdida de una arista en un conjunto completo: se deshace la camarilla y se generan dos camarillas menores, cuando la ausencia de la arista puede deberse a un error de medición o a que el grupo es tan grande que algunos miembros no se relacionan con todos los demás. Una definición más relajada de este tipo de grupos son los *k-cliques*, que permiten identificar los grupos de nodos maximales conectados, por lo menos, a k otros nodos dentro del subgrafo. Es decir, se sustituye la condición de completitud por otra aproximada.

Los métodos más habituales de detección de comunidades, que se construyen a partir de las ideas anteriores, son aquellos que se basan en el concepto de modularidad. Es importante destacar que también son hasta cierto punto vagos o confusos porque se basan en algoritmos que a menudo incluyen aleatoriedad, pero esto les otorga cierta flexibilidad. El concepto de modularidad es el siguiente: asumiendo que se dispone de un grafo partido en comunidades que no se solapan, la modularidad es la fracción de las aristas que caen dentro de las comunidades menos la fracción que se esperaría en una distribución aleatoria que mantuviese la distribución de grados (Newman, 2006). En palabras más simples, la modularidad da una idea de la cantidad de relaciones intra-comunidad frente a las relaciones inter-comunidad. Se entiende que en caso de tener una partición con muchas relaciones internas en las comunidades y pocas entre comunidades es mejor partición que una que mantenga muchas relaciones entre comunidades.

Idealmente, se intenta generar una partición con la modularidad lo más alta posible (que indicaría una

estructura de comunidades razonable). Sin embargo, este es un proceso difícil y costoso computacionalmente, para el que se han publicado diversos algoritmos (cada uno con sus pros y sus contras):

- El problema de la partición óptima es NP-Completo respecto al tamaño de la red; se puede decir que para encontrar la mejor partición se deberían calcular todas las particiones posibles del grafo y seleccionar la máxima, pero el número de particiones escala de forma no lineal y no tiene solución exacta.
- Las soluciones aproximadas, como la generada por el popular algoritmo de Blondel et al. (2008), funcionan bien pero en general se basan en modelos probabilísticos, lo que genera soluciones inestables, que varían en cada ejecución.
- Aun así, la detección en grafos grandes sigue siendo costosa y, además, acaba evitando la creación de comunidades pequeñas que suelen terminar integradas en comunidades mayores. A veces es recomendable partir la red en redes más pequeñas si se quiere evitar el efecto, pero no es lo más óptimo.

1.4. Procesamiento de Lenguaje Natural

El procesamiento del lenguaje natural (PLN por sus siglas en español y NLP en inglés) es una rama de la Inteligencia Artificial que une la computación con la lingüística, la ciencia con las humanidades. Es una técnica que viene motivada por un problema importante: la interpretación de lenguaje humano, tanto para su comprensión como para su creación. A nivel histórico, gozó de gran popularidad a mediados del siglo XX, pero la falta de resultados (y de capacidad de computación) hicieron que pasara por un bache hasta el resurgir de los métodos estadísticos ya hacia finales de siglo. A día de hoy es, sin duda, una de las principales áreas de interés. Se puede dividir en dos grandes áreas, partes de un mismo todo:

- Análisis, que es la parte del NLP que pretende analizar un texto para extraer su significado.
- Generación, que es la complementaria; es decir, la que pretende generar respuestas de forma natural.

1.4.1. Principios

El procesamiento del lenguaje natural se articula mediante las llamadas cadenas de procesamiento: procesos separados que se enlazan unos con otros para ir resolviendo problemas más pequeños y que, conjuntamente, dan solución al problema general. A nivel teórico se pueden considerar las siguientes fases en el análisis lingüístico:

- Sintaxis, para obtener la estructura de cada frase.
- Semántica, para su significado literal.
- Pragmática, para la adaptación del significado al contexto.

En realidad, no obstante, la extracción de información a partir de lenguaje natural es bastante más compleja y el proceso termina separado en múltiples pasos menores:

1. Un preprocesamiento, subdividido en limpieza del texto, codificación de caracteres e identificación del idioma.
2. La segmentación de los textos, que empieza con la *tokenización* (segmentación de un texto en palabras o *tokens*) y la segmentación de frases para separarlas entre sí.
3. El análisis léxico o morfológico, que asigna a un lema y una etiqueta morfológica a cada token. A menudo se utiliza la lematización o el *stemming*, que son dos procesos distintos, para obtener los lemas o raíces de las palabras.
4. El análisis sintáctico, que es el que se encarga de relacionar y agrupar las palabras entre ellas y que intenta trabajar con todas las combinaciones posibles.
5. Y finalmente, el análisis semántico, que es el que pretende entender los textos y que, como tal, es el más complejo de todos, porque no tan solo tiene que resolver la ambigüedad del paso anterior, sino que también tiene que resolver elementos de lógica, de relaciones semánticas y ontologías.

1.4.2. Topic Modeling

Una técnica particular que se aplica de forma habitual a "textos" es el llamado *topic modeling,* que en realidad es un conjunto de algoritmos destinados a asignar cada texto o documento a uno (o más) "temas", habitualmente de forma no supervisada (como un *clustering*). Una aplicación habitual de la técnica es para categorizar noticias de periódico, por ejemplo.

En realidad, cuando se habla de *topic modeling* suele hacerse de una implementación concreta, el llamado LDA (*Latent Dirichlet Allocation*), propuesto por Blei et al. 2003. Explicado de manera resumida, el proceso LDA intenta encontrar grupos de palabras (a los que se llama "temas") que aparecen de frecuentemente de forma conjunta. Hay que entender que los temas de un documento no son únicos, sino que se entienden como una mezcla; una noticia puede ser, por ejemplo, 80% deporte y 20% economía. Las categorías o temas tampoco tienen porqué entenderse de la misma forma que temas "al uso"; una categoría como deportes puede tener más habitualmente palabras como "gol", "jugador" o "equipo", pero si sus artículos son escritos por dos personas distintas con una escritura característica (imaginen a una usando "cancerbero" y "guardameta" mientras otra habla siempre de "portero") es posible que aparezca otra subdivisión que tendría más que ver con el estilo que con el tema.

Esta técnica se utiliza para descubrir estructuras semánticas de textos ocultas en un corpus y es por este motivo por el que es una técnica adecuada para deducir las temáticas en foros online. En función de la aplicación puede requerir un paso manual. Si el objetivo es agrupar de la manera más automática posible, será suficiente con la salida en crudo del algoritmo. Sin embargo, si se quiere etiquetar mensajes o *posts* con una categoría concreta (por ejemplo, si el administrador del foro quiere simplificar la búsqueda de mensajes y les añade unas etiquetas de un listado cerrado), entonces será necesario analizar los temas extraídos por LDA, asociar cada tema con una categoría, y posteriormente usar estas asociaciones para clasificar. Esta aproximación automática o semiautomática ha sido comentada por Semberecki and Maciejewski (2016) o Nesi et al. (2015). Otra opción es etiquetar los textos manualmente y aplicar un algoritmo de *clustering* (Zhang et al. (2017)); requiere un intenso trabajo manual inicial además de ser difícil de extrapolar a otros entornos.

1.5 Infraestructuras de computación distribuida

El análisis de redes sociales y de procesado de lenguaje natural ofrece oportunidades a quien tenga datos susceptibles de analizar. Estas disciplinas no son nuevas: se ha comentado en los apartados anteriores que los orígenes de ARS se remontan a la década de 1930 y los de NLP a los 1950. Pero si bien la teoría ha estado presente desde hace tiempo, su eclosión fuera del ámbito puramente académico se debe en parte a la aparición de bases de datos masivas y al desarrollo de infraestructuras capaces de procesar tal cantidad de datos. Las siguientes secciones repasan la aparición de uno de estos sistemas, Hadoop, y la rápida evolución

hasta Spark, uno de los entornos de computación distribuida más extendidos en la actualidad.

1.5.1 Hadoop

El paradigma *mapreduce*, presentado por Dean and Ghemawat 2008, y su implementación de código abierto, *Hadoop/HDFS*, permitió el procesamiento de ingentes cantidades de información en entornos distribuidos construidos a base de hardware comercial (por contraposición a hardware específico, White 2011). Este paradigma se apoyaba en 4 pilares:

- Un sistema de ficheros distribuido tolerante a fallos. *Google File System* fue originalmente propuesto por Ghemawat et al. 2003 y posteriormente implemente en código abierto como *Hadoop Distributed File System* o HDFS. Este sistema de ficheros está orientado a almacenar ficheros de texto (típicamente tabulares) de tamaño arbitrariamente grande. Los divide en bloques que distribuye en un cluster de N nodos de manera que cada bloque esté alojado por al menos 3 nodos. Es además capaz de recuperarse ante la caída, temporal o permanente, de uno o más nodos, redistribuyendo y replicando los bloques automáticamente.
- Un modelo de programación en paralelo. El paradigma *mapreduce* intenta generalizar el procesado en paralelo ofreciendo al programador dos conceptos: tareas *map* y tareas *reduce* (ver figura 5). Las tareas *map* tienen como objetivo convertir cada registro de los datos de entrada en una o varias tuplas intermedias con un formato de {clave: valor}. En el ejemplo más típico, el contador de palabras, una línea de texto se convierte en tantas claves como palabras hay en la línea, siendo 1 el valor de cada clave. Las tareas *reduce* se encargan de agregar los valores de una misma clave con alguna operación concreta, devolviendo como salida una única tupla. Siguiendo con el ejemplo anterior, una tarea *reduce* sumará todas los valores de aquellas tuplas con la misma clave (es decir, la misma palabra) y el resultado será una tupla de tipo {palabra: *num_palabras*}. Ambas tareas están enlazadas por una fase de mezcla (denominada *shuffling* o *merge and sort*) en la que se agrupan todas las tuplas de la misma clave, para que sea una única tarea *reduce* la que las procese. No es necesario programarla, pero es habitual que la tarea *reduce* se pueda usar también en la fase de mezcla.
- Una infraestructura de coordinación de las tareas *mapreduce*. Siempre y cuando el programador sea capaz de expresar su algoritmo en forma de las dos citadas tareas, Hadoop es capaz de distribuir el código, arrancar las ejecuciones *map*, mezclar las salidas, ejecutar las tareas *reduce*, y consolidar el resultado final. También es capaz de detectar si una tarea ha fallado y ejecutarla de nuevo, dejando el

resto de tareas en espera, pero sin ejecutar todas de nuevo.

- La coubicación de datos y procesamiento. La infraestructura de coordinación es capaz de arrancar las tareas de *map* y *reduce* en los nodos que alojan los datos de entrada, reduciendo así la cantidad de datos transmitida por la red.

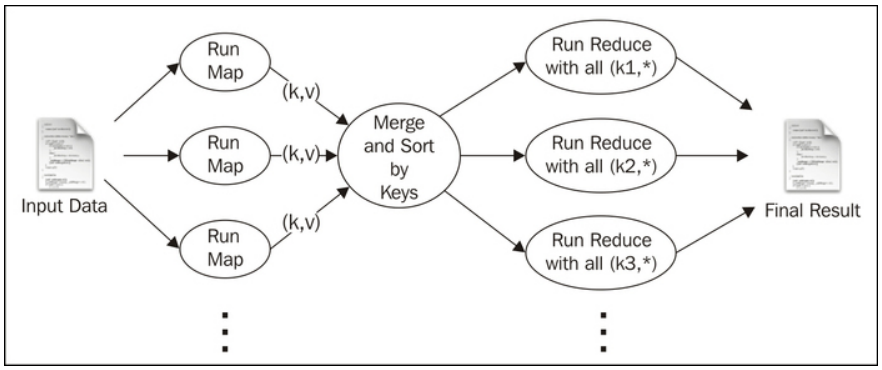


Figure 5: Esquema de *map/reduce*. Fuente: Gunarathne 2015

1.5.2 El ecosistema Hadoop

La aparición de Hadoop bajo el paraguas de la fundación Apache fue el pistoletazo de salida para una plétora de herramientas de código libre que han formado el denominado *ecosistema Hadoop* (Landset et al. 2015). Algunas herramientas se centraron en ofrecen a los desarrolladores una sintaxis similar (Apache Pig, Apache Hive) o incluso compatible (Apache HBase) con SQL, con el objetivo de abstraer el paradigma mapreduce sin perder sus beneficios. Otros proyectos generalizaron la capa de coordinación (Apache Hadoop YARN, Apache Mesos) para que fuera capaz de ejecutar tareas de varios motores de consulta. De hecho, el motor de *mapreduce* se independizó de la capa de coordinación y pasó a llamarse Hadoop MapReduce. YARN o Mesos se convirtieron en la base de varios proyectos (siendo Hadoop MapReduce un proyecto más). Aparecieron también bibliotecas especializadas que se ejecutan sobre Hadoop MapReduce, como Apache Mahout para aprendizaje automático o Apache Giraph para procesamiento de grafos. HDFS sigue siendo el sistema de referencia para almacenamiento distribuido. Han aparecido formatos compatibles que permiten aprovechar sus ventajas a la vez que ofrecen optimizaciones para datos tabulares (Apache Parquet) o para serialización de datos en memoria (Apache Avro). Completan el ecosistema proyectos centrados en el procesado de flujos continuos de datos (o *streams)* como Apache Kafka, Apache Flume o Apache Flink.

1.5.3 Crítica de MapReduce

A pesar de abrir las puertas a la computación distribuida, Hadoop ha sido criticado, entre otros, por Zaharia et al. 2009. En este artículos los autores alaban las bondades de Hadoop, como la alta disponibilidad, la coubicación de datos y la paralelización automática, a la vez que listan las que consideran como principales deficiencias:

- No tiene soporte para los trabajos iterativos típicos de algoritmos de aprendizaje automático. Si el algoritmo lo necesita, el programador debe encargarse de expresar cada iteración como funciones *map/reduce* y de coordinar el arranque de cada trabajo, indicando como datos de entrada los datos de salida de la fase anterior. Además de la falta de esta coordinación global, la alta disponibilidad de Hadoop impacta de manera negativa en el rendimiento, ya que los datos intermedios se persisten a disco. Los accesos a disco tienen una latencia mucho más alta que los acceso a memoria.
- No está preparado para consultas interactivas. Aunque hemos visto que es posible consultar datos en HDFS con sentencias SQL gracias a Hive o HBase, estas herramientas siguen el modelo de Hadoop: leen de disco, ejecutan las fases de *map* y de *reduce*, y vuelven a escribir a disco. Esta latencia impide que se puedan usar para análisis exploratorios en los que el analista ejecuta muchas consultas seguidas, habitualmente cambiando los parámetros de una a otra.

1.5.4 Spark

Tras iniciar la discusión anterior, los autores pasaron a proponer Spark: una infraestructura de computación de código abierto que combina un motor de coordinación de trabajos en un clúster, un modelo de programación unificado y un conjunto de librerías (Sandy Ryza, 2017, White 2011). Spark introduce dos conceptos principales:

- Los *resilient distributed datasets* o *RDDs*, datos distribuidos resistentes a fallos: un RDD es una colección de objetos distribuida y de sólo lectura. Spark es agnóstico en cuanto a los objetos de dicha colección y da libertad al usuario para expresar las operaciones necesarias sobre los mismos. Las operaciones que ofrece por defecto están inspiradas en programación funcional: map, filter, flatMap, reduceByKey, etc. Cada RDD está particionada en memoria de una manera análoga a los bloques de ficheros en HDFS.
- Un *directed acyclic graph* o *DAG*, grafo dirigido acíclico: un conjunto de operaciones encadenadas sobre

RDD. Spark aplica operaciones de manera *perezosa*: sólo inicia los cálculos en el momento en que se solicita la salida. Si hay varias operaciones de filtrado y transformación y sólo una de imprimir por pantalla al final del programa, Spark genera un único DAG y lo ejecuta al final. Esto permite optimizar las operaciones y reconstruir los resultados a partir de los datos iniciales en caso de fallo.

Ambos conceptos se podrían asimilar al paradigma *map/reduce* y al sistema de ficheros HDFS, con dos diferencias notables. Por un lado, Spark mantiene las RDDs intermedias en memoria, es decir, no escribe datos intermedios a disco, reduciendo así la latencia que criticaban sus autores. Por otro, Spark implementa un motor de optimización capaz de reordenar las operaciones sobre RDDs. Este motor, *Catalyst*, ha ganado importancia a partir de la incorporación de los *Dataframes* y *Datasets* en versiones posteriores de Spark. Estas clases se construyen sobre RDDs, a las que imponen un esquema fijo; es decir, cada objeto de la colección debe ser de una clase concreta. Lo que parece una reducción en la libertad del programador ha permitido que Spark ofrezca muchas más operaciones sobre dataframes y datasets que sobre RDDs. Si el programador es capaz de expresar sus datos como dataframes, ahorrará tiempo programando sus operaciones. Además, Catalyst es capaz de aplicar mejores optimizaciones ya que puede hacer asunciones sobre los tipos de datos que no son posibles sobre RDDs.

Junto a los dataframes y datasets, Spark introdujo *Spark SQL*: una capa de abstracción que permite expresar las operaciones sobre dataframes como sentencias SQL. Como hicieron Hive y HBase con Hadoop, Spark SQL acerca la computación distribuida y las consultas interactivas a equipos con experiencia en bases de datos. Las optimizaciones de Catalyst se basan en un modelo de coste, tal como los optimizadores de motores SQL tradicionales: en base a un plan lógico definido por la consulta se generan varios planes físicos y se escoge el de menor coste. Gracias a la ejecución *perezosa* de Spark, las operaciones se pueden reordenar sin afectar al resultado final. El ejemplo más habitual es un *join* seguido de un filtrado: el optimizador se encargará de evaluar si es más eficiente filtrar antes de combinar tablas. El analista puede escribir la consulta de manera declarativa (es decir, indicar qué debe devolver la consulta) sin plantearse problema de optimización, evitando escribir código imperativo (es decir, no necesita indicar cómo debe ejecutarse la consulta).

Al igual que ocurre con el ecosistema Hadoop, Spark se ha rodeado de librerías de propósito específico como MLlib y GraphX. La primera implementa algoritmos de aprendizaje automático como selección de atributos, clusterización, clasificación y regresión. La segunda ofrece funcionalidades de análisis de grafos.

Estas librerías implementan un número limitado de algoritmos permiten trabajar sobre datos distribuidos, mientras que otras librerías de aprendizaje automático o de análisis de grafos en otros entornos (como scikit-learn o NetworkX para python), con una oferta de algoritmos más amplia, están limitados a trabajar a un único nodo.

2. Objetivos y aportaciones

2.1. Objetivo principal

El objetivo principal del presente trabajo es diseñar una metodología que permita, de forma sistemática, tratar y analizar comunidades en foros online desde la perspectiva del análisis de redes sociales y haciendo uso de técnicas de tratamiento del lenguaje natural.

La metodología estará centrada en analizar individuos (por influencia, problemática, etc.), identificar temas populares y su evolución en función del tiempo, así como analizar las comunidades que puedan surgir. Se trabajará sobre casos de estudio reales y se comprobará la aplicabilidad de estos análisis y los resultados de los mismos. El entregable final será una propuesta de metodología a partir de una generalización de los pasos seguidos en los casos de estudio.

Este objetivo principal se puede concretar en varios objetivos específicos que consisten en:

- Identificar los usuarios más relevantes del foro.
- Identificar las comunidades que forman el foro así como las temáticas que las unen.
- Identificar los temas populares.

Adicionalmente, se plantea como objetivo complementario el dar respuesta técnica a los retos que presentan las redes masivas de los foros más grandes.

2.2. Aportaciones

Como principio fundamental, se busca que -si bien utilizaremos un caso de ejemplo real como apoyo- la metodología propuesta sea generalizable a foros de diferentes temáticas y aplicable a diferentes objetivos de negocio. Esto permitiría una extracción de conocimiento al alcance de pequeños foros a los que este paso les supone una barrera.

Se tratará de identificar las tecnologías necesarias para aplicar la metodología a tanto a redes que se puedan almacenar en memoria como a aquellas para las que esto no sea posible. Se intentará, asimismo, que la tecnología sea capaz de integrar todos los aspectos del análisis: la captura de datos, operaciones de NLP y redes sociales y presentación.

A nivel de negocio, la metodología puede aportar utilidad para:

- Identificar, en un campo de interés, cuáles son los usuarios más relevantes con los que trabajar (por ejemplo, una marca deportiva en un foro de deportes).
- Identificar los temas más relevantes y populares en un cierto momento de tiempo.
- Aprovechar la popularidad cambiante de los temas para dirigir acciones de marketing en diferentes direcciones.
- Identificar las comunidades que se generan alrededor de un producto.

3. Metodología

En el resto del trabajo se plantean una serie de tareas sobre los datos originales (que podrían proceder de cualquier foro online) con la intención de alcanzar los objetivos mencionados en el punto anterior. En primer lugar se desarrollará una metodología general que considerará la estructura habitual de los foros de usuarios para establecer los pasos a seguir. Partiendo de los pasos definidos, se trabajará sobre dos casos de estudio para, iterativamente, refinar la propuesta hasta considerarla generalizable.

3.1. Flujo de trabajo

La metodología propuesta por He et al., 2015 sirve de esquema para el flujo del trabajo de los apartados y . Así, se ha dividido el trabajo en las fases que se resumen a continuación.

1. **Identificación de la estructura del foro** (categorías, subforos, etc.). Es posible que interese analizar sólo una categoría concreta, aplicar diferentes métricas a cada categoría o analizar todos los hilos sin distinguir categorías. También se identificarán hilos que interese descartar, como los de Preguntas Frecuentes (*Frequently Asked Questions* o *FAQs*), los que marcan las reglas o la pauta a seguir por los usuarios que quieran escribir en el foro o aquellos en los que los usuarios se presentan y se dan la bienvenida unos a otros pero no entran en discusiones.
2. **Revisión de la fuente de datos, preparación y limpieza**. El origen será habitualmente la base de datos de la aplicación web, bien sea una aplicación comercial o un desarrollo a medida. En estos casos es necesario identificar el esquema que sigue la base de datos: probablemente mantendrá los hilos en una única tabla o colección, con un mensaje por registro o documento. Para los casos de estudio ha sido necesario recopilar el contenido mediante *web scraping*, lo que ha permitido decidir cómo almacenar los datos (si bien tiene sus propios problemas como se verá más adelante). El esquema incluirá datos habituales en todos los foros como el título del foro/subforo, el título del hilo, el autor, la fecha y el texto del mensaje, pero puede incluir información adicional: si incluye una cita a otro mensaje, si tiene imágenes o vídeos embebidos, etc.
3. **Extracción de temas**. En este paso hay que implementar una cadena NLP y decidir la técnica para extraer temas a partir del texto de los mensajes. Los pasos típicos de la cadena NLP serán *tokenización*, descarte de *stopwords* y vectorización. Los datos de salida de la cadena alimentarán el algoritmo de

extracción de temas, que será necesario optimizar.

4. **Construcción de grafos**. La colección de hilos y mensajes, junto con los temas, servirán de base para construir uno o varios grafos en función de la manera en la que identifiquen las relaciones entre usuarios. Por ejemplo, un grafo unipartito que relacione cada usuario que responde en un hilo con el usuario que lo inicia, o un grafo bipartito que relacione cada usuario con el tema del hilo.
5. **Cálculo de métricas de red**. Las métricas habituales de red se puede aplicar directamente sobre grafos unipartitos. Los grafos bipartitos requieren de una fase de aplanado.
6. **Identificación de comunidades**.
7. **Identificación de usuarios relevantes**.
8. **Validación**. La identificación de temas se puede validar revisando los textos de varios hilos de cada tema, mientras que el ranking de los usuarios se puede comparar con las etiquetas que les asigna el foro.

Otros pasos deseables podrían ser el análisis temporal de una temática concreta, la evolución de la influencia de los usuarios, la integración del análisis de sentimiento acerca de un tema o incluso la extracción automática de información, como los detalles de un producto en un foro de compraventa.

Tras aplicar los pasos anteriores en los dos casos de estudio, se analizarán las dificultades, diferencias y puntos en común de ambos casos, con el objetivo de proponer una metodología general que cumpla con los objetivos propuestos en el apartado .

3.2. Casos de estudio

Los dos casos de estudio que se han elegido son los foros de Planet Virtual Boy y AtariAge. Ambos portales son puntos de encuentro de aficionados a los videojuegos y sistemas *retro*, descritos brevemente a continuación. La elección de estos ejemplos concretos se basa principalmente en la temática y el tamaño. PlanetVB se centra en un único producto, por lo que es de esperar que el análisis de temas no se complique demasiado, mientras que en AtariAge hay discusiones de muchos productos distintos que es probable que hagan crecer el número de temas. Por otro lado, PlanetVB tiene un tamaño relativamente pequeño que probablemente admita todo el tratamiento en un único nodo, mientras que AtariAge acumular más de 3.3 millones de mensajes, lo que

añadirá complejidad al trabajo y permitirá evaluar la validez de la metodología para grandes volúmenes de datos. Se consideraron otros foros del sector como VectorGaming y AmiBay, pero los dos foros seleccionados eran suficientemente significativos en cuanto a temática y tamaño.

Sobre cada uno de ellos se aplicarán varias iteraciones para refinar la metodología:

- aplicación de métricas y algoritmos NLP,
- descripción de resultados,
- identificación de problemas,
- planteamiento de mejoras para la siguiente iteración.

Finalmente, en base a las conclusiones de cada uno de los casos, se generalizará el planteamiento para que sea aplicable a foros de otras temáticas y de diversos tamaños.

3.2.1. Planet VB

La Virtual Boy fue una videoconsola de Nintendo presentada en 1994 y lanzada al mercado japonés y americano (nunca llegó de forma oficial a Europa) en verano de 1995. En su corta vida (fue descatalogada en 1996) se calcula que vendió unas 770.000 unidades, aunque no existen cifras oficiales. Las que para otra compañía podrían haber sido cifras respetables eran un fracaso para Nintendo, que decidió enterrarla silenciosamente y que, a menudo, se olvida completamente de ella en su propia historia. Cuatro años más tarde, en el 2000, una pequeña comunidad de aficionados empezó con Planet VB para "construir un memorial a esta rareza de la historia de los videojuegos" (www.planetvb.com). Aunque su fundador sea alemán (y muchos de sus miembros europeos), el foro se puede considerar punto de encuentro internacional y el idioma usado en sus hilos pasó rápidamente del alemán al inglés.

A nivel general, el sitio incluye no solo los foros, sino que también incorpora noticias relacionadas con la máquina, listas completas de juegos y documentación varia sobre la Virtual Boy y su historia, como capturas de revistas de la época. Definen su misión como la de "ofrecer a los fans, coleccionistas y desarrolladores de Virtual Boy una plataforma dedicada a su pasión". La estructura del sitio refleja esta subdivisión aparente entre tipos de usuarios, como aficionados que expresan su aprecio por la consola, desarrolladores que

comparten sus proyectos y coleccionistas que entran en dinámicas de compra-venta de artículos. Así, se encuentra un foro general en el que tienen cabida muchos temas, pero también subforos más específicos para la compra-venta (*marketplace*) y para el desarrollo (*homebrew*).

PlanetVB es un foro de reducido tamaño, con algo más de 3.000 usuarios registrados (aunque solo la mitad sean realmente activos) y unos 40.000 mensajes, que permitirá ejecutar el flujo múltiples veces en un caso manejable en prácticamente cualquier entorno de procesamiento.

3.2.2. AtariAge

AtariAge (http://atariage.com/) es otro punto de encuentro para aficionados a los videojuegos y máquinas clásicas, pero esta vez con origen en Estados Unidos. Nacido en 1998 como rincón de fans de la Atari 2600, es uno de los portales más longevos del sector. Su éxito (dada la gran cantidad de usuarios de la Atari 2600 en el mundo) lo llevó a reconvertirse en AtariAge en 2001, un portal más generalista dedicado al *retro* en todo su espectro. Las máquinas de Atari (desde la 2600 hasta la Jaguar, pasando también por los microordenadores de la casa americana) tienen un protagonismo especial en sus foros, con secciones específicas dedicadas a cada una de ellas, pero también se encuentra una cantidad ingente de mensajes e hilos dedicados a otras máquinas populares.

Como en el caso de PlanetVB, el sitio incluye también otros elementos aparte de los foros, como documentación, descargas o tienda de productos. De la estructura de los foros también se intuye la posibilidad de encontrar temas o perfiles de usuarios distintos, con una sección dedicada a los sistemas Atari, otra a "jugar en general" (con hilos de otros sistemas pero también sobre emulación y similares), una parte de compra-venta (*marketplace*), desarrollo (*game programming*), comunidad e incluso noticias.

En comparación con PlanetVB, AtariAge es un foro mucho más masivo, con más de 3,3 millones de mensajes (dos órdenes de magnitud mayor) y más de 20.000 usuarios activos, lo que requerirá el procesamiento distribuido para poder ejecutar ciertos algoritmos.

4. Propuesta inicial

Los pasos descritos en el apartado necesitan una definición sobre unos algoritmos y una plataforma concretos. Se ha optado por Spark porque a) propone un modelo de programación unificado (ver apartado) y b) puede escalar horizontalmente sin cambios en el código. En un estudio preliminar se identificaron funcionalidades necesarias para el flujo de trabajo: la biblioteca MLlib dispone de utilidades de NLP como tokenización y eliminación de stopwords y una implementación de LDA para detección de temas, y la biblioteca GraphFrames está preparada para análisis con redes con diversas métricas y algoritmos de centralidad y de identificación de comunidades. La elección de una plataforma concreta limita la implementación del código a la misma, pero es necesaria para poder trabajar con un caso de estudio real. No obstante, el objetivo es proponer una metodología general, de manera que incluso sea posible sustituir unos algoritmos por otros más acordes al caso de uso.

4.1. Temas

La extracción de temas seguirá el modelo de cadena NLP comentado en . El primero paso de toda cadena NLP es definir sobre qué documentos va a actuar. En el caso de un foro hay dos alternativas: considerar cada mensaje como un documento, o considerar el texto agrupado de todos los mensaje de un hilo como un único documento. Esta segunda opción requiere un paso adicional de concatenación del texto de los mensajes agrupados por hilos.

La cadena como tal constará de fases de tokenización, limpieza de términos y descarte de *stopwords.* Los mensajes de los casos de estudio están escritos principalmente en inglés, así que no se incluirá ningún paso de identificación de lenguaje. No es necesario una fase de análisis sintáctico, así que la tokenización se limitará a términos, no a frases y términos como es habitual en una cadena NLP general. A continuación, una fase de eliminación de *stopwords* descartará palabras comunes que pueden añadir ruido a la identificación de temas sin aportar información: artículos, pronombres, preposiciones, números y palabras habituales como *are, much* o *someone*. El último paso será la vectorización de los textos: se construye una matriz con tantas

columnas como palabras en todos los mensajes del corpus. El valor de cada columna en cada mensaje será el número de veces que aparece dicho término en ese mensaje.

La cadena NLP estará conectada al bloque de identificación de temas propiamente dicho. En la introducción teórica se explicaron dos aproximaciones a la identificación de temas:

- Etiquetado manual y aplicación de un algoritmo de clustering.
- Extracción automática mediante LDA y revisión manual de términos.

El procedimiento manual deja de ser efectivo si aparecen temas nuevos que no han aparecido durante el proceso de etiquetado, si bien es útil en otras situaciones (por ejemplo, para diferenciar los mensajes en *petición* y *respuesta*). Por tanto, en este caso se ha decidido aplicar una extracción automática. No obstante, tras aplicar LDA se analizará cada tema (que no es más que un conjunto de términos relevantes comunes a todos los documentos asociados a dicho tema) y se le asignará una etiqueta o título. Este paso hace posible referirse a un tema con un nombre concreto en lugar de un identificador numérico, y permite agrupar temas similares. Este etiquetado manual de unas pocas decenas de temas es comparativamente mucho menos costoso que el etiquetado de varios cientos o miles de mensajes.

4.2. Análisis del grafo de usuarios

4.2.1. Construcción

La construcción de grafos se realizará utilizando la librería *GraphFrames* de *Spark*, que extiende la librería nativa *GraphX* y permite utilizar *dataframes* para definir los enlaces y vértices, aportando también algunas funcionalidades extra.

Los pasos que se seguirán para construir el grafo que representa las relaciones de usuarios en cada uno de los foros se describen a continuación.

1. Se agrupan los posts por hilo y se ordenan por fecha para obtener una secuencia de mensajes.
2. Dentro de cada hilo, se considera que todos los participantes contestan al usuario que inicia el hilo. Por tanto, se generan tantos enlaces como usuarios hayan participado en el hilo y se etiquetan con el

author	**thread_code**	**...**	**quoted_user**	**topic_id**
u1	1234	...		5
u2	1234	...		5
u3	1234	...	u2	5
u4	1234	...		5
u1	1234	...		5
u3	1234	...	u1	5

Table 1: Sucesión de posts en un hilo

src	dst	topic_id	weight	relationship
u2	u1	5	1	replies_to
u3	u1	5	2	replies_to
u4	u1	5	1	replies_to
u3	u2	5	1	quotes
u3	u1	5	1	quotes

Table 2: Lista de enlaces generados

atributo *relationship = replies_to.*

3. Se añaden además enlaces para las ocasiones en las que un usuario cita a otro dentro del mismo hilo, aplicando la etiqueta *relationship = quotes.*
4. Se añade otro atributo *topic_id* para indicar el tema principal del foro.
5. Se eliminan los enlaces que tengan como origen y destino el mismo usuario.
6. Se genera el grafo utilizando a partir de los enlaces.

En las Tablas 1 y 2 se puede observar un ejemplo de mensajes en un hilo y la lista de enlaces entre usuarios que se obtendrían tras aplicar este proceso, en el que se ha generado un atributo *weight* que representa el peso del enlace. Finalmente, en la Figura 6 se muestra el grafo resultante de representar gráficamente dichos enlaces.

4.2.2. Usuarios relevantes

De cara a identificar los usuarios relevantes en el grafo se aplicará PageRank (Page et al. 1999). Se trata de un algoritmo iterativo que se aplica de la siguiente manera:

- En la primera iteración ($t = 0$) se asigna un peso inicial a cada nodo. Este peso habitualmente se fija en función del número total de nodos que existen en el grafo, N; típicamente, $\frac{1}{N}$.
- En las siguientes iteraciones, el valor de PageRank de cada nodo se actualiza en función del valor de

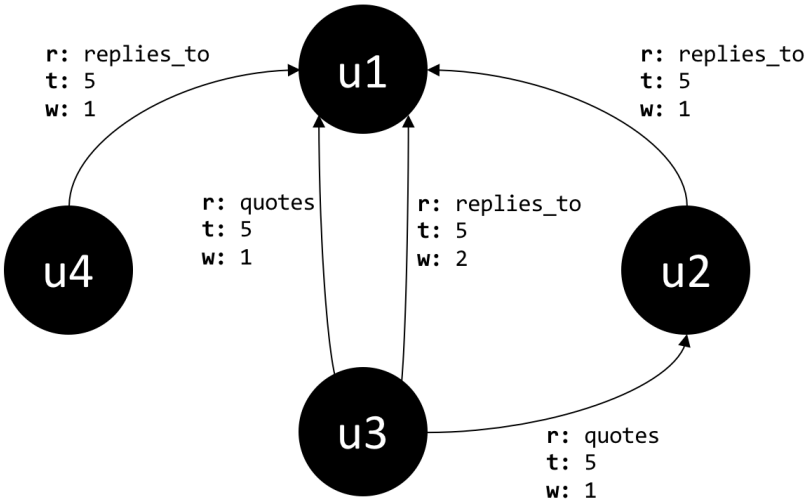


Figure 6: Generación del grafo a partir de un hilo

	u1	u2	u3	u4
t=0	0.25	0.25	0.25	0.25
t=1	0.125	0.5	0.25	0.125
t=2	0.0625	0.375	0.3125	0.25
t=3	0.125	0.375	0.3125	0.1875
t=4	0.09375	0.4375	0.28125	0.1875
t=5	0.09375	0.375	0.3125	0.21875
t=6	0.109375	0.40625	0.296875	0.1875
...	...	...	...	...
t=[?]	0.1	0.4	0.3	0.2

Table 3: Pesos calculados para cada iteración de PageRank

PageRank de los nodos con los que tiene un enlace entrante. Se aplica también un factor de amortiguación d que limita el crecimiento del valor de PageRank para los nodos sin enlaces.

Se puede ver un ejemplo de la aplicación de este algoritmo sobre un grafo sencillo en la figura 7 junto con el cálculo detallado para las primeras iteraciones en la tabla 3. En este ejemplo se ha utilizado una simplificación del algoritmo que no tiene en cuenta el factor de amortiguación. A medida que se realizan iteraciones se puede comprobar que el valor del PageRank de cada nodo tiende a estabilizarse en torno a un valor concreto, siendo los nodos que tienen un PageRank más elevado los más relevantes. Se podrá decidir cuántas iteraciones se realizarán siguiendo estos criterios bien fijando un número fijo de iteraciones t o bien fijando una tolerancia ϵ tal que la diferencia de los valores entre iteraciones sea menor que dicha tolerancia.

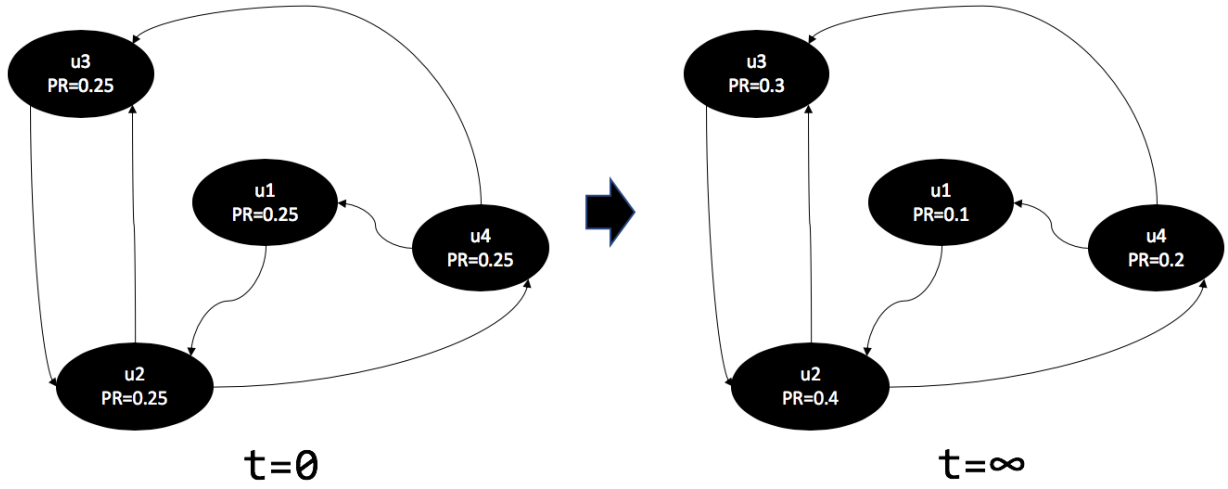


Figure 7: Ejemplo de PageRank

El algoritmo HITS - Hyperlink-Induced Topic Search (Kleinberg 1999) - proporciona otra métrica adicional. Se trata también de un algoritmo iterativo, pero se diferencia de Pagerank en que permite medir la importancia de los nodos en dos aspectos. En primer lugar, identifica los nodos que actúan como *hubs,* es decir, usuarios que son referenciados por muchos usuarios relevantes. En estos casos de estudio un *hub* corresponde con aquél nodo que es frecuentemente citado o cuyos posts reciben multitud de respuestas. En segundo lugar identifican las *authorities*, que corresponden con usuarios que tienen multitud de enlaces con *hubs* diferentes. HITS no garantiza la convergencia de los resultados, así que es necesario limitar el número de iteraciones con un valor fijo o con un umbral de diferencia mínima entre iteraciones, al igual que PageRank.

En este caso, cada iteración del algoritmo actualiza el valor de *hub* y *authority* de cada nodo de la siguiente manera:

- El valor de *authority* será la suma de los valores de *hub* de todos los nodos que apuntan a ese nodo.
- El valor de *hub* será la suma de los valores de *authority* de todos los nodos que apuntan a ese nodo.

En los casos de estudio se usará PageRank como métrica para obtener un ranking de usuarios influyentes, ya que, aparte de métricas directas como el grado, *GraphFrames* sólo dispone de este algoritmo. Aunque existen implementaciones de PageRank que sí que permiten tener en cuenta el peso de los enlaces (como por ejemplo, en NetworkX), en GraphFrames no existe esta posibilidad.

En la metodología propuesta se calculará PageRank sobre GraphFrames y se validarán los resultados obtenidos de la siguiente manera:

- Se comparará el PageRank obtenidos con GraphFrames con el obtenido con NetworkX.
- Se aplicará el algoritmo HITS en NetworkX y se comparará este ranking con el generado a partir de PageRank.

4.2.3. Detección de comunidades

Una red tiene una estructura de comunidades si sus nodos pueden ser fácilmente agrupables en conjuntos de nodos tal que cada grupo tiene una densidad de enlaces internos elevado. Si se considera que cada nodo pertenece a un único grupo o comunidad, esto significa que no tan solo las conexiones internas son densas, sino que también las conexiones externas (con otros grupos) deben ser débiles. La identificación de comunidades se basa en un principio muy simple: dos nodos tienen más probabilidades de estar conectados si ambos pertenecen a la misma comunidad y menos si no comparten comunidades.

La detección de comunidades es muy relevante para el análisis de redes sociales (al fin y al cabo intenta identificar estructuras subyacentes en la red) pero, por otro lado, presenta diversos problemas:

- Es computacionalmente compleja.
- El número de comunidades (si es que las hay) raramente se conoce de antemano.
- Los grupos son habitualmente heterogéneos y poco balanceados, es decir, suelen existir comunidades muy grandes que conviven con otras muy pequeñas.

Afortunadamente, existen múltiples algoritmos desarrollados para la detección de comunidades (Radicchi et al., 2004). Aquí solo se van a detallar tres de ellos, porque son los que se encuentran implementados en las herramientas que se van a utilizar: LPA (Raghavan et al., 2007), Girvan-Newman (Newman and Girvan, 2004) y Louvain (Blondel et al., 2008).

4.2.3.1 LPA (Label Propagation Algorithm)

El algoritmo LPA se guía únicamente con la estructura de la red, así que no necesita ningún parámetro externo. Cada nodo decide si pertenece o no a una comunidad en función de las etiquetas asignadas a sus vecinos inmediatos (Raghavan et al., 2007). Su funcionamiento es el siguiente:

1. Se asignan etiquetas únicas a cada uno de los nodos.
2. En cada paso, se ordenan los nodos de forma aleatoria y se tratan secuencialmente.
3. Para cada nodo, su nueva etiqueta pasa a ser la que tenga la mayoría de sus vecinos.
4. Y si cada nodo tiene una etiqueta que coincide con el máximo de las etiquetas de sus vecinos, termina el algoritmo. En caso contrario, se vuelve al paso 2.

Este algoritmo tiene su principal ventaja en la velocidad, ya que se ejecuta en tiempo próximo a lineal; su inconveniente es que no tiene una solución única y que, además, no produce resultados estables (es decir, cambian notablemente entre ejecución y ejecución). La librería *GraphFrames* de *Spark* implementa LPA como algoritmo de detección de comunidades.

4.2.3.2 Girvan-Newman

El algoritmo de Girvan-Newman es uno de los más utilizados y extendidos por la calidad de sus resultados y porque se encuentra implementado en multitud de herramientas y programas, como es el caso de NetworkX. Este algoritmo identifica las aristas que se sitúan entre comunidades y las elimina, dejando únicamente la estructura de comunidades aisladas. Esta identificación se hace a través de la medida de la centralidad de intermediación, que como se ha visto anteriormente tiene un valor elevado cuando los nodos se encuentran en la frontera de las comunidades.

El problema principal del algoritmo de Girvan-Newman es que, para grafos grandes, sufre en tiempo de ejecución, puesto que la complejidad es de orden $O\left(m^2 n\right)$, siendo n el número de nodos y m el de aristas, así que es poco práctico para redes que tengan más allá de unos miles de nodos (Newman and Girvan, 2004).

4.2.3.3 Louvain

A medio camino entre la relativa ingenuidad del LPA y la complejidad de Girvan-Newman se encuentran los algoritmos basados en el concepto de la modularidad. La modularidad es una función beneficio que mide cuan buena es una partición particular de los nodos en comunidades. La idea es que es capaz de medir la relación entre enlaces internos (intracomunitarios) y externos (extracomunitarios). Los algoritmos basados en la modularidad se basan en buscar la forma de maximizarla entre todas las particiones posibles. Esta es una tarea que no se puede abarcar de forma racional, así que los algoritmos se centran en métodos de optimización y cada uno de ellos balancea de forma distinta la velocidad y la exactitud. Quizás el más popular de todos ellos es el Louvain (Blondel et al., 2008), que optimiza de forma iterativa las comunidades locales hasta que no hay forma de mejorar la modularidad mediante perturbaciones sobre el estado actual.

Este balance tiene un coste, no obstante; los métodos que se basan en la modularidad presentan problemas al detectar comunidades menores que cierta escala (en función del tamaño de la red) (Sun, 2014); es el llamado límite de resolución. La herramienta de visualización Gephi se basa en Louvain y la librería Community API para NetworkX también lo implementa.

4.3. Análisis del grafo usuarios-temas

4.3.1. Construcción

La construcción del grafo que relaciona los usuarios con los temas sobre los que escriben se ha realizado en *NetworkX*. Para cada usuario U, una vez identificados los temas T de sus mensajes, se crea una arista entre U y T de peso w_{UT}, donde el peso es el número de mensajes escritos en ese tema en particular. El grafo resultante es bipartido a dos niveles: el superior, los temas, y el inferior, los usuarios.

Se muestra la construcción del grafo con los datos de la tabla 4 en la figura .

Usuario	Tema	Número de mensajes (peso)
u1	t1	3
u1	t2	7
u1	t3	2
u2	t1	2
u2	t2	4
u3	t1	3
u3	t3	2

Table 4: Ejemplo de construcción del grafo bipartido de usuarios-temas

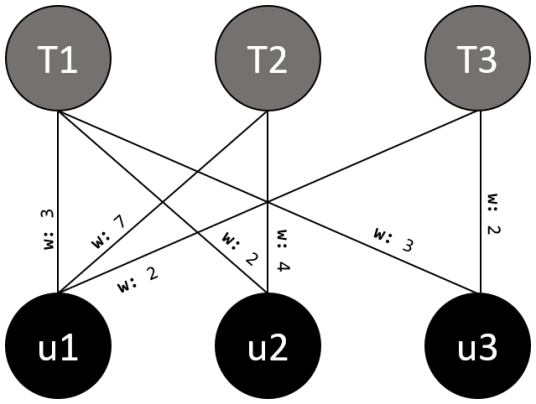


Figure 8: Ejemplo de construcción del grafo usuarios-temas. El usuario 2, por ejemplo, ha escrito 2 mensajes del tema 1 y 4 del tema 2.

4.3.2. Proyección

Tal y como se ha desarrollado en el apartado 1.3.3.1, los grafos bipartidos pueden proyectarse (o aplanarse) para obtener dos redes nuevas a partir de la original, con la ventaja que a ambas redes unimodales se les pueden aplicar los algoritmos más habituales. Como se ha visto anteriormente, la metodología más adecuada para proyectar una red de mensajes de foros de usuarios es la que tiene en cuenta los pesos de la red original. Así, para esta propuesta, se procede a utilizar el principio del método de Newman (2001b) teniendo en cuenta los pesos de la red. En la figura 3 se ha visto el funcionamiento de dicha proyección, pero los algoritmos de detección de comunidades no funcionan (al menos sin modificación) para grafos dirigidos, así que en este caso se va a implementar la variación que se verá a continuación.

Cada par de usuarios (u, v) coincide en un tema T_i si existen, a la vez, las aristas $(u,\ T_i)$ y $(v,\ T_i)$. En caso de coincidir pueden, además, coincidir en más de un tema. Así, podemos definir la intensidad de esa relación

en función del número de mensajes que cada uno de los usuarios ha realizado con ese tema con la suma de sus pesos, $w_{uT_i} + w_{vT_i}$. Hasta aquí tendríamos una proyección no dirigida parecida a una versión no dirigida de lo propuesto en la figura 3b. No obstante, en la mayoría de foros hay hilos o temas más populares (con multitud de usuarios coincidiendo en ellos) mientras otros quedan dominados por comunidades más pequeñas. Así, se considera también relevante incluir la ponderación que propone Newman, incluyendo la división por el número total de usuarios que han escrito en un mismo tema. La ecuación final para determinar el peso de las relaciones queda, pues, de la siguiente forma:

$$w_{uv} = \Sigma_T \frac{w_{uT} + w_{vT}}{N_T - 1}$$

La ecuación se puede leer de la siguiente forma: el peso entre los nodos (usuarios) u y v es igual al sumatorio, para todos los temas en los que coinciden, de los pesos de sus aristas con los respectivos temas dividido por el número de usuarios (menos uno) que han colgado algún mensaje en dicho tema.

NetworkX implementa varios algoritmos de proyección, entre los que se encuentra la proyección ponderada de Newman para el caso de grafos sin peso (collaboration_weighted_projected_graph). Este algoritmo, no obstante, no tiene en cuenta los pesos de la red original; en su lugar opta por asumir todos los enlaces como de peso 1 si están presentes y de peso 0 si no lo están. Como el algoritmo descrito anteriormente no se encuentra disponible, se procede a modificar el primero para que tenga los pesos en consideración.

Como referencia, el cambio supone sustituir las siguientes dos líneas del código fuente original de la implementación de la proyección con colaboración en *NetworkX*:

```
common_degree = (len(B[n]) for n in unbrs & vnbrs)
weight = sum(1.0 / (deg - 1) for deg in common_degree if deg > 1)
```

Por las que se muestran a continuación:

```
common_degree = ((len(B[n]), B[u][n]['weight'] + B[v][n]['weight']) for n in unbrs & vnbrs)
weight = sum(wt / (deg - 1) for deg, wt in common_degree if deg > 1)
```

Así, el ejemplo mostrado en la figura 8 quedaría proyectado sobre los usuarios de la siguiente forma (figura 9):

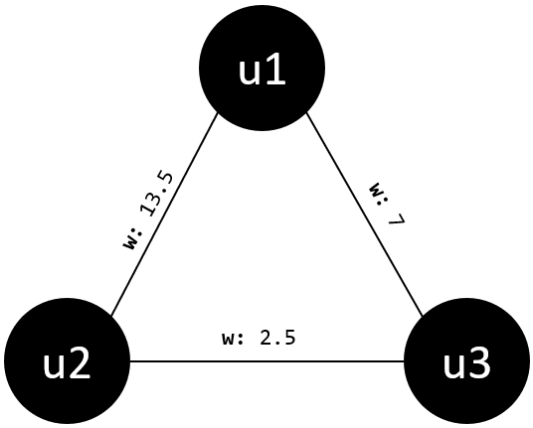


Figure 9: Aplanado del grafo de la figura 8 utilizando el algoritmo definido en .

En ella se puede apreciar como, aunque *u2* y *u3* presenten perfiles parecidos (un mensaje en un tema compartido con *u1* y un mensaje en un tema compartido con los otros dos), la relación entre *u1* y *u2* es mucho más fuerte que entre *u1* y *u3*. Eso es debido a que tanto *u1* como *u2* han escrito mucho en un tema (*t2*) en el que no aparece *u3*; así, sus opciones de haber establecido conexión son mucho más altas que si se conectan por un tema en el que han escrito comparativamente poco (*t3*). El algoritmo Girvan-Newman sin considerar el peso inicial del grafo hubiese dado la misma relevancia a ambas conexiones, algo que en el contexto de los foros, al menos, no parece ser la mejor aproximación.

4.3.3. Comunidades de interés

Una vez proyectado el grafo bipartido, las comunidades de interés entre los usuarios se detectan aplicando los mismos algoritmos que los descritos en el apartado . En el caso concreto de *NetworkX*, el algoritmo implementado por defecto es Girvan-Newman, aunque en este trabajo se utilizará el método *Louvain* implementado en el paquete python-louvain.

5. Casos de estudio

5.1. Obtención de los datos

Para obtener los datos de ambos foros se ha seguido un proceso de *web scrapping,* que consiste en recopilar información de forma automática de la Web. Hay que tener en cuenta que si un administrador de un foro desease ejecutar la metodología para estudiar su comunidad podría hacerlo más fácilmente mediante una descarga de su propia base de datos. Al tratarse de foros públicos pero ajenos, no obstante, hemos tenido que recurrir a la programación de arañas en *Scrapy* para extraer la información requerida en estos casos de estudio.

5.1.1. Proceso

El funcionamiento de las arañas es el siguiente:

1. Se empieza por un sitio (o una lista de sitios) a visitar, considerado la semilla.
2. La araña extrae la información deseada utilizando selectores de información que recogen, por ejemplo, nombres de usuario y títulos de foros.
3. A su vez, identifica los nuevos enlaces de la página y los añade a la cola de URLs a visitar mediante una nueva llamada a la araña.
4. Así, recursivamente, la araña visita todos los enlaces y recupera la información deseada, que se guarda estructuradamente para el análisis posterior.

Es importante destacar también la necesidad de introducir políticas de funcionamiento que eviten saturar el servidor con demasiadas llamadas: es lo que se llama hacer *scrapping* "educado" (*polite*), que aumenta el tiempo de proceso pero disminuye la posibilidad de causar molestias al servicio o a sus usuarios.

5.1.2. Información extraída

El proceso de extracción de información del foro se ejemplifica a continuación. La estructura típica en árbol de los foros online se muestra en forma de esquema en la figura 10. Básicamente se parte un foro que forma la raíz y que se divide en uno o varios temas. Estos temas, a su vez, pueden estar subdivididos en subtemas de uno o varios niveles de profundidad. Así, por ejemplo, PlanetVB no se estructura en subtemas (los temas están directamente formados por hilos), pero AtariAge sí puede tener subniveles (como *Atari 2600 Programming* dentro de *Atari 2600*). Cada tema o subtema, entonces, agrupa a una serie de hilos que son las discusiones que se forman entre usuarios, con un usuario iniciador (el que abre el hilo) y una serie de respuestas al mismo. Cada uno de estos mensajes tiene una fecha, un autor y un contenido.

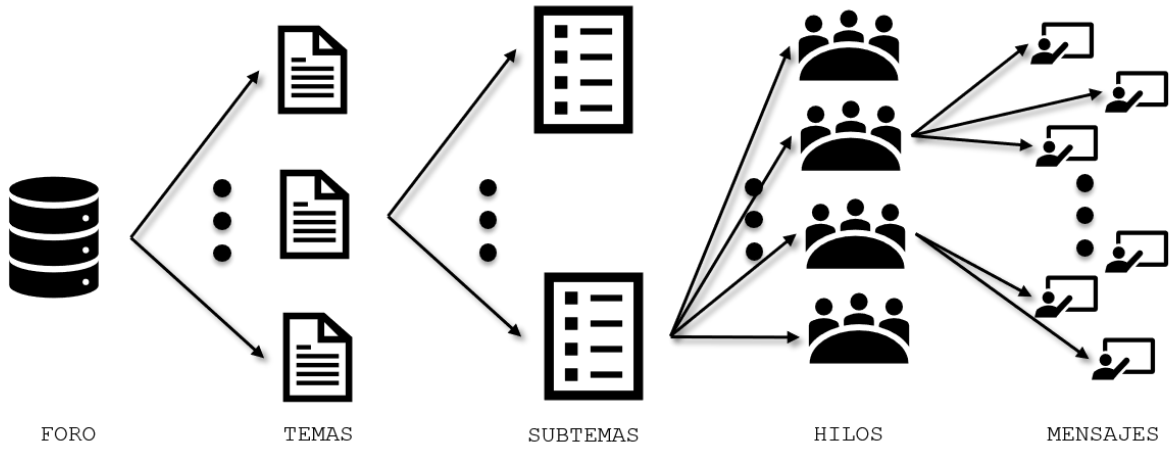


Figure 10: Estructura típica de un foro online, desde los temas a los mensajes en forma de árbol.

En primer lugar, se parte de la página de origen, que es la URL de entrada al foro. En el caso de PlanetVB se trata de la página inicial (`https://www.planetvb.com/modules/newbb/`). En ella se encuentran unas categorías generales que encapsulan los temas y enlaces a los subforos. En la figura 11 se muestra parcialmente dicha página inicial y se señala, en un recuadro rojo, uno de los temas a los que se puede acceder desde ella. Aquí la araña de *Scrapy* captura el título del tema y su enlace y lanza una nueva llamada para procesar el resultado de acceder a él. Solo como ejemplo, se muestra la parte del código que ejecuta esta parte, que es equivalente al que se ejecuta recursivamente en los pasos siguientes.

```
def parse(self, response):
    for line in response.xpath('//div[@id="index_forum"]'): # Se procesa cada tema
```

```
forum_title = line.css('b::text').extract_first() # Se obtiene el título del tema
url_threads = line.xpath('./a/@href').extract_first() # Se obtiene la URL del tema
yield scrapy.Request(url = url_threads,
                     callback = self.parse_sublevel,
                     meta = {'forum_title': forum_title}) # Y se llama de nuevo a la araña para
```

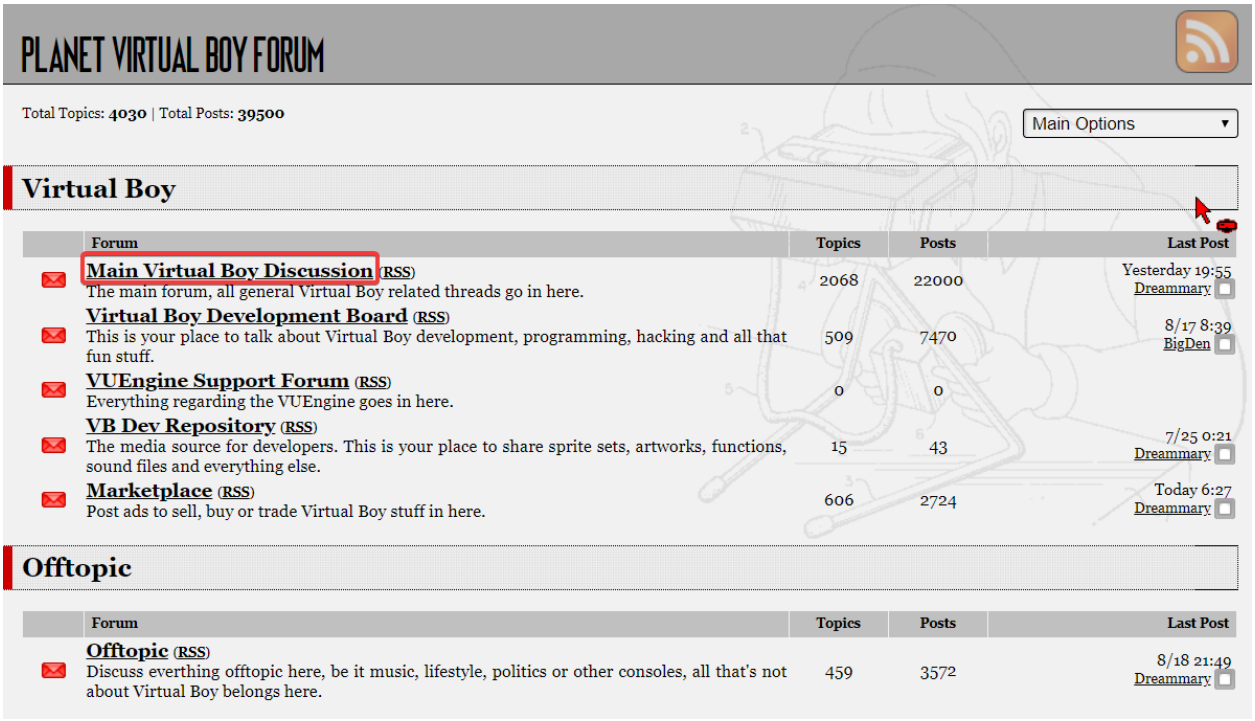


Figure 11: Página inicial y principal del foro PlanetVB, mostrando los temas.

Una vez procesados los títulos, se pasa al siguiente nivel del árbol, que en el caso de PlanetVB corresponde directamente a los hilos que forman cada tema. En la figura 12 se anotan en un recuadro rojo algunos de los hilos a modo de ejemplo. El proceso es muy parecido al paso anterior, aunque con una ligera diferencia. Aquí se anota el título del hilo y la URL a la que habrá que acceder para tratar sus mensajes (1), pero también se considera si hay más páginas de hilos bajo el mismo tema (2). En caso de que las haya, hay que recorrerlas todas, así que la araña procesará la página 1, pondrá en la cola todos sus hilos, se irá a la página dos, hará lo propio, página tres y así sucesivamente hasta llegar a la última.

Cada uno de los hilos está formado por uno o más *posts* como el de la figura 13. Los elementos de interés para la recolección de información se listan a continuación:

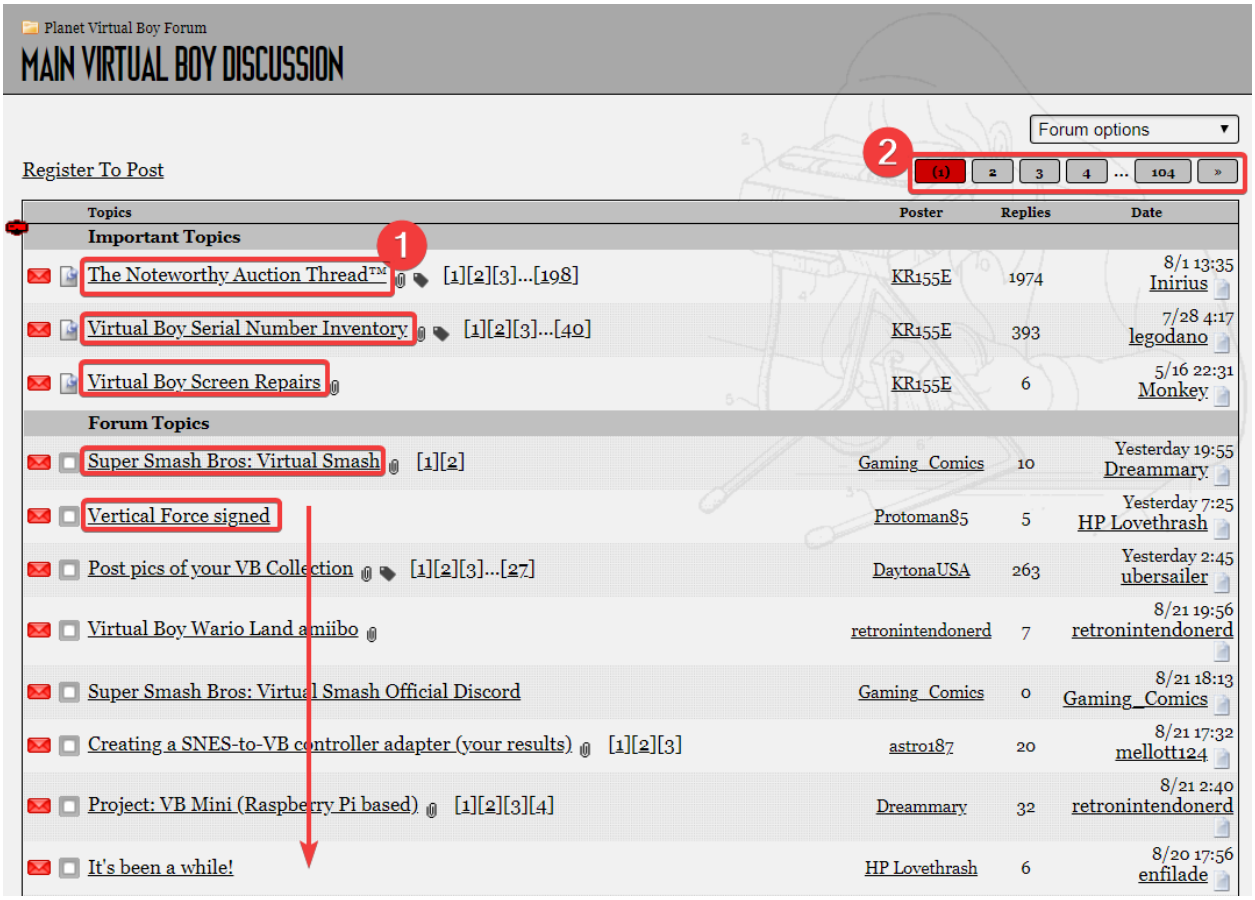


Figure 12: Hilos dentro de uno de los subtemas de PlanetVB. (1) Acceso a los hilos y (2) Paginación de hilos.

1. El título del hilo, que viene de la figura anterior.
2. La fecha del mensaje.
3. El nombre del autor.
4. Su reputación (según lo establecido por la página).
5. El cuerpo del mensaje.
6. Las páginas que forman el conjunto de mensajes del hilo en caso de extenderse más allá de una página.

Toda la información anterior se guarda en formato JSON para su tratamiento posterior, tanto para el trabajo en procesamiento del lenguaje natural como para la construcción de grafos. Adicionalmente, algunos *posts* pueden contener referencias (citas) a mensajes anteriores, como se puede apreciar en la figura 14.

Dichos usuarios se copian a una columna adicional como se muestra en la Tabla 5.

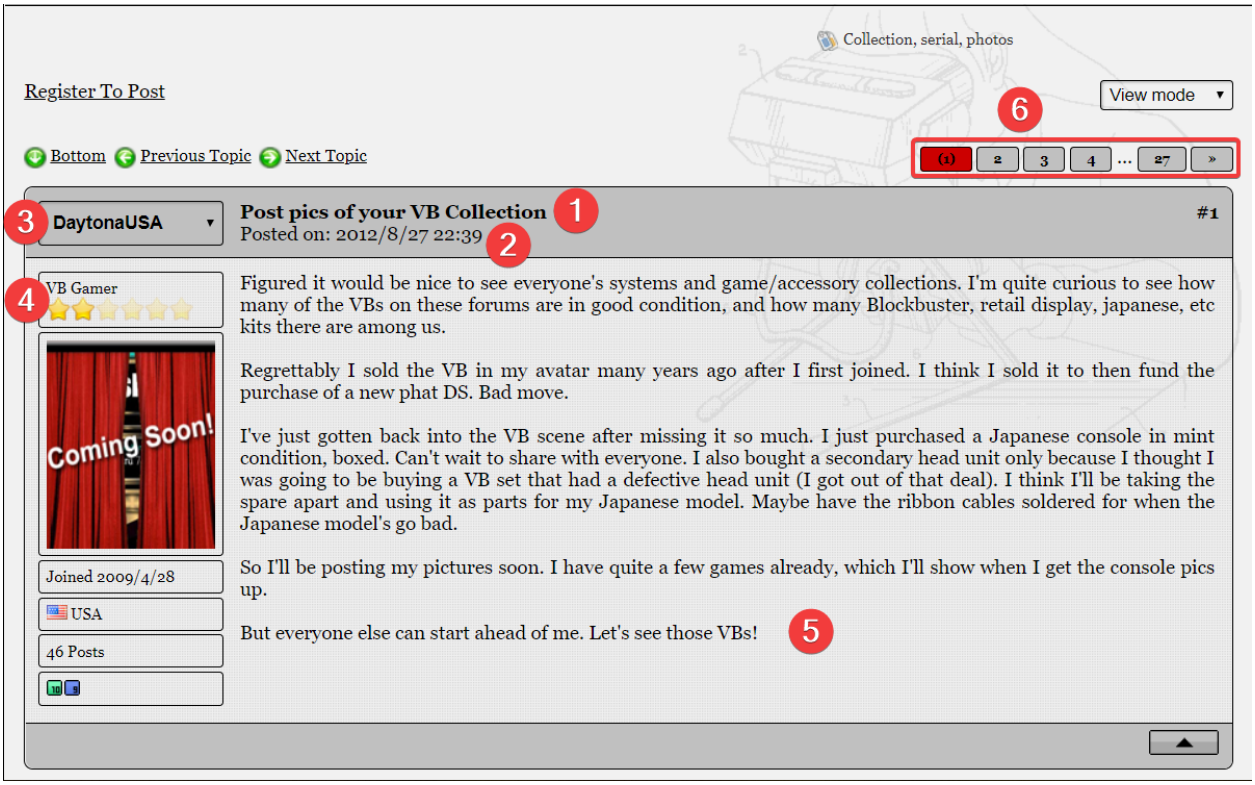


Figure 13: Extracción de los datos de un post dentro de un hilo en PlanetVB.

author	post_date	post_text
retronintendonerd	01/10/2018 15:43	This year I intend on reviewing every...
Dreammary	01/10/2018 18:21	Adventures in the 3rd Dimension
retronintendonerd	01/10/2018 19:36	Quote:Dreammary wrote:Adventures in the 3rd Dimension OMG I friggin love th

Table 5: Columna quoted_user

5.2. Caso de estudio 1: PlanetVB

A lo largo de este caso de estudio se aplicarán los pasos presentados en el apartado . Servirá como banco de pruebas del flujo de trabajo propuesto y como entorno de laboratorio a una escala más reducida de cara al segundo caso de estudio.

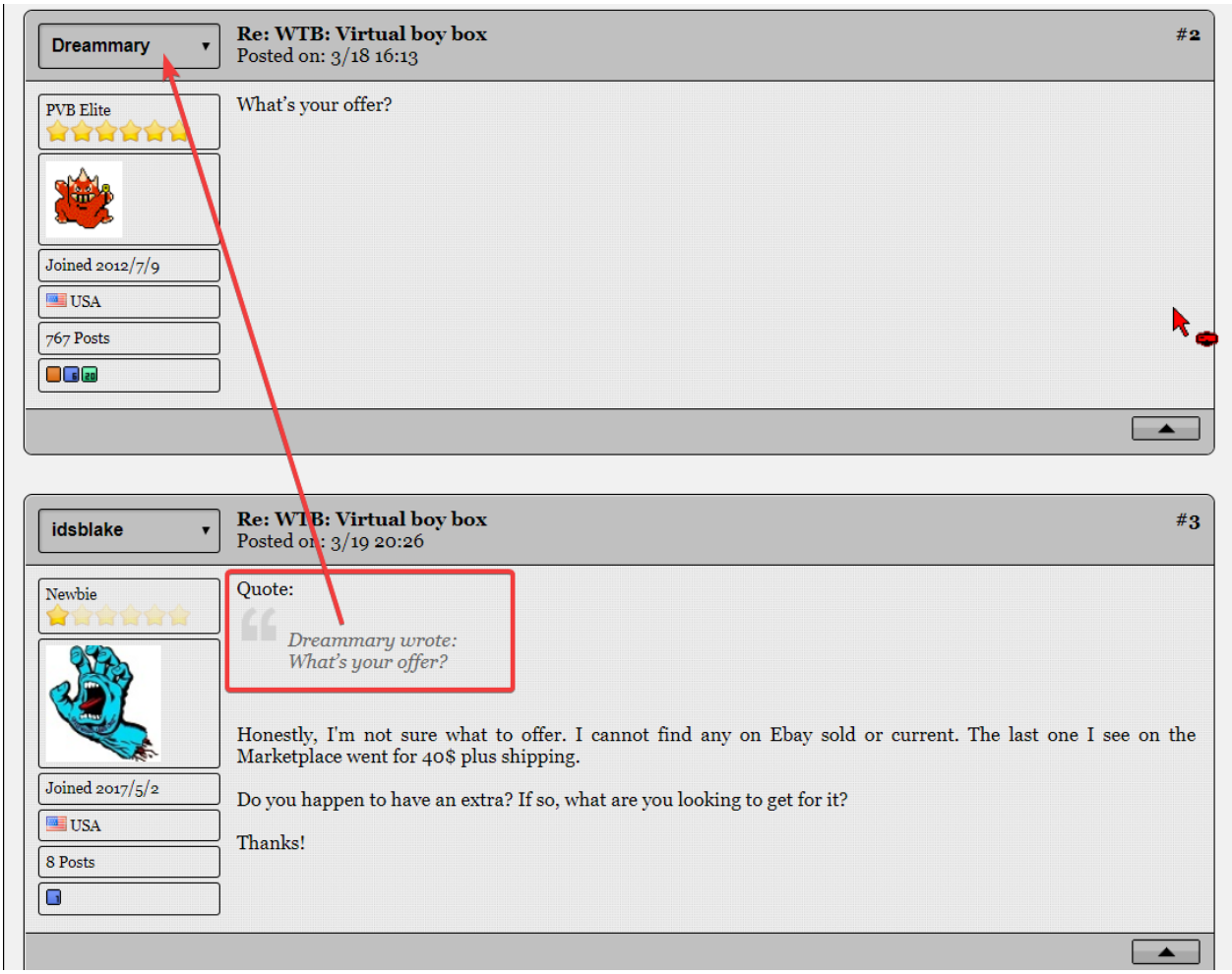


Figure 14: Ejemplo de cita a otro usuario en PlanetVB.

5.2.1 Estructura del foro

En el momento de captura de los datos (enero de 2018) PlanetVB constaba de unos 37.000 mensajes a lo largo de uno 3.800 hilos, repartidos en 10 subforos y escritos por una base de 1.600 autores. Más del 55% de los mensajes se concentran en el subforo principal de Virtual Boy (bajo un título relativamente genérico, *Main Virtual Boy Discussion*) y aproximadamente un 22% en los foros de programación, *Virtual Boy Development Board* y *PVB Coding Competition 2008*, *2010* y *2013* (ver tabla 6).

El tamaño medio de los mensajes está en torno a 400 caracteres, aunque hay auténticos artículos de más de 18.000 caracteres. Hay mensajes con longitud cero que suelen contener sólo imágenes o emoticonos. La longitud de los hilos varía mucho, con una media de 10 mensajes y una desviación típica de 36. El más concurrido se inició en 2003 y actualmente tiene 1879 mensajes. La actividad de los usaurios también es muy

Elemento	Cantidad
Subforos	10
Hilos	3852
Posts	37584
Autores	1618

Subforo	Número de posts	Porcentaje de posts
Main Virtual Boy Discussion	21039	55,8%
Virtual Boy Development Board	7024	18,6%
Offtopic	3416	9,1%
Marketplace	2538	6,7%
FlashBoy	1419	3,8%
Feedback	1233	3,3%
PVB Coding Competition 2013	424	1,1%
PVB Coding Competition 2010	293	0,8%
PVB Coding Competition 2008	246	0,7%
VB Dev Repository	40	0,1%

Table 6: PlanetVB: dimensiones del foro

	Núm. caracteres por mensaje	Núm. mensajes por hilo	Núm. mensajes por autor
Máximo	18172	1879	1969
Media	411	10	23
Desviación típica	612	36	110

Table 7: PlanetVB: estadísticas

desigual: dos de ellos han publicado aproximadamente 1900 mensajes mientras que la media es de 23. Estas estadísticas se reflejan en la tabla 7 y en las gráficas de la figura 15.

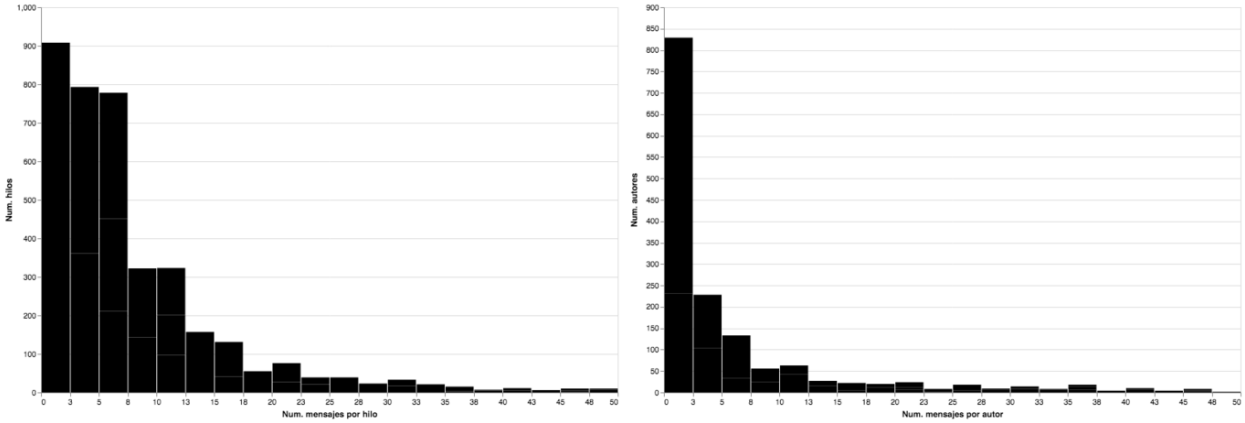


Figure 15: PlanetVB: mensajes por hilo y por usuario (eje x truncado en 50).

Elemento	Cantidad	% eliminado
Hilos	3.358	13%
Posts	36.784	2.5%
Autores	1.540	5%

Table 8: PlanetVB: dimensiones del foro tras la limpieza

5.2.2 Limpieza

Los datos que son objeto de estudio en este trabajo se obtuvieron a través de la tarea de *scrapping* anteriormente citada y requieren cierto procesamiento antes de proceder a realizar un análisis más detallado. A continuación se describen las tareas que han sido necesarias para estandarizar y corregir los datos.

Limpieza de hilos

Se han eliminado del conjunto de datos los siguientes hilos:

- Aquellos hilos fijos con contenido estático como las normas del foro o un inventario de números de serie.
- Los hilos que no tienen ninguna respuesta, conteniendo únicamente el *post* inicial, puesto que se considera que no generan interacción social alguna.
- Los mensajes de longitud cero.

Los hilos arriba mencionados no son relevantes a la hora de estudiar interacciones entre usuarios ni permiten extraer información acerca de los temas, por lo que se ha procedido a eliminarlos. En este paso se han eliminado cerca de 1000 mensajes, 500 hilos y 80 usuarios (ver tabla 8).

Limpieza del campo de fecha

Ya que los datos proceden de un proceso de *scrapping*, la información del campo de fecha se ha capturado con el mismo formato en que se muestra en la página web. Ha sido necesario corregir las fechas de esta manera:

- Los posts que se escribieron el mismo día o el día anterior al que se ejecutó la tarea de *scrapping* muestran en el campo de fecha “Today” o “Yesterday” en lugar de la fecha real. Se han sustituido dichos campos por el valor correspondiente, teniendo en cuenta que la tarea se ejecutó el 24 de Enero

de 2018.

- Los posts escritos durante 2018 incluyen información acerca del día y el mes, pero no muestran el año. Se han modificado de manera que aparezca la fecha completa.

Extracción de citas

Los usuarios que participan en el foro de Planet Virtual Boy tienen la opción de citar un fragmento de texto escrito por otro usuario. El proceso de *scrapping* incluye las citas como parte del texto del post, delimitándolas con palabras clave en inglés y también en alemán ("wrote" y "schrieb"). El nombre de usuario se ha extraido con una expresión regular y se ha guardado en una columna adicional (ver figura 14). No obstante, el usuario tiene la opción de editar el contenido después de citarlo, mantener el texto sin incluir el nombre del autor original o incluso de copiar y pegar texto de otro mensaje sin referencia alguna. En estos casos no es posible identificar la cita.

5.2.3 Extracción de temas

La cadena NLP consta de las siguientes fases:

- **Tokenización**: divide cada documento en tokens con una expresión regular que detecta espacios, saltos de línea, tabulaciones, puntos, comas y comillas dobles. Un token es por tanto el texto delimitado por dos elementos detectados por la expresión regular. Cada documento se convierte en un array con todos los tokens del mismo.
- **Descarte de stopwords**: parte de un diccionario de stopwords para inglés que contiene pronombres, preposiciones, artículos, complementos verbales (como *can*, *shall* y *will*) y verbos habituales (como *be*, *have* y *do*), así como flexiones de los mismos. Este paso recibe cada array de tokens y devuelve otro array sin los tokens que aparecen en el diccionario de stopwords.
- **Vectorización**: este paso difiere de los anteriores en cuanto a que trabaja en dos fases. La primera fase genera un vocabulario con todos los términos que aparecen en el corpus. Este vocabulario se representa como las cabeceras de una matriz inicializada a cero con tantas columnas como términos y tantas filas como documentos. La segunda fase "rellena" la matriz con el número de apariciones de cada término en cada documento. La implementación de Spark no usa matrices sino vectores dispersos (*sparse vectors*), ya que la mayoría de las celdas de la matriz tiene valor cero y es más eficiente representar cada fila

como un conjunto de claves-valor. Las claves identifican el término del vocabulario y los valores, el número de apariciones. Los términos que no están presentes en el documento no aparecen en su vector disperso.

A la ejecución de esta cadena sigue la aplicación del algoritmo LDA. Este algoritmo asignará a cada documento la probabilidad de pertenencia a cada tema. Una característica de LDA, así como de los algoritmos de clustering usados en NLP, es que necesita el número de temas (o clusters) *a priori*. Esto obliga a ajustar el número de temas hasta que se consiga un resultado que se pueda considerar útil. Un número de temas muy reducido va a consolidar muchos documentos bajo el mismo tema, incluso aunque para un humano haya una clara diferencia. Un número muy elevado, por lo contrario, puede acabar generando temas muy específicos para unos pocos documentos.

En un primer intento se ejecuta la cadena NLP considerando un documento por hilo, seguida de LDA con los parámetros por defecto y 10 temas. Se escoge este número con la idea de comprobar si los mensajes de los 10 subforos están perfectamente separados en temas. A continuación se muestra la relación de términos por temas y la cantidad de documentos en cada tema (tabla 9).

```
#0: [u'(jpn', u'(us)', u'manual)', u'vb', u'ebay', u'games', u'(usa', u'vn', u'one', u'(us']
#1: [u'ich', u'0x03ff', u'ist', u'das', u'und', u'adapter', u'seite', u'zu', u'ein', u'mal']
#2: [u'game', u'virtual', u'vb', u'one', u'boy', u'like', u'3d', u'also', u'games', u'think']
#3: [u'vb', u'one', u'game', u'like', u'virtual', u'get', u'games', u'-', u'boy', u'know']
#4: [u'game', u'like', u'one', u'virtual', u'games', u'boy', u'get', u'play', u'vb', u'make']
#5: [u'virtual', u'vb', u'games', u'boy', u'game', u'one', u'like', u'really', u'get',
u'think']
#6: [u'(pal)*', u'vb', u'like', u'game', u'get', u'battery', u'controller', u'know', u'games',
u'%1%2']
#7: [u'tbd', u'title)', u'(working', u'(tentative',
u'\u4fa1\u683c\u672a\u5b9a/\u30ab\u30fc\u30c8\u30ea\u30c3\u30b8virtual',
u'4900', u'5800', u'price)', u'5300', u'yen']
#8: [u'game', u'like', u'one', u'vb', u'make', u'use', u'see', u'got', u'display', u'good']
#9: [u'game', u'vb', u'jpg', u'kb)\xa0', u'get', u'one', u'random_text[number++]\xa0=\xa0',
u'like', u'games', u'3d']
```

topic_id	documentos
3	3773
1	69
5	4
9	2
2	2
6	1
4	1

Table 9: PlanetVB: Temas en el primer intento

En esta prueba se detectan los siguientes problemas:

1. Aparecen signos de puntuación como si fueran palabras aunque no aportan significado.
2. El tokenizador sólo divide textos pero no elimina caracteres como (,) o %, por lo que un mismo término que aparezca rodeado de diferentes caracteres será considerado como un término diferente. En el ejemplo anterior, *title)* y *title* son dos términos diferentes, pero aportan el mismo significado y deberían comportarse como un único término.
3. No se ha aplicado aplicado *stemming* o extracción de raíces; por tanto, *game* y *games* son términos diferentes aunque en un documento aporten el mismo significado.
4. Hay muchas palabras que no aportan significado al tema como *get*, *think* o *make*; hay otras que tienen significado, pero son tan comunes en el contexto que no ayudan a separar temáticas, como *vb*, *virtual* o *boy*.
5. Hay varias ocurrencias de lo que parecen ser cadenas de texto en UTF-8 que no se pueden representar en ASCII. Tras revisarlo en detalle se descubre que son textos en japonés.
6. Aunque LDA ha detectado 10 temas, ha repartido los documentos en sólo 7. Además, uno de ellos acapara la gran mayoría de los documentos.

También se detecta que uno de los temas, el **#1**, contiene varias palabras en alemán y agrupa unos 70 hilos. Como se comenta en la descripción del foro, tiene su origen en Alemania, y hay algunos hilos (especialmente entre los más antiguos) en dicho idioma. Dado que la clasificación realizada por el algoritmo LDA no considera los idiomas, los textos en alemán aparecen como otra categoría simplemente porque, como es de esperar, la palabras en alemán aparecen habitualmente en los mismos documentos con otras palabras también en alemán.

Para solucionar estos problemas se toman las siguientes acciones en sucesivas iteraciones:

1. Se amplía el diccionario de *stopwords* con los signos de puntuación (tanto individuales como en grupos de 2 y 3 caracteres), con las palabras que bien no aportan significado (*get, think, make*) y con las que son demasiado comunes en todos los hilos (*vb, virtual, boy*). La implementación de LDA de Spark usa el mismo diccionario base de NLTK, que para este caso de aplicación deja pasar demasiadas palabras con poco valor. También se añaden los nombres de los usuarios como palabras a evitar, ya que muchos de ellos firman en sus mensajes y debido a ello aparecen con cierta frecuencia.
2. Se implementa un transformador de MLlib para limpiar los caracteres de puntuación alrededor de tokens y descartar términos que no se pueden representar en ASCII. Además transforma términos que contienen **0x** en el token **MEM**, ya que los foros de desarrollo tienden a estar cargados de direcciones de memoria.
3. Se ejecuta LDA con variaciones de los parámetros de configuración. No es posible aplicar un algoritmo de optimización de hiperparámetros porque no se dispone de una métrica de evaluación que indique cuán buenos son los temas; simplemente se procede al análisis y contraste manual por parte de tres evaluadores de los resultados de cada combinación de parámetros. Otra variante que se ha probado ha sido la forma de identificar documentos entre dos opciones: por hilo o por mensaje.

- Optimizador *(method)*: Spark implementa dos optimizadores, *Online Variational Bayes* (Hoffman et al. 2010) y *Expectation-Maximization* (Asuncion et al. 2009). En la experiencia resultante del análisis de este foro, el primero tiende a agrupar la mayoría de los documentos en uno o dos temas principales, mientras que el segundo reparte los documentos de manera más homogénea entre todos los documentos.
- Número de temas (k): se ha probado con cifras desde 5 hasta 15 temas.
- Alfa (*doc_concentration*): para valores altos de este parámetro los documentos contendrán una mezcla de varios temas, mientras que para valores bajos tendrán un tema principal destacado.
- Beta (*topic_concentration*): para valores altos de este parámetro los temas contendrán una mezcla de los mismos términos, mientras que para valores bajos tendrán términos no compartidos con otros temas.

La distribución de temas más razonable, a la luz de la revisión de la intención y significado de los mensajes, se generó para **12** temas, método ***Expectation-Maximization***, doc_concentration de **1.0** (por defecto), topic_concentration de **4.0** y agrupando los mensajes de un hilo en un único documento. La distribución de términos por tema es la siguiente.

```
0: [code, screen, file, version, size, level, sound]
```

```
1: [page, english, translations, instruction, magazine, 1995, preliminary]
2: [cart, nintendo, flashboy, adapter, power, snes, love]
3: [price, MEM, working, released, title, release, space]
4: [cable, display, problem, fix, displays, lines, solder]
5: [ebay, box, collection, buy, sell, case, cart]
6: [nintendo, area, red, space, love, video, cool]
7: [ich, die, und, stage, ist, das, button]
8: [radiation, jugem, sauce, microwave, radiant, peppers, heating]
9: [puck, flipper, combp, flippers, spots, scramble, sweep]
10: [sauce, samsung, peppers, radiation, cheapskate, seeds, banned]
11: [famits, jugem, 1995, issue, number, page, dated]
```

Se asigna una etiqueta a cada tema para poder referenciarlo fácilmente. En el Apéndice I () se muestran extractos de varios hilos de cada tema. Se analiza el contenido de los mismos para validar que los temas que LDA ha extraído tienen sentido. Se ha dado el caso de que varios de los temas identificados por el algoritmo hablan de conceptos similares, así que se procede a agruparlos en un único tema. Se opta, además, por asignar a cada hilo el tema al que LDA asigna mayor probabilidad. Opcionalmente se podrían extraer más temas para cada hilo. Hemos incluido la distribución de temas por foros en la figura 16.

Tema	ID LDA	# posts	# hilos
Social	2	10321	816
Marketplace	5	9429	854
Dev	0	8500	742
Hardware/Repairs	4	4456	525
Misc	6/8/9/10	1818	203
Information/Scans	1/3/11	1625	160
Non-English	7	635	59

Table 10: PlanetVB: número de mensajes e hilos por tema

El esquema de la cadena NLP que se acaba de detallar está representado en la figura 17. Los temas extraídos servirán como otro dato de entrada para la construcción del grafo que sigue a continuación.

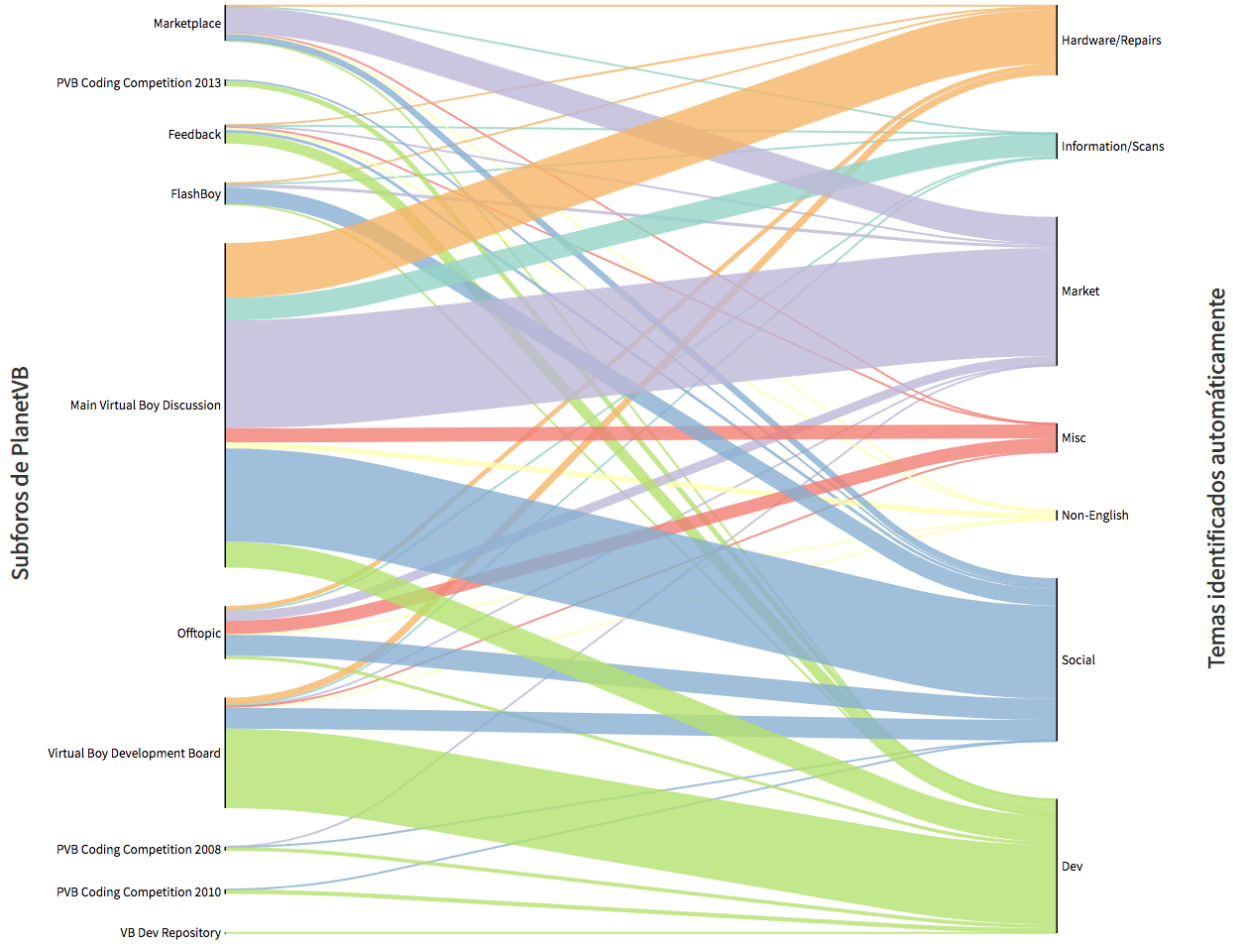


Figure 16: PlanetVB: relación de temas discutidos en cada foro

5.2.4 Construcción del grafo

La construcción del grafo se realiza siguiendo el procedimiento descrito en el apartado , a partir de los datos originales limpios a los que se añade la etiqueta de tema para cada post obtenida en la fase anterior.

Se genera el grafo mediante la librería *GraphFrames* usando dos aproximaciones:

- Grafo valorado: se agrupan los enlaces por la terna (*origen*, *destino*, *relationship*) y se cuenta el número de ocurrencias para asignar el peso de cada enlace como un atributo adicional llamado *weight*.
- Grafo no valorado: no se agrupan los enlaces, el grafo resultante no dispone de atributo adicional para el peso de los enlaces. Al no estar agrupados, la lista de enlaces del grafo no valorado es más numerosa que la del grafo valorado.

Figure 17: Cadena NLP

La representación del grafo generado a partir de cada una de las aproximaciones se puede observar en la figura 18. El objetivo de utilizar estos dos métodos a la hora de crear el grafo es determinar si el hecho de agrupar los enlaces tiene algún efecto sobre los resultados finales.

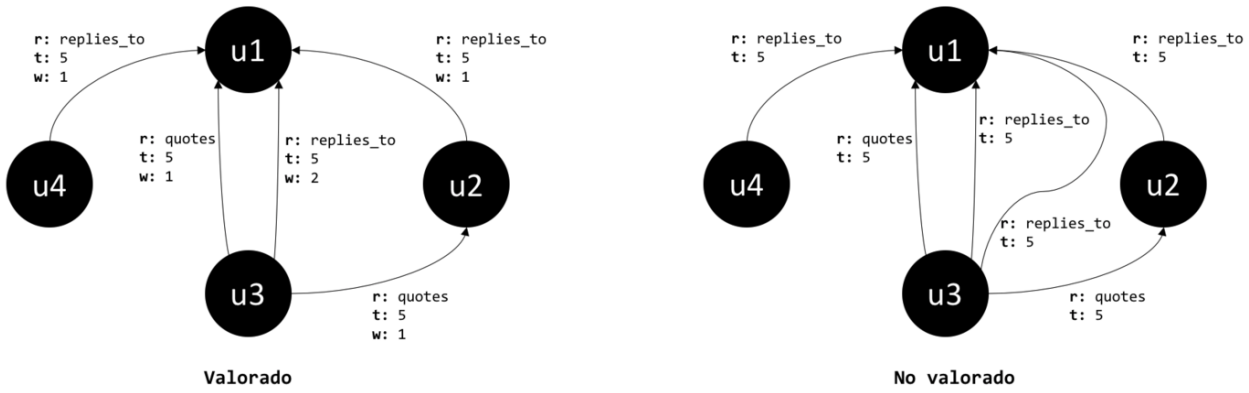


Figure 18: Representación del grafo para cada aproximación (r: relación, t: tema, w: peso).

User	Rank Grafo Valorado	Rank Grafo No Valorado
KR155E	1	1
DogP	2	2
VirtualChris	3	4
thunderstruck	4	6
bigmak	5	5
Benjamin Stevens	6	3
MineStorm	7	7

Table 11: Resultados de PageRank sobre GraphFrames para las dos aproximaciones (en negrita los usuarios con mayor actividad)

5.2.5 Identificación de usuarios relevantes

Los usuarios relevantes se identifican mediante el algoritmo PageRank ejecutado sobre GraphFrames. En la tabla 11 se muestra una comparación de los resultados obtenidos en las dos aproximaciones de construcción del grafo: la primera con los enlaces no valorados, y la segunda con los enlaces valorados y un atributo adicional para el peso.

A la vista de los resultados de la tabla 11, se puede observar que ambas aproximaciones seleccionan el mismo conjunto de usuarios influyentes para los siete primeros puestos. El orden de los primeros usuarios es muy similar también, por lo que se puede concluir que ambos métodos obtienen resultados equivalentes.

Validación

Autor	Número de posts
Benjamin Stevens	1969
KR155E	1883
DogP	1497
RunnerPack	1366
thunderstruck	1035

Table 12: PlanetVB: usuarios más activos

Usuario	**PageRank**	**Authority**	**Hub**
KR155E	1	1	2
DogP	2	2	4
VirtualChris	3	3	32
thunderstruck	4	5	7
bigmak	5	6	8
Benjamin Stevens	6	4	3
MineStorm	7	9	23

Table 13: PlanetVB: Comparación de rankings de usuarios relevantes

Los resultados obtenidos en el apartado se validarán en función de tres aspectos.

Comparación con la actividad de cada usuario

El ranking obtenido está muy relacionado con el número de posts que ha escrito cada usuario (se puede consultar en la tabla 12, se han destacado en negrita los usuarios con mayor actividad en la tabla 11) aunque no coincide exactamente. Esto se debe a que el ranking de PageRank no depende únicamente del número de enlaces, sino que además tiene en cuenta con qué nodos está conectado cada usuario.

Comparación con los resultados del algoritmo HITS

GraphFrames no dispone del algoritmo HITS para obtener un ranking de *hubs* y *authorities*, así que se utiliza NetworkX para realizar los cálculos. Se puede comprobar en la tabla 13 que el ranking de PageRank es muy similar al obtenido con HITS para los nodos considerados autoridades. Se muestran también los resultados de los nodos clasificados como *hubs*, que como es de esperar no tiene por qué coincidir con los usuarios más influyentes.

Comparación con el ranking de foro

Los foros suelen asignar *medallas* o *títulos* a los usuarios en función del número de publicaciones o de alguna participación especial (por ejemplo, si son administradores del foro, moderadores o han ayudado al foro con

Comunidad	Num. miembros	% Miembros
1	654	44.4%
2	636	43.2%
3	100	6.8%
(resto)	82	5.6%

Table 14: PlanetVB: Comunidades LPA

contenido específico). La figura 19 muestra las categorías de los usuarios en PlanetVB y su valor de PageRank (no su posición según el ranking de PageRank). Lo más reseñable es que hay usuarios de menor categoría que son más relevantes que otros que han escrito muchos más posts.

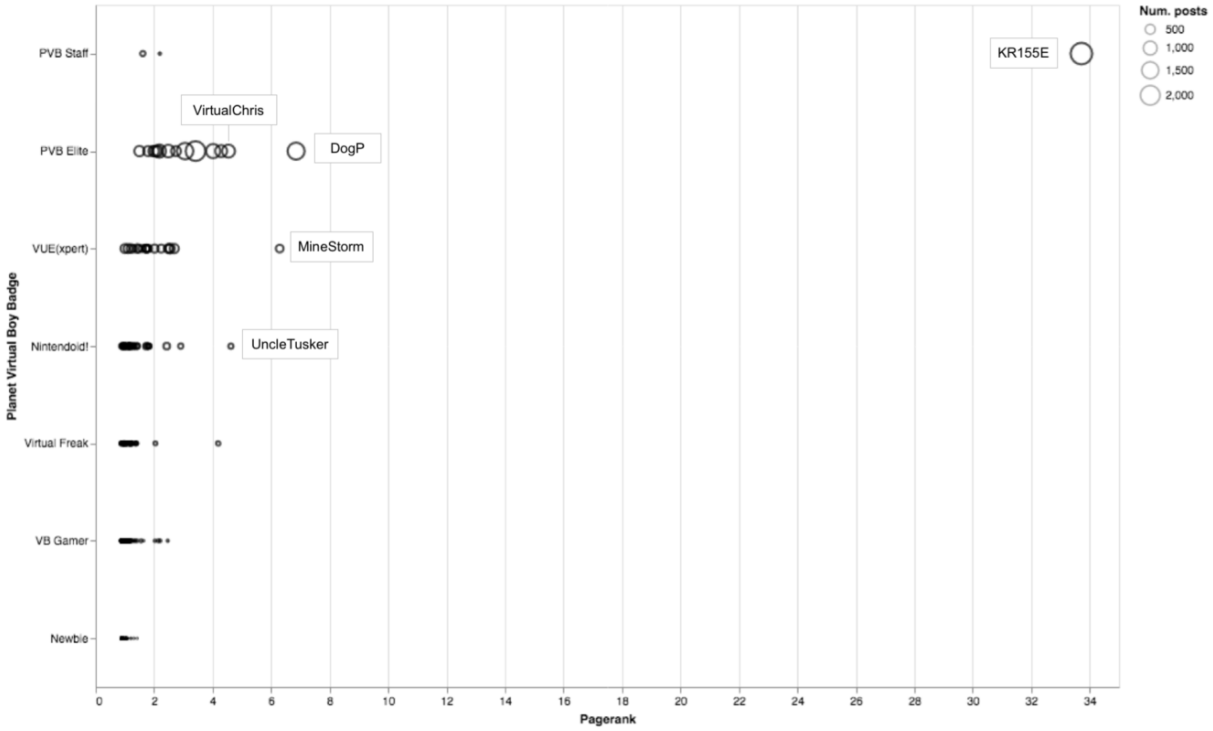


Figure 19: PlanetVB: ranking del foro vs. Pagerank

5.2.6 Identificación de comunidades de usuarios

La identificación de comunidades de usuarios se ha realizado aplicando dos de los algoritmos descritos en , LPA y Louvain. En primer lugar se ha utilizado LPA, que es el algoritmo disponible en *GraphFrames*, obteniendo los resultados de la tabla 14. Se identifican tres comunidades grandes de más de 100 usuarios y numerosas comunidades más pequeñas que corresponden con el 5.6% del número total de usuarios.

A continuación se importan los datos en *Gephi* y se calculan las comunidades con Louvain para comparar

Comunidad	Num. miembros	% Miembros
1	665	45.2%
2	517	35.1%
3	112	7.6%
(resto)	178	12.1%

Table 15: PlanetVB: Comunidades Louvain

los resultados, obteniendo las comunidades de la tabla 15. De nuevo se obtienen tres grupos grandes con un número similar de usuarios, y numerosas comunidades pequeñas que suponen el 12.1% del número total de usuarios.

La representación gráfica de las comunidades más numerosas obtenidas mediante LPA y Louvain se puede observar en la figura 20. Se puede comprobar que ambos algoritmos identifican grupos de usuarios similares, encontrando diferencias especialmente en los nodos que están situados entre ambos grupos. En el gráfico *c)* se ha asignado a cada usuario un color en función del tema más habitual en el que ha participado para poder estudiar si tiene relación con las comunidades encontradas. Se observa que los miembros de las comunidades encontradas no participan necesariamente en el mismo tema de manera mayoritaria. Para identificar comunidades de intereses será necesario un estudio más exhaustivo del grafo, realizado en el apartado .

Sí que cabe destacar el caso concreto de la comunidad de color azul, que en el gráfico *c)* representa al grupo de usuarios que utilizan idiomas distintos del inglés para comunicarse. Dicha comunidad se identifica como grupo independiente en las tres representaciones, y se puede confirmar que los resultados obtenidos en la extracción de temas de son consistentes con los extraidos en este apartado.

5.2.7 Identificación de comunidades de intereses

Con el fin de identificar las comunidades de interés que se encuentran en PlanetVB, se combinan los siguientes pasos:

1. En primer lugar, se construye el grafo entre usuarios y temas según lo descrito en .
2. El grafo se proyecta siguiendo la propuesta desarrollada en .
3. Se analiza el grafo con la proyección superior (temas) y se aplican algoritmos de búsqueda de comunidades en el inferior (usuarios).

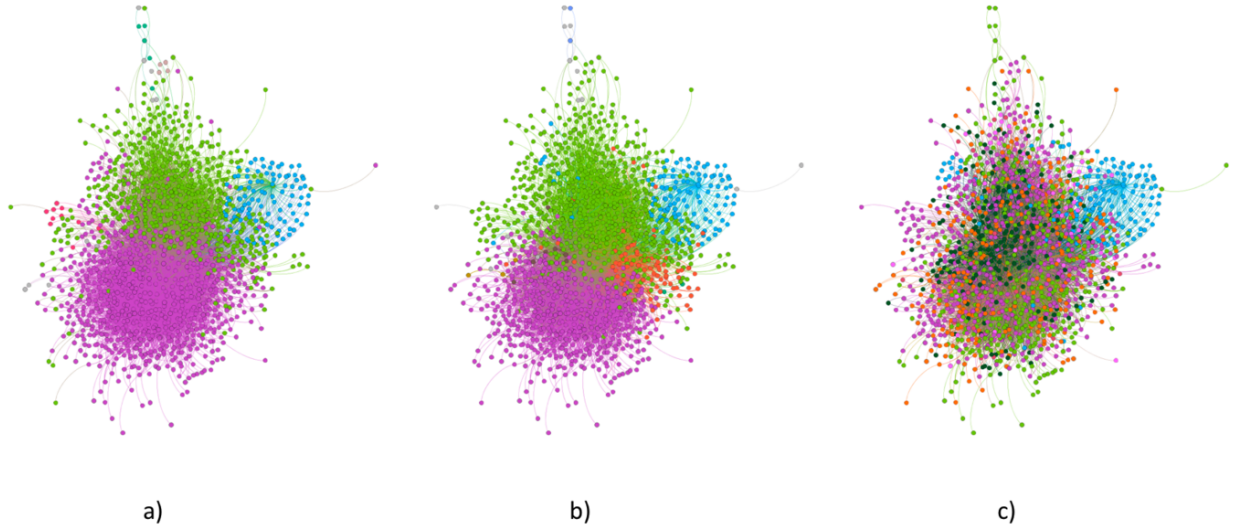


Figure 20: Comunidades de usuarios: a) LPA, b) Louvain, c) Tema mayoritario

4. Se analizan las comunidades obtenidas en el paso anterior mediante el perfil temático de los mensajes agregados de cada uno de los grupos de usuarios.

El grafo resultante de la proyección sobre los temas es el mostrado en la figura 21:

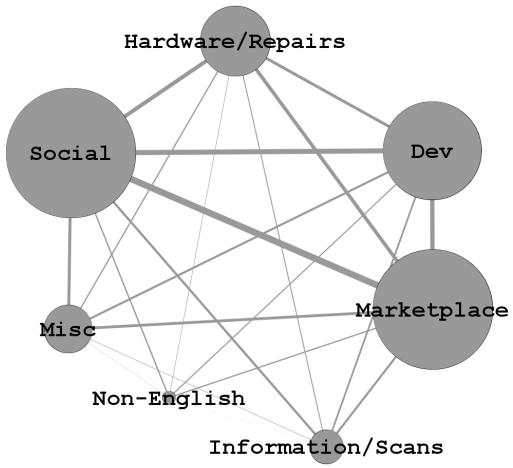


Figure 21: Proyección del grafo de usuarios-temas sobre los temas para el caso PlanetVB. La relación más fuerte se halla entre Social y Marketplace y el triángulo más fuerte incorpora el Desarrollo (Dev).

Las relaciones más relevantes son:

- El vínculo entre “Social” y “Marketplace” es el más fuerte. Es decir, es la pareja de temas más común

entre todos los usuarios.

- La triada más relevante se completa con "Dev". Esto significa que, en la comunidad de PlanetVB, es bastante probable que un usuario que utiliza los foros para dos de estos temas (por ejemplo, como desarrollador y socializador) participe también en el tercero (en este caso, la compra-venta).
- El cuarto tema en todas estas relaciones es la discusión sobre elementos de *hardware* o reparaciones.

Por lo que a la proyección sobre los usuarios respecta, representarlo es poco útil al estar formado por 1.162 nodos pero 390.165 aristas. La ejecución del algoritmo de detección de comunidades (*Louvain*) en *NetworkX* resulta en la identificación de cinco comunidades distintas, distribuidas sobre el total de usuarios según lo mostrado en la tabla 16. El posterior análisis se ha basado en clasificar cada usuario según su comunidad y realizar un conteo agrupado del porcentaje que representa cada tema para cada comunidad. Los resultados y su interpretación se muestran, también, en la misma tabla 16.

5.2.8 Resumen del caso PlanetVB

La estructura del foro de Planet Virtual Boy es relativamente ordenada y constituye una buena representación de un foro básico, centrado en un tema concreto y con una cantidad de información adecuada para una primera iteración de la metodología. Durante el análisis se han detectado algunos puntos destacados, como subforos con una actividad muy superior al resto o usuarios especialmente participativos.

Los temas identificados automáticamente están alineados con el subforo al que pertenecen, especialmente en foros específicos como *Marketplace* o *Dev*. Esta consistencia permite validar que la detección de tema ha funcionado de manera razonable en este caso de estudio concreto y su uso en la detección de comunidades está justificado.

El análisis de usuarios en base a su actividad en las conversaciones ha identificado correctamente los usuarios más relevantes (el administrador del foro es el usuario más activo y el de mayor ranking según PageRank). Adicionalmente, permite concluir que el número de mensajes no implica directamente una mayor relevancia.

Con respecto a las comunidades de usuarios identificadas a partir de las respuestas y citas presentes en los hilos, los grupos encontrados no atienden a una organización por temas y sería necesario realizar un estudio más exhaustivo. El análisis en base a respuestas y citas ha permitido identificar diferentes comunidades pero

sólo de manera limitada, siendo más completas las conclusiones que se extraen mediante el grafo bipartito de temas y usuarios. En el caso AtariAge no se continuará la vía del análisis de comunidades.

A nivel de la relación entre temas y usuarios, las principales conclusiones son:

- Si un usuario participa en más de un tema, lo más probable es que lo haga en hilos de carácter social y en hilos de compra-venta. Tras ellos, se encuentran el desarrollo y los mensajes sobre *hardware* y reparaciones.
- Aparecen cinco comunidades que, si se obvia la comunidad de habla no inglesa, se reparte en aproximadamente un tercio de usuarios interesados principalmente en la compra-venta (probablemente coleccionistas o vendedores profesionales), otro tercio de usuarios que se dedican al desarrollo de software o que preguntan por los últimos desarrollos y un último tercio que se divide en dos partes: una de aficionados al sistema, que principalmente usa el foro para socializar, y otra que accede para solicitar reparaciones o buscar soluciones a sus problemas con la máquina.

5.3. Caso de estudio 2: AtariAge

La elección de este foro como segundo caso de estudio servirá para extrapolar el trabajo realizado sobre PlanetVB. Es especialmente interesante comprobar qué modificaciones debe sufrir al flujo de trabajo al saltar de un foro de pequeñas dimensiones a otro con un tamaño dos órdenes de magnitud superior.

5.3.1 Estructura del foro

En Agosto de 2018, momento en el que se capturaron los datos, AtariAge ofrecía 32 foros, algunos de ellos divididos en subforos, con más de 3,3 millones de mensajes (tabla 17), escritos por cerca de 23.400 usuarios. La clasificación en foros y subforos es casi obligada dado el volumen de mensajes. La página de inicio muestra un nivel superior al de foros, que se podrían llamar categorías, donde se agrupan los diferentes foros: sistemas Atari, otros sistemas, compraventa, desarrollo, comunidad y noticias del sitio. Algunos foros aparecen en

varias categorías; por ejemplo, los foros de programación aparecen tanto en la categoría de desarrollo como en sistemas Atari u otros sistemas, según el sistema al que estén dedicados.

Aunque AtariAge supera en números absoluto a PlanetVB, los hilos y los mensajes tienen una longitud media similar (en número de mensajes y de caracteres, respectivamente; ver tabla 18 y figura 22). Los usuarios son más activos, con una media de mensajes superior. Tres usuarios han escrito más de 20.000 mensajes, y el 3% de los usuarios ha escrito cerca del 67% de los mensajes.

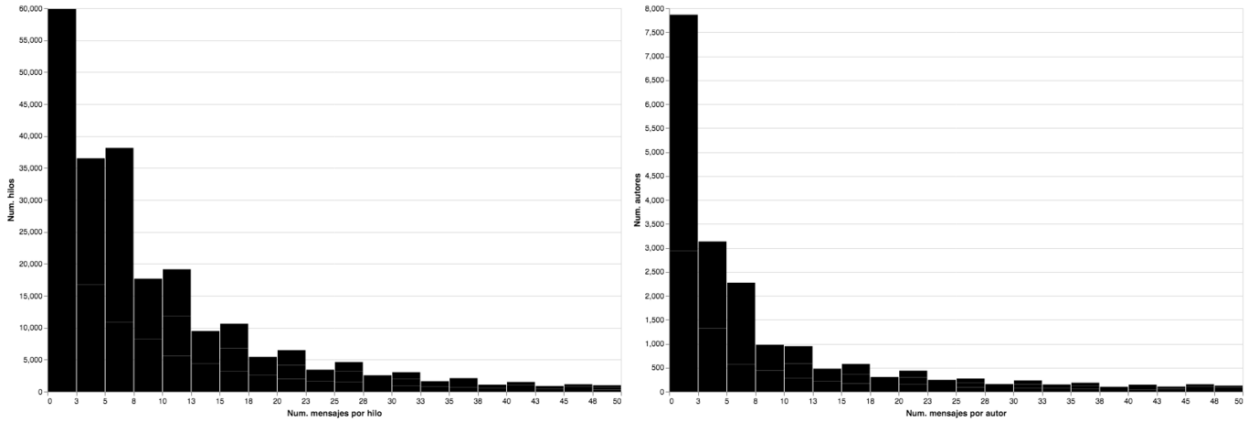


Figure 22: AtariAge: mensajes por hilo y por usuario (eje x truncado en 50)

5.3.2 Limpieza

Si bien los datos de ambos foros fueron descargados mediante *scrapping*, la gran cantidad de metadatos embebidos en el código HTML de AtariAge ha permitido extraer muchos campos en un formato usable sin apenas preprocesamiento. Por ejemplo, no ha sido necesario interpretar las fechas ni extraer las citas a partir del texto de los mensajes, ya que en ambos casos estos datos estaban presentes en etiquetas HTML.

No obstante, sí se han eliminado hilos estáticos con normas de uso o con noticias internas (cambios en el estilo de la web, nuevas secciones, etc.) y varios listados que contienen enlaces a blogs externos a AtariAge, números de serie de consolas y nombres de usuario en otras redes (que los usuarios comparten para conectar en PlayStationNetwork o EA Online). En total se han borrado cerca de 170.000 mensajes, 44.000 hilos y 800 usuarios, un porcentaje ligeramente superior al de PlanetVB.

5.3.3 Extracción de temas

El trabajo sobre PlanetVB ha permitido construir una cadena NLP para modelado de temas que ha sido fácilmente adaptable a este segundo caso de uso. Ha sido posible reusar prácticamente todo el código salvo ciertos detalles. El principal reto de la extracción de temas fue trasladar el código de PlanetVB, escrito para PySpark pero probado en un único nodo (un clúster local de Databricks con 6 GB de RAM), a un clúster real con varios nodos, ya que el trabajo enviado a Spark provocaba errores de límite de memoria. Aunque era posible cargar el foro en memoria y aplicar consultas SQL, la fase de entrenamiento del algoritmo LDA provocaba los citados errores de memoria. LDA tiene una complejidad (en este caso) de $O(b \cdot (N + k)^3)$, donde b es el número de términos asignados a un tema, N el número de términos por documento y k el número de términos con el que se inicializa el algoritmo (Sontag and Roy 2011), por lo que la duración del proceso era varios órdenes de magnitud superior en AtariAge respecto a PlanetVB. A esto hay que sumarle que la matriz términos-documentos, aun con una implementación dispersa, es también más grande para AtariAge. Para solucionar esta limitación se optó por migrar el código a un cluster EMR (Elastic MapReduce) de Spark en AWS. La migración de código como tal necesitó modificar la ubicación de almacenamiento de los datos (en local vs. en AWS S3) y tratamiento de cadenas Unicode debido a la diferencia de versiones Python entre los clusters locales y los provisionados en AWS.

En cuanto a la lógica de la cadena NLP como tal, fue necesario modificar el diccionario de *stopwords* en dos aspectos: hubo que añadir términos del nuevo vocabulario y eliminar los nombres de los autores, ya que muchos de ellos usaban nombres de equipos y consolas como identificadores de usuario, y estos términos eran útiles para identificar temas. El resto de pasos de la cadena NLP fueron idénticos, ya que la naturaleza de ambos foros es muy similar. No obstante, se optó por etiquetar manualmente los hilos de los subforos *International* y *High Score Clubs*. Los mensajes del primero están en muchos lenguajes diferentes y podían *ensuciar* los temas identificados por LDA. El segundo es realmente un conjunto de subforos en los que los usuarios añaden semanalmente su puntuación más alta en diversos juegos. Estos mensajes pueden ser interesantes para encontrar relaciones entre usuarios, pero el contenido suele ser un número y una captura de pantalla, así que se decidió no incluirlos en el entrenamiento del *topic modeling*.

Al igual que con PlanetVB, se ejecutó la cadena completa varias veces con diferentes combinaciones de

hiperparámetros. El mayor volumen de documentos parecía indicar un mayor número de temas, y finalmente se decidió usar un valor de *k=45*. Los parámetros *method*, *doc_concentration* y *topic_concentration* se mantuvieron en **Expectation-Maximization**, **1.0** y **4.0** respectivamente.

Los temas con mayor presencia fueron el #3, relativo a compraventa, #25, con discusiones sobre juegos, y #4, también sobre juegos pero con un marcado interés en los publicados para la consola Atari 2600. La relación de temas y términos completo está en en apartado .

```
#3: ['ebay', 'auction', 'item', 'seller', 'price', 'shipping', 'bid']
#25: ['points', 'arcade', 'super', 'played', 'mario', 'star', 'wii']
#4: ['2600', 'minutes', 'team', 'ebay', 'week', 'label', 'min']
```

Cada uno de estos temas se etiquetó con un título con el mismo procedimiento que se utilizó para PlanetVB.

Hay dos conexiones especialmente relevantes en la figura 23. Por un lado, la mayoría de los mensajes del foro *Marketplace* han sido etiquetados con el tema *Market*. Por otro, también la mayoría de los mensajes de *Programming* han caído bajo el paraguas del *Dev*. Hay una segregación de foros en temas más marcada que para PlanetVB, pero hay que resaltar que foros como *Atari 2600, Atari Jaguar* o *Atari 8-bit Computers* tratan de englobar todas las conversaciones en torno a un único producto. Estas conversaciones versan sobre discusiones de juegos, desarrollo, reparaciones o compraventa: es decir, hay una división en temáticas con un lenguaje homogéneo, independiente del producto que se trate.

5.3.4 Construcción del grafo

Al igual que en la extracción de temas descrita en el apartado , la construcción del grafo para el caso de estudio de AtariAge se hizo utilizando un cluster EMR de Spark en AWS. Todos los conjuntos de datos que se utilizan durante la extracción de temas y construcción del grafo (datos de partida, resultados intermedios y datos procesados finales) se guardan en AWS S3, con el objetivo de persistir los datos y poder retomar la ejecución en cualquier punto del proceso. El procesamiento de un volumen de datos tan amplio como el que se maneja en este caso de estudio es muy costoso computacionalmente, y es útil disponer de los resultados intermedios para evitar ejecutar todo el *pipeline* si es necesario realizar cualquier cambio en el código.

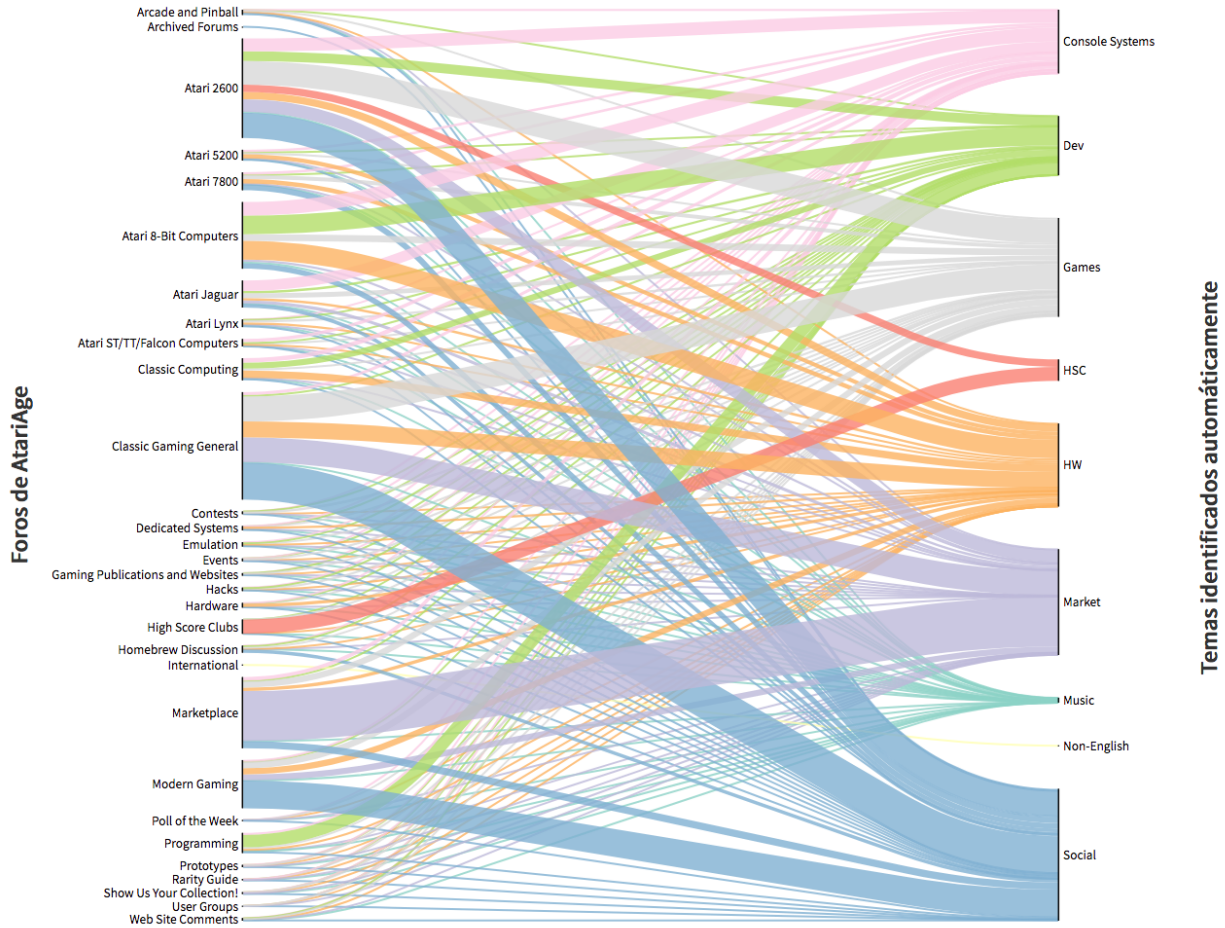


Figure 23: AtariAge: distribución de temas por foros.

El procedimiento que se ha seguido es similar al seguido en el caso de estudio de Planet Virtual Boy, siendo solamente necesario modificar la manera en que se procesan las referencias de unos usuarios a otros dentro de un post. La diferencia radica en que en Planet Virtual Boy sólo se permite incluir una referencia dentro del cuerpo del post, mientras que en AtariAge se permite el uso de múltiples referencias. El proceso de *scrapping* separa las referencias a una columna que contiene un array de usuarios; un array vacío significa que no hay citas en el cuerpo del mensaje.

En la tabla 20 se puede consultar el aspecto que tiene la columna *quoted_user* antes de ser procesada, y los resultados obtenidos tras el procesamiento en la tabla 21. Los arrays vacíos se ignoran, y se separan en líneas diferentes las referencias de un autor a cada usuario, replicando el resto de la información de cada registro.

Por otra parte, en el apartado se observó que el hecho de trabajar con un grafo valorado o no valorado no afecta de manera relevante a la hora de extraer conclusiones, por lo que para este caso de estudio se trabajará exclusivamente con el grafo generado a partir de los enlaces no valorados.

5.3.5 Identificación de usuarios relevantes

Se aplican PageRank y HITS sobre el grafo generado en GraphFrames y se comparan sus resultados. En la tabla 22 se puede observar que el usuario más influyente coincide tanto para PageRank como para el ranking de authorities de HITS, y dicho usuario también se clasifica como un buen hub. En este caso de estudio los rankings de authorities y PageRank son similares, al igual que ocurría en el caso de Planet Virtual Boy. Por otra parte, se vuelve a confirmar que los usuarios con mayor actividad (ver tabla 23) tienen mayor probabilidad de convertirse en un usuario influyente de acuerdo a los resultados de los algoritmos.

Validación

AtariAge también asigna títulos a los usuarios en función del número de mensajes que publican, existiendo un total de nueve categorías estándar. Adicionalmente se permite que el usuario escoja su propia categoría personalizada mediante un campo de texto libre, lo que resulta en numerosas categorías compuestas por un único usuario. Se han englobado dichas categorías personalizadas en un grupo llamado *Custom* para el estudio realizado en la figura 24, donde se puede comprobar que hay usuarios de estamentos inferiores con un ranking más elevado que usuarios que han publicado más mensajes, al igual que ocurría con PlanetVB.

5.3.6 Identificación de comunidades de intereses

Siguiendo la misma idea que en el caso anterior se procede a identificar las comunidades de interés que se encuentran en AtariAge. El grafo que relaciona los usuarios con los temas en los que interactúan se elabora de forma análoga al caso anterior; debido al tamaño de AtariAge, no obstante, algunos pasos se ven modificados en la proyección sobre los nodos de usuarios.

Proyección sobre los temas

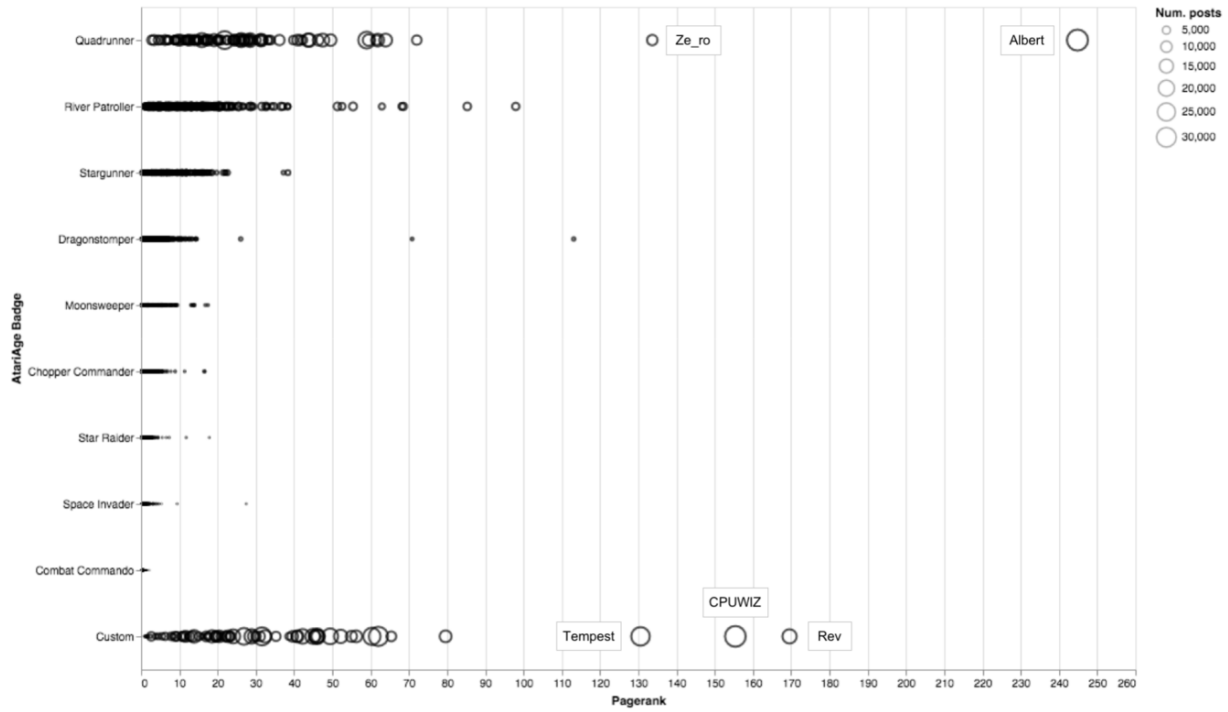


Figure 24: AtariAge: ranking del foro vs. Pagerank

Siguiendo el algoritmo descrito en , se realiza la proyección sobre los temas sin ninguna modificación adicional. El resultado de la operación se muestra en la figura 25. En este caso, las relaciones más relevantes que aparecen son:

- El vínculo entre "Social" y "Market" vuelve a destacar como el más fuerte, aunque los cuatro enlaces más destacados tienen siempre "Social" como uno de sus nodos (Social - Market, Social - Games, Social - Dev y Social - Hardware).
- La tríada principal añade la relación entre "Market" y "Games", lo que parece indicar que la temática principal de los foros de AtariAge es aquella que combina las actividades sociales (que se podrían dar por supuestas) con la compra y venta de juegos, marcando un cierto perfil de coleccionismo de artículos de *retrogaming*.

Proyección sobre los usuarios

Debido al tamaño de AtariAge, algunos pasos adicionales son necesarios para poder elaborar la proyección sobre los nodos de usuarios. En particular:

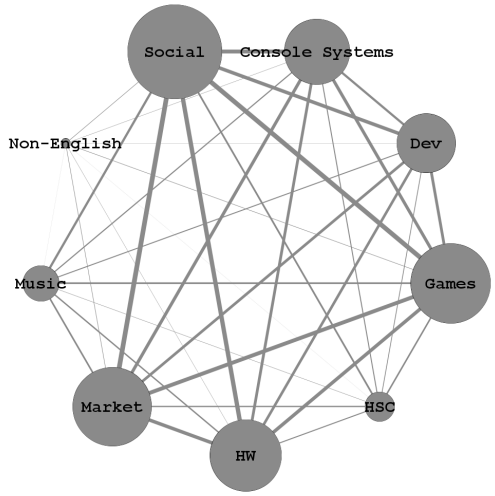


Figure 25: Proyección del grafo usuarios-temas sobre los temas para el caso de AtariAge. La relación más fuerte está entre Social y Market; el triángulo de más peso incorpora también el tema "Games".

- El código propuesto en colapsa en una máquina con 16GB de RAM; tras unas primeras optimizaciones se consigue procesar aproximadamente un 15% de los usuarios, cantidad insuficiente. Con la misma estrategia empleada para PlanetVB harían falta unos 70GB de RAM para procesarlo.
- Así, se opta por trabajar con un fichero auxiliar al que se vuelcan las aristas obtenidas de la proyección. De las casi 300 millones de aristas que podría contener el grafo, se obtienen 100 millones, lo que indica un conjunto de relaciones relativamente densas (muchos usuarios comparten temas).
- Con una muestra de 100.000 aristas del fichero final, se calcula el percentil 90 para el peso, con el fin de seleccionar, de forma aproximada, el 10% de aristas de más relevancia (es decir, las conexiones más fuertes entre usuarios).
- Con el valor umbral del peso determinado en el paso anterior se procesa la lista de aristas y se obtiene una nueva, filtrada, con poco menos de 10 millones de aristas entre sus nodos. A nivel de nodos, existen exactamente los mismos usuarios que en el grafo sin filtrar (no se ha perdido ninguno con el filtro).
- Finalmente se ejecuta el cálculo de comunidades mediante *Louvain* en NetworkX (paquete *python-louvain*).

El posterior análisis se ha basado en clasificar cada usuario según su comunidad y realizar un conteo agrupado del porcentaje que representa cada tema para cada comunidad, de forma análoga a lo visto en el caso 1. Los resultados y su interpretación se muestran, conjuntamente, en la tabla 24.

5.3.7 Resumen del caso AtariAge

El caso de AtariAge ha servido para añadir dos niveles adicionales de complejidad al caso PlanetVB. En primer lugar, su escala, que es dos órdenes de magnitud superior respecto al número de mensajes. En segundo, la temática (al menos a priori) que se centra en un enfoque mucho más amplio, considerando aficionados a prácticamente todos los sistemas clásicos. Los resultados, sin embargo, han mostrado una estructura muy similar a la del primer caso.

Por un lado, los temas detectados por LDA han sido, en líneas generales, similares a los temas de PlanetVB. Incluso con la cantidad de productos con subforo en AtariAge, las discusiones de los usuarios estaban centradas en compraventa, desarrollo y reparaciones de hardware, tal y como ocurría en PlanetVB. La principal diferencia ha sido, como se ha comentado, el aspecto técnico: la mayor cantidad de datos ha requerido mayor capacidad de computación. La decisión de optar por un sistema escalable horizontalmente permitió migrar sin apenas modificaciones.

En cuanto a la detección de usuarios relevantes, se confirma que los usuarios con mayor actividad en el foro tienen mayor probabilidad de ser seleccionados como autoridades, aunque se encuentran numerosos ejemplos de usuarios con menor número de mensajes identificados como relevantes por el algoritmo.

A nivel de la relación entre temas y usuarios, las principales conclusiones son:

- El perfil más común es el del coleccionista, interesado en socializar y comprar (o vender) juegos, especialmente. Si un usuario participa en más de un tema, lo más probable es que lo haga en hilos de carácter social y en hilos de compra-venta, desarrollo o máquinas y juegos.
- Aparecen cinco comunidades bastante bien definidas; las dos mayores contienen al grupo de desarrolladores ("Dev", pero también "Hardware" y "Console Systems") y a los coleccionistas (básicamente "Market" y "Games"), y están formadas por algo más de la mitad del total de usuarios. La otra mitad se divide en tres comunidades algo más pequeñas pero equilibradas: la primera es altamente social, mientras que la segunda contiene usuarios que hablan de juegos y sus puntuaciones máximas, com-

partiendo sus hitos, con lo que se diferencia de la comunidad de coleccionismo porque aquí sí hablan de juegos para jugarlos y no para comprarlos y venderlos. La última comunidad muestra un equilibrio entre varios temas, lo que parece indicar que o bien son usuarios más polivalentes (con un interés general en los foros de AtariAge) o bien buscan ayuda, soporte u opinión en algunos de los temas complementarios, al parecerse a la primera comunidad pero con un mayor elemento social y menor contexto de desarrollo.

6. Discusión y propuesta de metodología

6.1. Discusión

Los casos de estudio previos han servido de ejemplo de aplicación de un flujo de trabajo para conseguir los objetivos propuestos en el apartado . Algunos de los pasos son fácilmente generalizables mientras que otros requieren de revisión en función de la morfología y el dominio del foro.

Los procesos de *scrapping* y limpieza han sido específicos de este proyecto. Cualquier entorno real dispondrá de los datos en una base de datos preparada a tal efecto, bien con un esquema personalizado para una aplicación *ad hoc* o generado a partir de un administrador de contenido específico para foros como phpBB o Discourse, que soportan bases de datos relacionales. La fase de recogida de datos sería por tanto inexistente o, en todo caso, fácil de implementar aplicando una fase de ETL que vuelque los datos estrictamente necesarios a otra instancia de base de datos. La decisión de alimentar el procesado a partir de una instancia diferente a la de producción puede atender a requisitos de seguridad (por ejemplo, diferentes equipos que acceden a datos para propósitos diferentes) o de latencia (para acercar los datos al cluster de procesado). La fase de limpieza puede incorporarse directamente en este proceso ETL - todos los pasos indicados en los apartados y han sido implementados en Spark SQL con la intención de aprovechar la optimización de consultas y el lenguaje declarativo habitual en entornos relacionales (Jurney 2017) y así acercar lo más posible el caso de estudio a un caso real.

La cadena NLP es generalizable a otros foros con un formato de hilos y mensajes. Otras estructuras, co-

mo mensajes con comentarios o hilos con una diferencia clara entre preguntas y respuestas requerían una personalización adicional. En cuanto a la temática, un ajuste correcto de los parámetros permitiría aplicar la cadena a foros con temas afines (como en los casos de estudio) o muy dispares. Sería necesario, por tanto, optimizar los parámetros de las diferentes fases, como la selección de *stopwords*, la generación de tokens específicos (como en el caso de **MEM** para direcciones de memoria), el número de temas y la configuración de LDA a partir de *doc_concentration*, *topic_concentration* y el método. Los dos primeros parámetros requieren un conocimiento en profundidad del dominio de trabajo: incluso en foros con una temática aparentemente similar ha sido necesario identificar *stopwords* específicas.

La principal debilidad de esta cadena se encuentra en foros multilenguaje. Para PlanetVB ha conseguido identificar mensajes en otro lenguaje como un tema concreto, y para AtariAge se han obviado los mensajes que no estuvieran en inglés. Una posible solución sería introducir un paso de detección de idioma en la cadena NLP; otra opción sería aplicar la cadena dos veces: una primera ejecución sin stopwords y un valor de *topic_concentration* muy bajo, con el objetivo de identificar lenguajes como si fueran temas, seguida de una segunda ejecución similar a la propuesta ***para cada idioma***.

La estrategia de construcción del grafo es, al igual que la cadena NLP, generalizable a cualquier foro con una estructura de hilos y mensajes, sea cual sea la temática. La relevancia de usuarios en un foro de preguntas y respuestas no se podría evaluar correctamente con esta estrategia. La elección de un algoritmo concreto no es inamovible, ya que se han validado los resultados de PageRank con HITS. No obstante, la elección de Spark en las primeras fases del estudio minó en cierto modo la versatilidad de este análisis, y fue necesario usar NetworkX. La facilidad de escalado horizontal de Spark fue una ventaja en la fase de modelado de temas, pero no así para el cálculo de métricas de redes: NetworkX trabaja sobre un único nodo y aún así fue capaz de trabajar con las redes de ambos foros sin problemas aparentes.

El uso de citas como uno de los enlaces entre usuarios depende de la posibilidad de identificarlas en el contenido del mensaje. Incluso aunque el foro soporte las citas como una función nativa, los usuarios pueden mencionarse unos a otros sin usar dicha función, o incluso borrar el contenido o el autor de una cita. Por tanto, la presencia de citas puede estar falseada.

El análisis de las comunidades de usuarios identificadas a partir de las respuestas y citas presentes en los hilos no ofrece unos resultados tan completos como los que se obtienen mediante el grafo bipartito entre temas y usuarios, por lo que resulta más interesante explotar esta segunda opción.

Por último, la proyección del grafo bipartido entre temas y usuarios mediante un algoritmo diseñado para la ocasión y basado en las redes de colaboración de Newman resulta útil para identificar las comunidades de usuarios, que se agrupan según su perfil temático y que muestran ciertas similitudes en ambos casos, de lo que se intuye que es posible, quizás, que exista una estructura común para las comunidades de interés en el contexto del *retrogaming*. En particular, se ha visto cómo los temas principales tratados en ambos casos son prácticamente iguales: por un lado la socialización, prácticamente dada por supuesta en un entorno de red social, y por otro, la compra y venta, principal actividad de coleccionistas y aficionados a los sistemas antiguos y que parece ser también el principal motivo para incorporarse a los foros de estudio. El tercer elemento difiere ligeramente entre ambas comunidades; PlanetVB, una comunidad más pequeña y cerrada a un único sistema, parece estar más relacionada con los proyectos de desarrollo - con lo que toman un componente muy comunitario - mientras que en una comunidad más grande como AtariAge tiene lugar un mayor número de discusiones y debates sobre juegos de diversos sistemas, comparaciones de versiones y máximas puntuaciones.

A nivel de comunidades de usuarios según su interés aparece una estructura parecida, dividida en 5 comunidades que mantienen un balance también similar: las dos comunidades más grandes, que aglutinan a más de la mitad de usuarios, corresponden a los grupos de desarrollo y a los grupos de compra-venta (y coleccionistas). Estas actividades también se corresponden a los temas más habituales, con lo que demuestra coherencia. Una tercer conjunto, también común en ambos casos, es aquel que tiene más comportamiento social. La diferencia es que en PlanetVB se forma una comunidad única con este perfil de usuarios, mientras que en AtariAge se parte en dos: una puramente social y otra en la que la interacción se basa en el compartir datos y puntuaciones sobre juegos. El cuarto grupo es el de los usuarios que buscan ayuda sobre ciertos temas o reparaciones. En AtariAge este grupo es más heterogéneo, mientras en PlanetVB, dedicado únicamente a la Virtual Boy, consola plagada de problemas técnicos, se centra más en ofrecer y demandar reparaciones de *hardware*. En PlanetVB se ha detectado otra pequeña comunidad de usuarios de habla no inglesa: su tamaño tan pequeño en términos relativos es, probablemente, el causante de que esta comunidad no se haya

replicado en el caso de AtariAge.

Se encuentran, así, varias limitaciones. En primer lugar, el elevado número de usuarios que comparten más de un tema hace que la proyección del grafo sobre los nodos de usuarios tenga gran tamaño, obligando a filtrar o realizar un muestreo en el caso de AtariAge. En segundo, la complejidad computacional de los algoritmos de detección de comunidades, que escalan con el número de enlaces, fuerzan a limitarse a la aplicación de métodos como *Louvain*, un algoritmo voraz con el que se obtienen resultados prometedores pero que no garantizan que estén cerca del óptimo. Además, pueden presentar problemas en la detección de comunidades pequeñas; en el caso de AtariAge, por ejemplo, se han detectado grandes comunidades dedicadas al coleccionismo o al desarrollo, pero posibles comunidades temáticas como la de "Music" (que es un tema muy específico) o la de "Non-English" (que sí había sido detectada en PlanetVB) han quedado diluidas dentro de categorías mucho más grandes. Esta falta de granularidad (o la incapacidad del algoritmo para trabajar con comunidades de tamaños muy dispersos) se considera, pues, otra limitación del trabajo realizado.

6.2. Propuesta

La metodología propuesta a continuación pretende generalizar el trabajado sobre los dos casos de estudio. Estaría compuesta de las siguientes fases:

- **Análisis de la fuente**. En esta fase es necesario analizar la forma de los datos (estructura y esquema), el dominio y su ubicación. **Hacerlo nos proporciona información sobre el tipo de datos y el formato que tienen, así como la tecnología que habrá que usar para extraerla. En líneas generales, los mensajes de los foros con estructura tradicional se adaptan bien al formato JSON y, por lo tanto, también se podrían almacenar en MongoDB si hiciese falta por volumen u otros motivos. En ninguno de los dos casos analizados ha sido necesario almacenarlo más allá del fichero (aunque también es cierto que el segundo caso, AtariAge, genera más de 2GB de información).**
- **Elección del motor de procesamiento. En función del volumen de datos a analizar, se**

podrá elegir como motor de procesamiento un servidor independiente o bien un cluster de servidores con recursos más potentes para agilizar las fases más pesadas del proceso, especialmente para la identificación de temas. En el caso de foros pequeños como PlanetVB será suficiente con un único servidor. Sin embargo, si el volumen de datos es mayor (incluso unos pocos gigabytes marcaron el límite en un nodo aislado), es recomendable escoger una plataforma que sea paralelizable y extensible horizontalmente con una inversión de tiempo mínima, por ejemplo Spark.

- **Carga de información. De cara al acceso a los datos por parte del motor de procesamiento será necesario tener en cuenta la latencia de conexión entre la base de datos y los nodos de procesamiento. La ventaja de usar un entorno como Spark es que no escribe a disco en cada tarea, por lo que sólo es necesario tener en cuenta la latencia en la lectura inicial y en cualquier guardado intermedio. En el caso de trabajar con ficheros planos, el proces de carga en el sistema de ficheros de los nodos de procesamiento puede implicar un tiempo importante pero reducirá la latencia drásticamente.**
- **Limpieza**. En función del esquema de los datos, esta fase puede requerir transformaciones, para disponer de campos concretos en las siguientes fases, o filtrados. En caso de descartar registros hay que considerar aquellos que no aporten información relevante a la identificación de temas (hilos sin contenido textual o en idiomas que no se van a analizar) o al análisis de redes (hilos con reglas de conducta, noticias internas, etc.). En caso de disponer de los datos en su base de datos original es recomendable aplicar esta fase mediante comandos nativos del motor.
- **Identificación de temas**. En esta fase son fundamentales un conocimiento de los temas que se tratan en el foro y un trabajo de optimización de parámetros. Sin estos, la estrategia de identificación automática genera demasiado ruido y poco valor. El paso manual de revisión y etiquetado de temas permite validar el proceso y sirve de entrada para la identificación de comunidades. **Se propone el uso de LDA ya que es relativamente sencillo de optimizar (sólo fueron necesarios 4 parámetros en los casos de estudio) y ha demostrado ofrecer resultados útiles y fáciles de entender.**
- **Construcción de grafos.** Aquí es necesario elegir la estrategia de modelado del foro, la representación de los enlaces y las relaciones que se consideran (responde a, cita a, etc...). El uso de varios grafos de diferente construcción añade más posibilidades a los análisis posteriores. **Una de las formas más razonables de hacer la construcción es considerar que el iniciador del hilo es el usuario que escribe el primer mensaje y que el resto de usuarios contestan, de alguna forma, a**

este primer usuario, excepto en el caso de las citas directas, que marcan una relación clara. A nivel social, los usuarios que escriben en un hilo nuevo pero no reciben respuesta no tienen interacción, así que se eliminan.

- **Elección de métricas de red.** A partir del grafo obtenido en el paso anterior se trata de extraer las métricas necesarias tanto para describir la red como para proceder posteriormente a identificar los usuarios influyentes y las comunidades que se forman.
- **Obtención de usuarios influyentes.** La obtención de un ranking de influencia consiste en aplicar algoritmos al grafo construido a partir de las relaciones entre usuarios. Se sugiere utilizar PageRank (un algoritmo extendido y conocido) para detectar la influencia de los usuarios, ya que se corresponde bien con las propias categorías que tienen los foros pero, a la vez, es capaz de detectar usuarios más prestigiosos de lo normal en categorías inferiores (o con menos mensajes). En la misma línea, PageRank no presenta problemas con foros con categorías personalizables como AtariAge, detectando un ranking que la jerarquía del propio foro no considera.
- **Identificación de comunidades.** Se propone realizar la identificación a dos niveles; primero, por usuarios según sus relaciones y segundo, la agrupación de temas y de usuarios resultantes de la proyección o aplanado del grafo bipartido que forman los usuarios con los temas en los que comentan.
- **Comunidades de usuarios. Los usuarios se pueden distribuir en grupos mediante algoritmos existentes como LPA o Louvain**.
- **Comunidades temáticas. Para la identificación de comunidades temáticas se propone enlazar a los usuarios con los temas de los que comentan, una relación que además debería ser valorada con un peso correspondiente al número de mensajes escritos en cada tema. De no ser así, muchos usuarios tienen enlaces a todos los temas y tendría el mismo valor escribir una vez que centenares de veces. El grafo resultante es bipartido y se puede proyectar sobre los temas y los usuarios. Sobre los temas es poco problemático, porque no deberían ser muchos, pero la proyección sobre los usuarios tiene problemas de tres tipos que se resuelven en esta propuesta metodológica. 1) No debería valorarse de la misma forma un mensaje sobre un tema que tiene millones de mensajes que un mensaje en un tema que sólo tiene 10. Para ponderarlo, el algoritmo de proyección incorpora un descuento al estilo de lo propuesto por Newman. 2) En casos como AtariAge se generan centenares de millones de enlaces. A nivel de nodos no es un grafo problemático, pero a nivel de enlaces no cabe en la memoria RAM de un ordenador medio actual. Así, es**

necesario volcar los enlaces individualmente o en pequeños grupos a un fichero para evitar el bloqueo. 3) El fichero de aristas resultante es igualmente grande; intentar cargarlo en un array en memoria (un *dataframe* de *pandas*, por ejemplo), resulta en saturación. Así, es necesario establecer un filtro (de peso, por ejemplo) en el propio fichero para cargar únicamente las filas relevantes. 4) La carga en programas de visualización de un grafo con millones de aristas carece de sentido, así que el grafo se trabaja únicamente en código. 5) Dado el elevado número de aristas, el algoritmo de Girvan-Newmann no es adecuado (su complejidad es cuadrática en número de aristas por número de nodos). Louvain obtiene resultados interesantes, así que se propone el uso del módulo *python-louvain* en NetworkX (paquete que hay que cargar por separado al no estar incluido en la versión base). Una vez identificadas las comunidades es posible caracterizarlas a partir del balance entre temas que contienen, perfilando sobre qué temas comentan los usuarios (la estrecha relación entre ellos ya viene asegurada por el paso anterior).

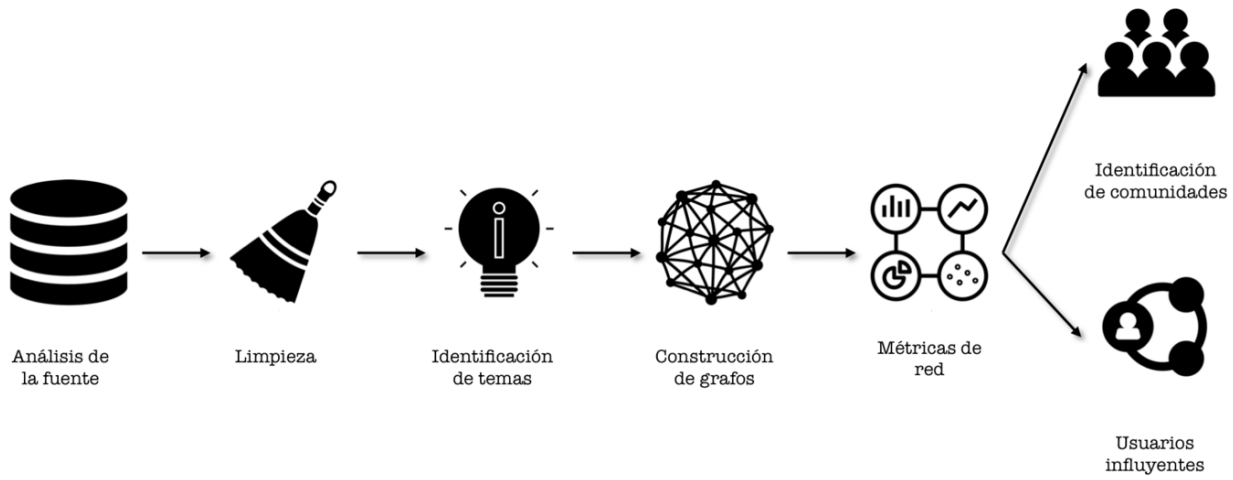


Figure 26: Flujo de trabajo propuesto

Uno de los puntos débiles de esta metodología es su limitada aplicabilidad a foros que no sigan la estructura tradicional de hilos y mensajes. Su aplicación en un foro de preguntas y respuestas o en un foro que permita comentarios como un nivel inferior al de los mensajes requerirá modificaciones en la construcción de documentos para la cadena NLP y en la generación de relaciones entre nodos del grafo. No obstante, el

flujo general no se vería afectado y los pasos de optimización de NLP y de identificación de comunidades se podrían aplicar sin modificaciones.

La fase de identificación de temas queda muy limitada en foros multilenguaje. En la discusión de se menciona una posible aproximación, aunque es probable que requiera de una fase completamente nueva entre las fases de limpieza e identificación de temas.

6.3 Consideraciones específicas

A continuación se enumeran una serie de consideraciones que se podrían considerar consejos a la hora de afrontar un proceso de análisis de foros online y que se han derivado del trabajo realizado:

- La extracción utilizando una herramienta de *scrapping* como Scrapy es una buena opción, pero aún así aparecen problemas derivados de su uso, como las fechas en formato no estándar (*Today*), campos que varían según los hilos o atributos que no existen para todos los usuarios. Los errores en la extracción son difíciles de detectar (normalmente no se encuentran hasta el análisis) lo que implica que son costosos de arreglar. En el caso de Atari Age, por ejemplo, la extracción original no capturó las citas (*quotes*) así que hubo que volver a realizar la extracción y el proceso posterior seis meses después de la primera extracción. Igualmente, la segunda extracción presentó problemas en algunas de las citas, puesto que los usuarios pueden modificar el código que genera la opción de cita en el foro. La forma de atajar este problema sería con un acceso directo a la base de datos del foro, algo posible en caso de que el análisis lo quiera realizar un administrador o propietario.
- Los datos extraídos con Scrapy se han almacenado formato JSON y en CSV. El proceso de análisis ha demostrado que lo más adecuado es hacerlo en JSON, puesto que el CSV presenta problemas evidentes cuando los usuarios incluyen comas en su nombre o en otros campos (que, en algunos casos, han incluso impedido la carga del fichero).
- **La limpieza de datos en origen (es decir, tan cerca de los datos originales como sea posible) permite aprovechar la optimización de consultas típica de motores de datos de datos (tanto SQL como NoSQL) y la versatilidad de cacheo de datos en memoria, incluso para conjuntos de trabajo que no caben simultáneamente en memoria. Aunque ha supuesto**

un reto aplicar las fases de análisis de lenguaje y de redes a un foro de cierto tamaño como AtariAge, cualquier motor de BBDD puede manejar consultas de filtrado, selección e incluso expresiones regulares sobre unos pocos gigabytes.

- **Algunas fases del proceso son bastante pesadas computacionalmente, por lo que es muy recomendable persistir los resultados intermedios de cara a retomar la ejecución en cualquier punto. Se puede utilizar para ello el propio sistema de ficheros del motor de procesamiento, o bien alojarlos en un sistema externo como soluciones de almacenamiento en la nube.**
- **El uso de un cluster de Spark ha sido fundamental para la extracción de temas del caso de estudio de AtariAge, ya que es la fase que necesita los recursos más potentes para obtener resultados. Sin embargo, el uso de Spark GraphFrames ralentiza considerablemente el proceso en comparación al tiempo de ejecución sobre NetworkX. Por ejemplo, en NetworkX la ejecución de LPA y PageRank sobre AtariAge tarda menos de 20 segundos, mientras que sobre GraphFrames tarda unos 10 minutos. No obstante, GraphFrames puede ser necesario para procesar grafos que no puedan ser cargados en memoria ya que NetworkX trabaja sobre un único nodo.**
- La identificación de temas tiene varios requerimientos que se consideran importantes: **(la enumeración ya la pondré bien en el latex)**

1. Se observa que, aunque foros como AtariAge tengan sus hilos divididos en foros a dos niveles, los subforos suelen estar contenidos dentro de los foros principales y los usuarios acaban escribiendo sus mensajes indistintamente en unos y otros, así que pierde el sentido considerarlos por separado al analizar los temas de discusión.
2. El modelo de LDA será más estable cuanto mejor calibrado esté, tanto a nivel técnico (para afinar el modelo) como a nivel conceptual (para identificar temas, agruparlos y eliminar *stopwords*). Una elección de parámetros de compromiso puede consitir en escoger a) un número de temas k similar al número de foros y/o subforos, b) un valor de *doc_concentration* cercano a 1 para que los documentos tengan un tema predominante y c) un valor de *topic_concentration* entre 5 y 10 para que los términos puedan aparecen en varios temas. La implementación con *Expectation-Maximization* ofreció unos resultados mucho mejores que *Online Variational Bayes* en todas las ejecuciones.
3. Aunque pueda parecer que la metodología es aplicable en general, debe añadirse un requerimiento: disponer de uno o más investigadores con dominio del campo semántico de estudio. En caso contrario

puede verse comprometida la clasificación de temas, eliminación de *stopwords* e iteraciones del proceso en general.

4. Es difícil escoger un diccionario de *stopwords* útil simplemente examinando los documentos. Una posible aproximación es aplicar LDA con un diccionario básico (por ejemplo, el diccionario por defecto del idioma), observar los términos que componen cada tema, y ampliar el diccionario con las palabras que no aportan información al tema o que aparecen en tantos temas que no ayudan a diferenciar unos de otros. En las primeras ejecuciones aparecerán muchas flexiones de verbos y muchos artículos y adverbios que no están en el dicionario por defecto (este fue el caso de *always* o *anyway*) así como palabras propias del dominio del foro pero tan repetidas que tampoco aportan información (ejemplos de este caso fueron *virtual*, *boy*, *play*, *jpg* o *kb*. En unas pocas ejecuciones, que incluso pueden lanzarse con un subconjunto de los datos, es posible construir un diccionario útil.
5. Hay términos que claramente ayudan a definir un tema pero que LDA no puede identificar como tal: por ejemplo, valores de rangos de memoria, direcciones web o precios. Estas cadenas de texto toman valores diferentes pero transmiten el mismo concepto a un humano: desarrollo, enlace web, compaventa. LDA identificará los temas de una manera más consistente si estas cadenas se transforman en un *token* único: las apariciones de 0xFFFFFFFF y 0x102D391B pasan a ser ***MEM*** en ambo casos, las cadenas que empiecen por http:// o https:// pasan a ser ***URL*** y los números que sigan el formato $19.95 se convierten en ***MONEY***. También puede ser razonable mantener ambas cadenas, o aplicar otro tipo de transformación: tanto los foros de compraventa como los sociales incluian muchas URLs, pero principalmente de *ebay* en el primer caso y de *youtube* en el segundo, así que extraer el nombre del dominio como un término individual fue fundamental en este paso.

- El trabajo con el grafo bipartido de temas-usuarios debe considerar el peso de la relación; de no hacerlo se tendrán problemas en la identificación de comunidades, porque a nivel binario de relación una gran mayoría están relacionados. Lo realmente interesante en este punto es detectar las relaciones "fuertes".
- Para intentar inferir información sobre los perfiles de las comunidades temáticas es útil ver qué tipo de mensajes escriben o en qué temáticas. Pero intentar visualizarlo es improductivo (más teniendo en cuenta que la mayoría de comunidades están también ligadas al resto de comunidades por múltiples enlaces).
- **La metodología propuesta aúna análisis con una base de funcionamiento muy diferente en una única cadena de procesamiento. Conviene definir la cadena de manera holística desde el principio, identificando especialmente las entradas y salidas de una fase. Un campo en**

origen que no se usa en la identificación de temas (por ejemplo, si el mensaje contiene citas) puede ser necesario en la fase de construcción del grafo. Una visión completa de la cadena evitará tener que reescribir (y sobre todo, volver a ejecutar) una fase intermedia que ya se había dado por terminada.

- **El uso de una plataforma única para todas las fases de la metodología facilita el punto anterior. En los casos de estudio se ha aprovechado la versatilidad de Spark para escalar horizontalmente en las fases de mayor carga computacional (extracción de temas), a la vez que ha permitido integrar la ejecución de código de NetworkX sobre el nodo master.**

7. Conclusiones y trabajo futuro

7.1. Conclusiones

El objetivo principal era el de diseñar una metodología que permitiese tratar y analizar las comunidades de usuarios en foros online de forma sistemática, aprovechando la combinación de dos técnicas de ciencia de datos: el tratamento del lenguaje natural y el análisis de redes sociales.

Tras el desarrollo teórico y la aplicación a dos casos de estudio, se ha obtenido una propuesta metodológica que, cumpliendo con los objetivos marcados, presenta una serie de etapas con las que se consigue:

1. Identificar las temáticas que se tratan en el foro mediante técnicas de *Topic Modeling*.
2. Modelar las relaciones entre usuarios mediante sus conexiones sociales, que resultan en la identificación de los usuarios más influyentes y/o prestigiosos.
3. Vincular a los usuarios con los temas que les interesan y, así, obtener información sobre las relaciones entre temas (conjuntos de temas más populares) y las relaciones temáticas (o de interés) entre usuarios (comunidades de interés).

Adicionalmente, se han aplicado técnicas de procesamiento distribuido (*Spark*) y optimizaciones técnicas (en los algoritmos de proyección, por ejemplo) para dar respuesta a los retos que presentan foros como AtariAge, con millones de mensajes y decenas de miles de usuarios.

7.2. Trabajo futuro

La metodología propuesta puede ser ampliada y refinada en multitud de aspectos. Algunos de ellos se resumen a continuación como campos para desarrollo futuro:

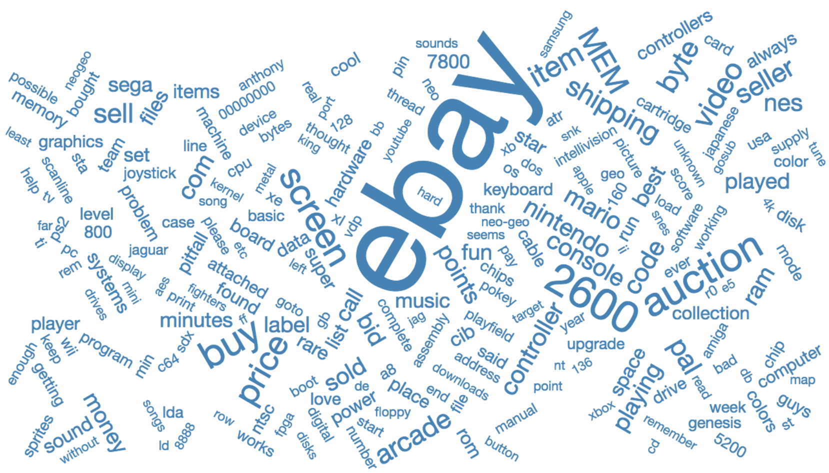
- El rendimiento del algoritmo LDA para la identificación de temas puede mejorarse dedicando más tiempo y recursos a la extracción de características. Hay palabras (*tokens*) que se pierden por errores

de escritura, por limitaciones de la extracción o del HTML (que concatena palabras o añade símbolos indeseados), etc. El *RegexTokenizer* utilizado es básico y también se podría mejorar. Se han encontrado también limitaciones técnicas derivadas de la codificación textual (Unicode, UTF-8,...).

- Incorporar el uso de *stemming* para agrupar las palabras según su raíz podría ser también beneficioso, puesto que, en la situación actual, palabras de uso habitual en el contexto como "game" y "games" se clasifican por separado.
- El diccionario de *stopwords* por defecto era claramente demasiado limitado, lo que ha obligado a añadir mucho trabajo manual para la correcta eliminación de palabras que no aportaban significado. La lista se podría ampliar o incluso se podría ver si en algunos contextos existen diccionarios específicos que eliminen términos que pueden no ser habituales en el lenguaje general pero sí en entornos dominados por usuarios con intereses comunes.
- Las ejecuciones de LDA son relativamente poco estables, lo que conlleva la necesidad de revisar manualmente los términos, ver mensajes de ejemplo y clasificar el contenido de forma manual (mediante más de un revisor). Si se pretende automatizar el proceso al máximo, este paso sería inviable en producción. La solución escapa del alcance de este trabajo, pero se podría estudiar la viabilidad de métodos de etiquetado automático, clusterización a partir de una muestra menor o diccionarios estáticos.
- Esta metodología se ha centrado en el componente textual, pero muchos de los hilos incorporan URLs, imágenes y, sobre todo, emoticonos. Todos estos elementos podrían ser incorporados al análisis, siendo los emoticonos también útiles para un hipotético análisis de sentimiento.
- Siguiendo en la línea anterior, la extensión quizás más directa sería la integración de un análisis de sentimiento a los contenidos de los hilos, mensajes o temas para poder extraer información adicional sobre los mismos.
- El análisis textual ha sido posible, en gran parte, gracias a su idioma: el inglés. Una línea de trabajo futuro sería la expansión a otros idiomas, como el castellano.
- El uso de algoritmos de comunidades se ve limitado a *LPA* y *Louvain* porque son los únicos algoritmos probados cuya complejidad permite la ejecución en tiempos razonables en los grafos de las dimensiones de los casos de estudio. Sería positivo, no obstante, comparar sus resultados con algoritmos mucho más lentos pero que se sabe que son más estables o robustos, como Girvan-Newman.
- De la misma forma, la detección de comunidades en foros grandes, como se ha visto en la discusión, obvia las comunidades pequeñas y las integra en comunidades más grandes. Otra perspectiva a tener en cuenta sería la de variar la granularidad del algoritmo para hacer surgir estas comunidades menores de

interés que, en la propuesta elaborada aquí, se unen en macrocomunidades que para análisis detallados quizás sean demasiado generalistas.

- Se podría añadir una componente temporal a los análisis para obtener una evolución de los temas más debatidos o de la influencia de los usuarios. Además, sería interesante aplicar este método periódicamente para comprobar si detecta correctamente nuevos temas y comunidades.



Apéndice I: Ejemplos de mensajes de Planet Virtual Boy

dev

- Yeah I looking for the Enemy Robot sprites.These are not in there.Would be very interesting to see them. Is there any chance to view the gfx in a Editor like Tilelayer Pro or Tilemolester?Or are they
- I believe not too many people are familiar with the fact that VB Mario's Tennis does have in-game credits! If I remember correctly, if a player beats tournament mode while having it set to the hardest
- I implemented a simple API in the new PVB version which outputs some info for a ROM identified by a given MD5 hash. Example URL:http://www.planetvb.com/rom-info/5b11d402f7e322c71a7d4fa6503631fa/It cur
- InstructionsOnce the CPU is initialized on reset, it can start processing program instructions. If you've ever taken a computer science course, you'll be told that the most general form of CPU operati

- That all sounds awesome. Good luck with everything. It will come in time. One thing VB fans are is patient! Promises I haven't technically failed to deliver on:EmulatorEmulators for video game systems
- Please forgive me for my noobular format. I am so lost. Maybe. Probably. I don't know. Gonna say yes. Files were downloaded and unzipped. I opened the demo1 folder then the demo1c. Computer asks me wh
- Quote:Shokwav wrote:Are you looking for a volunteer programmer for any odd work? I have several years experience in assembly & low-level. Made a few demos for the N64 & NES.That'd be amazing, thanks f
- I want to write a simple I/O program that extracts BGMaps from VB Wario Land ROM and therefore all sprites from the game automatically.Extracting chars from a rom is easy but manual areangement of the
- I agree that the current database could do a much better job at presenting translations, especially all the great work done by Benjamin Stevens.I got a new version of Planet Virtual Boy in the works c
- Hi everyone,Here is an album I released on Bandcamp late last month. All the music except maybe a track or two was designed as a soundtrack for particular independently developed video games made duri

information/scans

- Wow - that's amazing! I knew about the sega 3d glasses, but not the famicom ones.I didn't even know about "Famicom Grand Prix" - the artworks for "3D Hot Rally" are very nice!http://thevideogameartarc
- this is super cool. i can't wait to read this later! thank you for all of your efforts! Great work! Thanks!! :D I decided to make a nice PDF book of my English translations for the Virtual Boy Memoria
- https://www.youtube.com/watch?v=YDepDMiP2GM&t=626s You didn't have to look farther than the front page to see this has already been posted.
- Would be better if you post the articles about the VB GamePro en Español Vol.2 Num.4 Attach file: 01.jpg (791.87 KB) 04.jpg (796.02 KB) 06.jpg (1,031.10 KB) 09.jpg (818.28 KB) 79.jpg (864.52 KB)
- Quote:M-A-D-M-A-X wrote:Thanks for this contribution :)You're welcome. :-) Thanks for this contribution :) According to my friend in Japan, this book was bundled with a magazine called Dengeki Super F

- Virtual Boy Instruction Booklets / Manuals: Jack Bros. Instruction Booklet / Manual (Scans; PDF) Attach file: Jack Bros. Instruction Booklet.pdf Size: 11,454.34 KB; Hits: 47 Virtual Boy Instruction Bo
- Thanks for the great scans, Ben! Did anyone notice how page 39 refers to Mario Clash as Mario Smash (bottom left)? Theories anyone?Also great to finally have a good image of the prototype VB system bo
- A two-page Virtual Boy article appeared on pages 28 and 29 of the August 1995 issue of the Japanese magazine titled "Elementary School Fourth Grade." Attach file: Elementary School Fourth Grade (Fron
- Nice, that's some pretty rare (and even first hand) information! Really hard to find otherwise.Found one more issue mentioning Virtual Boy: http://www.ebay.com/itm/322237325634 Attach file: Nintendo
- Quote:Nanis149 wrote:Can you get "bigger" pictures of the previously lost screenshots?I was planning to create higher resolution scans for those screenshots whenever I get around to making English tra

social

- Quote:Lester Knight wrote:i do like adding pins to certain pieces of clothing. i would be interested depending on the final design. keeps us posted!If this is something people in the community want, I
- You mentioned he had pins. Does he make custom ones? I would kill for a nice enamel VB pin. I believe he ships globally, I would be surprised if not.Grabbed some pins from him before and those were sh
- Quote:JohnSReid wrote:Hi all,I just wanted to say "hi" as I've just signed up here.It's an exciting day in the household as a Virtual Boy that we ordered from Japan has just arrived! It's stubbornly r
- . . . where is it?Yes this has been asked before, possibly many times, but I think it's been a while enough now. Time for another regularly scheduled topic on this lost gem! lolThe thing that got me thin
- Quote:Wyndcrosser wrote:I really have disdain for Pat the NES punk. . . But nice work on the show.There's definitely moments where I wanna kick Pat in the balls but he's much better in person. Ian is wh
- Hello,I have just received the Flashboy+, when I plug it Windows 10 recognizes it, the Flashboy utility says connected but the Flash button stays grey. I tried with padded SD Gundam and Jack Bros on 3
- It must not have been all US ISP's that got blocked, as my girlfriend as well as my best friend were both able to access the site for as many days as I couldn't, but maybe it had something to do with
- I hate how poorly underrepresented the VB is at these events. When I did AVGC the last two years I brought the VB/VG and it got a lot of attention. The curiosity is there for plenty of people.I had a

- Quote:Dreammary wrote:Definitely, that's why I recommended it. Please review Water World next! :)I've got Galactic Pinball in the works next but I will do Waterworld afterwards :) When is the next epi
- on occasion we have also chatted in the channel #VirtualBoy on Efnet, IRC. it isn't widely populated with PlanetVB members, but can be a chill place to idle and communicate about VB stuff. It's an ama

information/scans

- Quote:KR155E wrote:I have briefly talked to her this morning. The video she posted was taken down by request of "stakeholders of the developer" and it's difficult to share information about the game s
- Hi,So I had a chance to go over the translation for the two Mexican pages (Dragon Hopper and Zero Racers) and I would rephrase some segments of the first one. The Zero Racers translation is much more
- First time I ever hear of this one. Unfortunately, the Wikipedia article does not provide any evidence or even proof of the claim that the game was planned for VB. Quote:KR155E wrote:First time I ever
- Hi Everyone!Here's a scan from the Spanish (Spain) edition of Club Nintendo that has a brief note on VB's announcement at Shohinkai. It's from issue 8, Dec/Jan 94/95on page 17. I've translated (from S
- If one wants to play VB games on their flash cart, they need to pad the ROM to some multiple of 2MB IIRC. I updated the build system on my VBdemo repo, and I've been looking at creating my own tool fo
- The other day, my brother-in-law showed me a photo of a school that is located not too far away from where I live. It's an aerial view taken right off of Google Maps. Do you see what I see? Attach fil
- typical modern day reporting... interpreting articles (WIKI and other) to report personal conclusions. i can not recall reading anything that relates eye strain to the automatic "pause" feature this a
- Shows how much research went into that article. reading stuff on the net came across a weird peripheral article, clicked and saw our beloved VB in it:http://www.techradar.com/news/11-weir ... erful-co
- Quote:Nitrosoxide wrote:Anyone make one of these that looks professional enough that they'd think of selling them?I'd love to buy a well made one that looked good aesthetically.I know there's this DIY
- Quote:What can you say about the gr.card Quadro?A Quadro is a heavy lifting card, but gamers won't get any more out of it from an equivalent consumer card, because Quadro is geared more towards profes

hardware/repairs

- Those are cool labels, very clever. I feel like everyone should have theirs soldered as my two VBs work perfectly. I just got two of my VBs back from NES Freak and they look great! He even took the ti
- Bought some card stock paper and printed color versions. Since I didn't have a laminate machine thingy I used a bit of packaging tape as a substitute. Got 8 done so far and they turned out nicely :) A
- Thanks! I used my iPhone 7 for the footage. It's the only decent camera I have for recording at the moment. It took me a good minute of getting things jusssst right but in the end it even shocked me
- I'm based in the US and could do the repairs as well if you're still looking. I bet @RunnerPack could do it too and he is in the US. He did the LED soldering repair for me a few years back and did a g
- Hey all,I have some questions about my vb. First off, is it normal that my vb shows two very different images at certain points in certain games? Second, is it normal that when I take my head away fro
- Um... The oven trick does work. I have two non-soldered VBs to prove it. So, I wouldn't flat out say that. I would not do the oven trick unless you've done your research though. And yes you already di
- Welcome Ben!! Thanks for the pics. After seeing more of the process I was definetly correct in thinking this is out of my skill set. I received my second set of fixed displays from Ben earlier this we
- Hi PlanetVB Community!I was hoping someone here can provide some guidance. I recently recovered my Virtual Boy from storage and set it up. I was disappointed to see no audio or video was being generat
- Do you mean like modifying a tv wall mount so you can attach a virtual boy to it? Or something like this custom vb arcade cabinet? https://s-media-cache-ak0.pinimg.com/7 ... ll-games-arcade-games.jpg
- Hello all.I bought a Virtual Boy near the end of its life cycle, around the time the N64 came out. I put it away about 20 years ago and just recently pulled it out of storage and dusted it off. To my

marketplace

- You'll find such numbers on many items, such as the bottoms of aluminum beverage cans. I'd be willing to bet that the stamped numbers on Nintendo game cartridges indicate the production line on which
- There is already a thread about those replacement parts:http://www.planetvb.com/modules/newbb/viewtopic.php?topi I browsed eBay to check what I need to get a second system - for the link cabl
- I work in the printing industry and I can tell you from my standpoint that these are common problems on any job. The galactic pinball manual has what we would refer to as a hickey. That's where a piec
- Please tell me you weren't the person who spent $12.90 on this ad. :)I have no idea what the ad's from,

but my guess would be any of the mid-90s gaming magazines. If I still had all my old GamePros, I

- Looking to buy another Virtual boy for some paint job and link cable action. I'd be looking for at least a head unit with the eye shade visor and bracket along with a controller. Original stand and ba
- Quote:Planlos1988 wrote:Hello and thanks for the welcome :)And/Und hallo RetroDan, wie meinst denn das? Fragen bezüglich Problemen oder Technik oder eher in Richtung verkauf von VB Sachen?(And hello,
- Thanks very much for this Ben - I will try and get a JP version translated to complement the sheer effort put into all of this. So are you looking for actual copies from the original limited productio
- Thanks for the reply, I've read it over, but I have a question about disassembling the unit. After doing some research, I see that I require a 4.5 gamebit, which I'll need to buy. Could you recommend
- this is the hat I own. Could rip the patch off of mine, put it on your purple one and have the one from the ad. Attach file: vbhat.png (569.33 KB) never heard of a contest. cool hat though. i have on
- As above I need a new right eye display. Sadly mine has some dead LED's that cannot be repaired :'(If anyone has one, or a unit they are looking to sell with a working right eye piece please drop me a

misc

- Found a Primal Rage figure/trophy (I think the character's name is Chaos?) today at a Goodwill and can't find anything about it online. Other than Time Warner Interactive and the Primal Rage logo, the
- I have no words to describe how much I hate Buzzfeed. I thought "as long as they picked a good game this shouldn't be too bad". Enter Mario Clash :/ https://www.youtube.com/watch?v=1Cu_9YPLuFk Seems t
- You are so right! They ARE backwards. Good eyes. Quote:L___E___T wrote:I know the artist that made that image it's Adam Rufino and he actually sells it as a poster. Here's his artwork page on Behance:
- Quote:ziggaboogi wrote:That one looks like they changed the recommended ages. The original print with the mom picture basically says to not let kids play the VB unsupervised. But that section is chang
- Wow-za... that's literally it.I can't believe he didn't even attempt to crop it or anything. xD Quote:Splain wrote:Already quotes VirtuousRage from this very thread. I'm famous˜!! :D Eh.Color me faith
- I was geekin' at this probably more than I should have. So I just thought to post it here, in case anyone

else can use a good chuckle! ?https://youtu.be/UFEq2_F9Gsg ? Brilliant. I think id like to hav

- That seems like a great site for buying incredibly outdated technology.I see they also have 10 "open box" FaxViews for $41.99 each:https://www.buymebuyme.com/buy/FaxView ... ments-Anywhere-47410.htmlT
- Quote:chicgamer wrote:Are you saying that Snake Eyes and Scarlett wouldn't enjoy some arcade gaming? ;)I don't have any of these newer machines, but my brothers had a lot of the old electronic games (
- I am in Philly I grew up in Texas. :DOriginally from New Braunfels!But, now I'm in the Seattle area. :S any other texan vb fans around? Quote:astro187 wrote:Never been before but always wanted to go.
- Quote:TerryJ wrote:Thanks for the suggestions so far!Just to clarify on my first post, I'll happily take any GBC & GBA recommendations too.Do you have a GBA flash kit? I can link you to some amazing h

non-english

- I was playing SD gundam today, and I was having a hard time figuring out how to use the "ambush" action. Every time I used it, nothing happened. Anyone know how this works? Wow, thank you for the thor
- Kollade runt pa tradera och denna kille fick mig att skrattahttp://www.tradera.com/Antenn-Klockra ... el-auktion_1708_128458026Han har ej testat det mesta och alla telefoner har ett eller annat proble
- Hallo ihr Lieben,ich befinde mich in einer misslichen Lage und muss umbedingt rausfinden ob es Verschidene Virtual Boy Versionen gibtSprich: einmal eine (von Oben ansicht) wo man den Text Fokus lesen k
- Thanks a lot - vielen Dank! Was ist die beste bzw. einfachste und zugleich sichere Methode, den VritualBoy an das deutsche Stromnetz anzuschliessen?Kann man z.B. ein handelsubliches Mulit-Netzteil an d
- Sa sant sa men kul att ha en trad dar vi kan vara lite hemliga av oss, man kanner sig lite mer ball trots allt :-Pglom inte nu finns speciela versionen av bound high sa passa pa att skaffa den medans
- I try make Virtual Boy Faceball (NikoChan Battle) Gameplay STAGE 1-6 to STAGE 1-10 video by same camera and video have low resolution like other but you can get idea of game movement and enemy attacks

- Ne, testade mest i reality boy och tex yeti3d blinkade allt som var javligt irriterande men addade dock chu chu rocket lanken pa videon i youtubeTack iofs att du gillade mina virtual boy videor pa you
- Trying to compile a demo... I got this:$ makev810-gcc -Wall -nodefaultlibs -mv810 -xc -o demo1.o demo1.cIn file included from ../libgccvb/affine.h:12, from ../libgccvb/libgccvb.h:21, from demo1.c:2:..
- Hi!I'm new here like my Virtual Boy which I got last week :)I was born 1986 and the 90s were my best time playing video games. But I never had the chance to get a Virtual Boy. So my wish come true and
- I haven't seen anything mentioning the game's story, yet. So I'm currently making a video on Virtual Lab and I got the story translated to English. I made it just for the VB fans, collectors and futur

misc

- 22 Vectrex eBay listings started last night, all with $9.99 opening bids. 19 games, 3 accessories. More coming later.www.ebay.com/sch/jasonbar/m.htmlThanks,-Jason [In addition to my eBay auctions, I'm
- Quote:WoLfMaN wrote:Thank you for taking your time and writing down all that. There's some very useful info in that for me.Quote:Benjamin Stevens schrieb:more and even get it safely resting on your loing your time and writing down all that. There's some very useful info in that for me.Quote:Benjamin Stevens schrieb:more and even get it safely resting on your lo
- How do you clean VB games? :)Thanks. They are supposed to self clean when you play them. How do you clean the system then?
- I don't feel it.Not like Steve Irwin or Christopher Reeve. http://www.timesonline.co.uk/tol/spor ... _sport/article2461339.eceNo comments...

information/scans

- Joypan October 1995 - Issue #46 Attach file: Joypad%20046%20-%20Page%20001%20(1995-10).jpg (549.67 KB) Joypad%20046%20-%20Page%20086%20(1995-10).jpg (634.23 KB) I figure it's probably just a mistran
- I was surprised to see that this magazine was not available here on the site so, i decided to throw a link to A LOT of french magazines that most likely wrote stuff about the Virtual Boy :-)But i'll p
- Included in this post are extremely high resolution (1200 dpi) scans of all screenshots for the lost Virtual Boy games appearing in Weekly Famitsu Magazine. By "lost," I mean unreleased Virtual Boy ga

- Jugemu: No. 5 - September 1995Virtual Boy Game Articles:Mario's TennisGalactic Pinball (2)Mario ClashTeleroboxerRed Alarm (2)V-Tetris (2)Hee-Haw in the Maze of Jack Bros.!Virtual FishingPop-Out! Panic
- Famitsu 349 Attach file: Famitsu 349 Cover.jpg (2,483.65 KB) Famitsu 349 p 029.jpg (2,517.04 KB) Famitsu 349 p 033.jpg (2,745.45 KB) Famitsu 349 p 092.jpg (2,581.31 KB) Famitsu 349 p 093.jpg (2,8
- Even though there are several games on R-Zone, only 4 types exist and all games are built on these three types... Side-scrolling FighterFPS (First Person Shooter)Multidirectional Combatant FighterRacer

Apendice II: Listado completo de temas de AtariAge

#0: ['byte', 'playfield', 'scanline', 'end', 'rem', 'player', '00000000']

#1: ['2600', 'drive', 'disk', 'board', '7800', 'power', 'run']

#2: ['screen', 'code', 'st', 'c64', 'graphics', 'disk', 'hardware']

#3: ['ebay', 'auction', 'item', 'seller', 'price', 'shipping', 'bid']

#4: ['2600', 'minutes', 'team', 'ebay', 'week', 'label', 'min']

#5: ['2600', 'pal', 'label', 'ntsc', '7800', 'controller', 'ebay']

#6: ['music', 'sound', 'song', 'e5', 'sounds', 'files', 'screen']

#7: ['buy', 'nintendo', '2600', 'jaguar', 'console', 'nes', 'wii']

#8: ['2600', 'arcade', 'played', 'playing', 'fun', 'best', 'screen']

#9: ['code', 'ti', 'file', 'program', 'data', 'address', 'r0']

#10: ['MEM', '800', 'ram', 'board', 'keyboard', 'os', 'upgrade']

#11: ['2600', 'arcade', 'video', 'label', 'collection', 'fun', 'list']

#12: ['call', 'downloads', 'print', '128', 'basic', 'goto', 'screen']

#13: ['data', 'map', 'row', 'rem', 'call', 'number', 'set']

#14: ['c64', 'screen', 'graphics', 'hardware', 'computer', 'video', '7800']

#15: ['2600', 'controller', '5200', '7800', 'joystick', 'controllers', 'best']

#16: ['nt', 'console', 'said', 'mini', 'buy', 'hardware', 'video']

#17: ['nes', '2600', 'minutes', 'console', 'controller', 'nintendo', 'systems']

#18: ['code', 'sta', 'lda', 'program', 'file', 'line', 'ram']

#19: ['drive', 'disk', 'card', 'pal', 'cable', 'disks', 'works']

#20: ['screen', 'color', 'player', 'colors', 'graphics', 'code', 'set']

#21: ['st', 'computer', 'pc', 'hardware', 'amiga', 'software', 'drive']

#22: ['video', 'computer', '2600', 'best', 'power', 'said', 'buy']

#23: ['MEM', 'db', 'de', 'cartridge', 'number', 'power', 'video']

#24: ['2600', '7800', 'screen', 'found', 'video', 'space', 'best']

#25: ['points', 'arcade', 'super', 'played', 'mario', 'star', 'wii']

```
#26: ['7800', 'controller', 'console', 'video', '5200', '2600', 'controllers']
#27: ['ebay', 'auction', 'price', '2600', 'sell', 'item', 'shipping']
#28: ['power', 'controller', 'console', 'cable', '2600', 'video', 'supply']
#29: ['ebay', 'buy', 'video', 'money', 'shipping', 'price', 'console']
#30: ['ebay', 'unknown', 'drive', '2600', 'problem', 'computer', 'working']
#31: ['disk', 'drive', '160', 'atr', 'dos', 'files', 'card']
#32: ['ff', 'screen', 'code', 'color', 'mode', 'line', 'hardware']
#33: ['jaguar', '2600', 'video', 'jag', 'best', 'console', 'playing']
#34: ['nes', '2600', 'buy', 'price', 'ebay', 'sega', 'super']
#35: ['2600', 'video', 'nes', 'arcade', 'nintendo', 'buy', 'console']
#36: ['guys', 'neo', 'anthony', 'geo', 'japanese', 'video', 'arcade']
#37: ['rom', '2600', 'space', 'cartridge', 'screen', 'list', 'pal']
#38: ['2600', 'video', '7800', 'best', 'power', 'buy', 'console']
#39: ['2600', 'controller', 'buy', 'video', 'console', 'said', 'best']
#40: ['guys', 'screen', 'always', 'video', 'thread', 'please', 'level']
#41: ['controller', 'nes', 'super', 'youtube', 'controllers', 'console', 'genesis']
#42: ['digital', 'gb', 'video', 'console', 'tv', 'buy', 'xbox']
#43: ['2600', '7800', 'best', 'said', 'computer', 'point', 'hardware']
#44: ['c64', 'sound', 'mode', 'a8', 'screen', 'sprites', 'cpu']
```

Apendice III: Transformador personalizado para Spark MLlib

```
import string
import re

from pyspark import keyword_only
from pyspark.ml import Transformer
from pyspark.ml.param.shared import HasInputCol, HasOutputCol
from pyspark.sql.functions import udf
```

```
from pyspark.sql.types import ArrayType, StringType



class WordNoiseRemover(Transformer, HasInputCol, HasOutputCol):

    @keyword_only
    def __init__(self, inputCol=None, outputCol=None):
        super(WordNoiseRemover, self).__init__()
        kwargs = self._input_kwargs
        self.setParams(**kwargs)

    @keyword_only
    def setParams(self, inputCol=None, outputCol=None):
        kwargs = self._input_kwargs
        return self._set(**kwargs)



    def _transform(self, dataset):

        def f(s):
            terms = list()
            domains = re.compile("https?:\/\/([a-z0-9.-]+)\/")
            for t in s:
                add = True
                term_to_add = ""
                url = re.match(domains, t)
                if url is not None:
                    term_to_add = url.group(0)
                elif "0x" in t:
                    term_to_add = "MEM"
                else:
```

```
                    stripped = t.strip(string.punctuation + " ")
                    add = True
                    for c in "()!#/\\:?!@":
                        if c in stripped:
                            add = False
                    try:
                        term_to_add = stripped.decode('ascii', errors='ignore')
                    except:
                        add = False

                if add and len(term_to_add) > 1:
                    terms.append(term_to_add)
            return terms

        t = ArrayType(StringType())
        out_col = self.getOutputCol()
        in_col = dataset[self.getInputCol()]
        return dataset.withColumn(out_col, udf(f, t)(in_col))
```

Apéndice IVa: Algoritmo de proyección sin optimización

```
def forum_projection(B, nodes):
    if B.is_directed():
        pred = B.pred
        G = nx.DiGraph()
    else:
        pred = B.adj
        G = nx.Graph()
    G.graph.update(B.graph)
    G.add_nodes_from((n, B.nodes[n]) for n in nodes)
```

```
    for u in nodes:
        unbrs = set(B[u]) #nodos a los que conecta del otro nivel
        nbrs2 = set(n for nbr in unbrs for n in B[nbr] if n != u) #nodos con los que coincide
        for v in nbrs2:
            vnbrs = set(pred[v])
            common_degree = ((len(B[n]), B[u][n]['weight'] + B[v][n]['weight']) for n in unbrs & vnbrs)
            weight = sum(wt / (deg - 1) for deg, wt in common_degree if deg > 1)
            G.add_edge(u, v, weight=weight)
    return G
```

Apéndice IVb: Algoritmo de proyección con optimización

```
def forum_projection_to_file(B, nodes, other_nodes, filename):
    lengths = {n: len(B[n]) for n in other_nodes}
    if B.is_directed():
        pred = B.pred
    else:
        pred = B.adj
    vnbrs_sets = {n: set(pred[n]) for n in nodes}

    with open('{}.csv'.format(filename), 'w', encoding="utf8") as file:
        for u in nodes:
            unbrs = set(B[u]) #nodos a los que conecta del otro nivel
            nbrs2 = {n for nbr in unbrs for n in B[nbr] if n != u} #nodos con los que coincide
            for v in nbrs2:
                vnbrs = vnbrs_sets[v]
                common_degree = ((lengths[n], B[u][n]['weight'] + B[v][n]['weight']) for n in unbrs & v
                weight = sum(wt / (deg - 1) for deg, wt in common_degree if deg > 1)
                file.write("{},{},{:.8f}\n".format(u,v,weight))
                del common_degree
                del weight
```

```
del unbrs

del nbrs2
```

Comunidad	Tamaño relativo	Tema	Porcentaje	Interpretación
Comunidad 0	33,69%	Dev	**40,2%**	Comunidad formada principalmente por desarrolladores, que comparten detalles sobre su código y dudas con la comunidad.
		Hardware/Repairs	10,5%	
		Information/Scans	2,8%	
		Marketplace	13,9%	
		Misc	5,0%	
		Non-English	1,3%	
		Social	26,2%	
Comunidad 1	3,40%	Dev	5,7%	Esta es una reducida comunidad de usuarios cuyos mensajes suelen ser escritos en un idioma distinto del inglés (se ha detectado alemán y sueco, básicamente). Cuando escriben en inglés lo hacen para socializar o participar en transacciones.
		Hardware/Repairs	5,0%	
		Information/Scans	0,4%	
		Marketplace	15,2%	
		Misc	6,6%	
		Non-English	**39,7%**	
		Social	27,2%	
Comunidad 2	34,15%	Dev	9,8%	Se trata de un numeroso grupo de usuarios centrados en la compra y venta de artículos. Sus mensajes son dedicados a las transacciones económicas o a la socialización - muy asociada a lo primero. Es probable que sean coleccionistas.
		Hardware/Repairs	6,0%	
		Information/Scans	3,2%	
		Marketplace	**48,3%**	
		Misc	5,8%	
		Non-English	0,8%	
		Social	26,1%	
Comunidad 3	17,40%	Dev	11,1%	Esta parte de usuarios está notablemente más dedicada que el resto a la interacción social. Intercambian comentarios y comparten su opinión, así que son aficionados, en general, a la Virtual Boy.
		Hardware/Repairs	11,6%	
		Information/Scans	2,8%	
		Marketplace	20,2%	
		Misc	4,7%	
		Non-English	0,6%	
		Social	**49,0%**	
Comunidad 4	11,35%	Dev	17,4%	El elevado número de mensajes relacionados con el hardware y sus reparaciones hace pensar que se trata de usuarios que son o bien "manitas"(gente que trabaja en proyectos o que repara consolas) o bien que buscan ayuda con un problema técnico.
		Hardware/Repairs	**32,1%**	
		Information/Scans	2,3%	
		Marketplace	17,6%	
		Misc	4,4%	
		Non-English	0,4%	
		Social	25,7%	

Table 16: Resumen de comunidades y análisis en PlanetVB. Aparecen cinco comunidades relativamente bien definidas.

Elemento	Cantidad
Foros	32
Subforos	66
Hilos	235.483
Posts	3.305.893
Autores	23.417

Subforo	Número de posts	Porcentaje de posts
Classic Gaming General	617752	18,6%
Atari 2600	571937	17,3%
Marketplace	424155	12,8/%
Atari 8-bit Computers	383301	11,5%
Modern Gaming	276998	8,3%
Atari Jaguar	155261	4,6%
Classic Computing	127088	3,8%
Programming	116055	3,5%
Atari 7800	82709	2,5%
High Score Clubs	62185	1,8%

Table 17: AtariAge: dimensiones del foro y número total de posts por subforo (top 10)

	Núm. caracteres por mensaje	Núm. mensajes por hilo	Núm. mensajes por autor
Máximo	124.609	16.229	25.790
Media	424	10	141
Desviación típica	74	85	733

Table 18: AtariAge: resumen de estadísticas

Elemento	Cantidad	% eliminado
Hilos	191.921	18%
Posts	3.135.557	5%
Autores	22.611	3%

Table 19: AtariAge: dimensiones del foro tras la limpieza

author	**quoted_user**
Albert	[]
Albert	[]
Artlover	[Albert]
Bryan	[]
Breakpack	[Artlover ,Albert]
video game addict	[]
JB	[Breakpack, Artlover, Albert]
Bruce Tomlin	[]

Table 20: AtariAge: Referencias antes de ser procesadas

author	**quoted_user**
Artlover	Albert
Breakpack	Artlover
Breakpack	Albert
JB	Breakpack
JB	Artlover
JB	Albert

Table 21: AtariAge: Referencias tras ser procesadas

Usuario	**PageRank**	**Authority**	**Hub**
Albert	1	1	1
Rev	2	14	-
CPUWIZ	3	3	3
Ze_ro	4	-	-
Tempest	5	2	2

Table 22: AtariAge: Resultados de PageRank y HITS (en negrita los usuarios con mayor actividad)

Usuario	**Número de posts**
CPUWIZ	25790
Albert	24446
Tempest	22114
Thomas Jentzsch	16289
Rev	15126

Table 23: AtariAge: Usuarios con mayor actividad

Comunidad	Tamaño relativo	Tema	Porcentaje	Interpretación
Comunidad 0	24,8%	Console S.	**18,4%**	Comunidad formada principalmente por desarrolladores, que comparten detalles sobre su código y dudas con la comunidad.
		Dev	**24,7%**	
		Games	11,6%	
		HSC	0,5%	
		Hardware	**22,6%**	
		Market	5,5%	
		Music	1,3%	
		Non-English	0,1%	
		Social	15,2%	
Comunidad 1	27,6%	Console S.	8,4%	Se trata de un numeroso grupo de usuarios centrados en la compra y venta de artículos. Sus mensajes son dedicados a las transacciones económicas, a los juegos o a la socialización - muy asociada a lo primero. Es probable que sean coleccionistas.
		Dev	2,4%	
		Games	**23,3%**	
		HSC	0,8%	
		Hardware	8,5%	
		Market	**34,7%**	
		Music	0,4%	
		Non-English	0,2%	
		Social	21,2%	
Comunidad 2	14,7%	Console S.	7,4%	Esta parte de usuarios está notablemente más dedicada que el resto a la interacción social. Intercambian comentarios y comparten su opinión, así que son aficionados, en general, a los sistemas retro.
		Dev	2,4%	
		Games	15,2%	
		HSC	1,0%	
		Hardware	14,1%	
		Market	12,1%	
		Music	0,4%	
		Non-English	0,0%	
		Social	**47,1%**	
Comunidad 3	15,5%	Console S.	8,0%	En esta comunidad se encuentran usuarios que habitualmente hablan de juegos pero que lo hacen desde una perspectiva más de jugarlos que de coleccionarlos, como se nota por su elevado índice en High Score Club.
		Dev	5,8%	
		Games	**19,7%**	
		HSC	**22,4%**	
		Hardware	9,5%	
		Market	12,3%	
		Music	0,7%	
		Non-English	0,0%	
		Social	21,5%	
Comunidad 4	17,4%	Console S.	13,2%	La última comunidad es un conjunto de usuarios que muestra cierto equilibrio en sus actividades, parecido a la comunidad 0 pero sustituyendo Dev por Social; es así probable que sea una comunidad con más soporte técnico que desarrollo.
		Dev	7,6%	
		Games	17,7%	
		HSC	0,1%	
		Hardware	15,4%	
		Market	18,5%	
		Music	0,8%	
		Non-English	0,1%	
		Social	26,4%	

Table 24: Resumen de comunidades y análisis en AtariAge. Aparecen cinco comunidades relativamente bien definidas.

References

Rasha A. Abdulla. Islam Jihad, and Terrorism in Post-9/11 Arabic Discussion Boards. *Journal of Computer-Mediated Communication*, 12(3):1063–1081, apr 2007. doi: 10.1111/j.1083-6101.2007.00363.x. URL https://doi.org/10.1111%2Fj.1083-6101.2007.00363.x.

Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. On smoothing and inference for topic models. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 27–34. AUAI Press, 2009.

Albert-László Barabási. *Network science*. Cambridge university press, 2016.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10): P10008, oct 2008. doi: 10.1088/1742-5468/2008/10/p10008. URL https://doi.org/10.1088%2F1742-5468%2F2008%2F10%2Fp10008.

Martin G. Everett Borgatti, Stephen P. and Jeffrey C. Johnson. *Analyzing social networks*. Sage, 2018.

Stephen P. Borgatti. Centrality and network flow. *Social Networks*, 27(1):55–71, jan 2005. doi: 10.1016/j.socnet.2004.11.008. URL https://doi.org/10.1016%2Fj.socnet.2004.11.008.

Ronald L. Breiger. The Duality of Persons and Groups. *Social Forces*, 53(2):181, dec 1974. doi: 10.2307/2576011. URL https://doi.org/10.2307%2F2576011.

Heith Copes and J. Patrick Williams. Techniques of Affirmation: Deviant Behavior Moral Commitment, and Subcultural Identity. *Deviant Behavior*, 28(3):247–272, mar 2007. doi: 10.1080/01639620701233167. URL https://doi.org/10.1080%2F01639620701233167.

Sabina Remmers de Vries and Albert A. Valadez. Let Our Voices Be Heard: Qualitative Analysis of an Internet Discussion Board. *Journal of Creativity in Mental Health*, 3(4):383–400, dec 2008. doi: 10.1080/15401380802530617. URL https://doi.org/10.1080%2F15401380802530617.

Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. *Commun.*

ACM, 51(1):107–113, jan 2008. ISSN 0001-0782. doi: 10.1145/1327452.1327492. URL http://doi.acm.org/10.1145/1327452.1327492.

M.G. Everett and S.P. Borgatti. The dual-projection approach for two-mode networks. *Social Networks*, 35 (2):204–210, may 2013. doi: 10.1016/j.socnet.2012.05.004. URL https://doi.org/10.1016%2Fj.socnet.2012.05.004.

Scott Feld and William C. Carter. Foci of activity as changing contexts for friendship. In Rebecca G. Adams and Graham Allan, editors, *Placing Friendship in Context*, pages 136–152. Cambridge University Press, 1998. doi: 10.1017/cbo9780511520747.008. URL https://doi.org/10.1017%2Fcbo9780511520747.008.

Aleksandra Galasińska. Leavers and stayers discuss returning home: Internet discourses on migration in the context of the post-communist transformation. *Social Identities*, 16(3):309–324, may 2010. doi: 10.1080/13504630.2010.482416. URL https://doi.org/10.1080%2F13504630.2010.482416.

Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The Google file system. In *Proceedings of the nineteenth ACM symposium on Operating systems principles - SOSP '03*. ACM Press, 2003. doi: 10.1145/945445.945450. URL https://doi.org/10.1145%2F945445.945450.

Jack Glaser, Jay Dixit, and Donald P. Green. Studying Hate Crime with the Internet: What Makes Racists Advocate Racial Violence? *Journal of Social Issues*, 58(1):177–193, jan 2002. doi: 10.1111/1540-4560.00255. URL https://doi.org/10.1111%2F1540-4560.00255.

Mark S. Granovetter. The Strength of Weak Ties. In *Social Networks*, pages 347–367. Elsevier, 1977. doi: 10.1016/b978-0-12-442450-0.50025-0. URL https://doi.org/10.1016%2Fb978-0-12-442450-0.50025-0.

Srinath Perera; Thilina Gunarathne. *Hadoop MapReduce Cookbook*. Packt Publishing, 2015.

Frank Harary. *Graph theory*. Addison-Weasley, 1969.

Wu He, Harris Wu, Gongjun Yan, Vasudeva Akula, and Jiancheng Shen. A novel social media competitive analytics framework with sentiment benchmarks. *Information & Management*, 52(7):801–812, nov 2015. doi: 10.1016/j.im.2015.04.006. URL https://doi.org/10.1016%2Fj.im.2015.04.006.

R. A. Hill and R. I. M. Dunbar. Social network size in humans. *Human Nature*, 14(1):53–72, mar 2003. doi: 10.1007/s12110-003-1016-y. URL https://doi.org/10.1007%2Fs12110-003-1016-y.

Matthew Hoffman, Francis R Bach, and David M Blei. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864, 2010.

Peter Holtz and Wolfgang Wagner. Essentialism and attribution of monstrosity in racist discourse: Right-wing internet postings about Africans and Jews. *Journal of Community & Applied Social Psychology*, 19 (6):411–425, nov 2009. doi: 10.1002/casp.1005. URL https://doi.org/10.1002%2Fcasp.1005.

Peter Holtz, Nicole Kronberger, and Wolfgang Wagner. Analyzing Internet Forums. *Journal of Media Psychology*, 24(2):55–66, jan 2012. doi: 10.1027/1864-1105/a000062. URL https://doi.org/10.1027%2F1864-1105%2Fa000062.

Russel Jurney. *Agile Data Science 2.0*. O'Reilly, 2017.

Charles Kadushin. *Understanding social networks: Theories, concepts, and findings*. OUP USA, 2012.

Jon M Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5): 604–632, 1999.

David Krackhardt and Jeffrey R. Hanson. Informal Networks: The Company behind the Chart. In *Creative Management and Development Creative management and development*, pages 191–196. SAGE Publications Ltd, 2006. doi: 10.4135/9781446213704.n15. URL https://doi.org/10.4135%2F9781446213704.n15.

Haewoon Kwak, Yoonchan Choi, Young-Ho Eom, Hawoong Jeong, and Sue Moon. Mining communities in networks. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference - IMC '09*. ACM Press, 2009. doi: 10.1145/1644893.1644930. URL https://doi.org/10.1145%2F1644893.1644930.

Sara Landset, Taghi M. Khoshgoftaar, Aaron N. Richter, and Tawfiq Hasanin. A survey of open source tools for machine learning with big data in the Hadoop ecosystem. *Journal of Big Data*, 2(1), nov 2015. doi: 10.1186/s40537-015-0032-1. URL https://doi.org/10.1186%2Fs40537-015-0032-1.

P.F. Lazarsfeld and R.K. Merton. *Freedom and Control in Modern Society*. Berger M., 1954.

Margaret F. Moloney, Alexa S. Dietrich, Ora Strickland, and Stuart Myerburg. Using Internet Discussion Boards as Virtual Focus Groups. *Advances in Nursing Science*, 26(4):274–286, oct 2003. doi: 10.1097/00012272-200310000-00005. URL https://doi.org/10.1097%2F00012272-200310000-00005.

J. L. Moreno. *Who shall survive?: A new approach to the problem of human interrelations*. Nervous and

Mental Disease Publishing Co, 1934. doi: 10.1037/10648-000. URL https://doi.org/10.1037%2F10648-000.

Paolo Nesi, Gianni Pantaleo, and Gianmarco Sanesi. A hadoop based platform for natural language processing of web pages and documents. *Journal of Visual Languages & Computing*, 31:130 – 138, 2015. ISSN 1045-926X. doi: https://doi.org/10.1016/j.jvlc.2015.10.017. URL http://www.sciencedirect.com/science/article/pii/S1045926X15000749. Special Issue on DMS2015.

M. E. J. Newman. Scientific collaboration networks. II. Shortest paths weighted networks, and centrality. *Physical Review E*, 64(1), jun 2001a. doi: 10.1103/physreve.64.016132. URL https://doi.org/10.1103%2Fphysreve.64.016132.

M. E. J. Newman. Scientific collaboration networks. II. Shortest paths weighted networks, and centrality. *Physical Review E*, 64(1), jun 2001b. doi: 10.1103/physreve.64.016132. URL https://doi.org/10.1103%2Fphysreve.64.016132.

M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, may 2006. doi: 10.1073/pnas.0601602103. URL https://doi.org/10.1073%2Fpnas.0601602103.

M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2), feb 2004. doi: 10.1103/physreve.69.026113. URL https://doi.org/10.1103%2Fphysreve.69.026113.

Bart Noteboom. Simmel's Treatise on the Triad (1908). *Journal of Institutional Economics*, 2(03):365, oct 2006. doi: 10.1017/s1744137406000452. URL https://doi.org/10.1017%2Fs1744137406000452.

Tore Opsahl. Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. *Social Networks*, 35(2):159–167, may 2013. doi: 10.1016/j.socnet.2011.07.001. URL https://doi.org/10.1016%2Fj.socnet.2011.07.001.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

Brea L. Perry, Bernice A. Pescosolido, and Stephen P. Borgatti. *Egocentric Network Analysis*. Cambridge University Press, feb 2018. doi: 10.1017/9781316443255. URL https://doi.org/10.1017%2F9781316443255.

F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences*, 101(9):2658–2663, feb 2004. doi: 10.1073/pnas.0400054101. URL https://doi.org/10.1073%2Fpnas.0400054101.

Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76(3), sep 2007. doi: 10.1103/physreve.76.036106. URL https://doi.org/10.1103%2Fphysreve.76.036106.

Sean Owen Josh Wills Sandy Ryza, Uri Laserson. *Advanced Analytics with Spark, 2nd edition*. O'Reilly, 2017.

Piotr Semberecki and Henryk Maciejewski. Distributed Classification of Text Documents on Apache Spark Platform. In *Artificial Intelligence and Soft Computing*, pages 621–630. Springer International Publishing, 2016. doi: 10.1007/978-3-319-39378-0_53. URL https://doi.org/10.1007%2F978-3-319-39378-0_53.

Georg Simmel. *The sociology of georg simmel*, volume 92892. Simon and Schuster, 1950.

Linda J. Skitka and Edward G. Sargis. The Internet as Psychological Laboratory. *Annual Review of Psychology*, 57(1):529–555, jan 2006. doi: 10.1146/annurev.psych.57.102904.190048. URL https://doi.org/10.1146%2Fannurev.psych.57.102904.190048.

Petra Sneijder and Hedwig F. M. te Molder. 'Health Should Not Have to be a Problem': Talking Health and Accountability in an Internet Forum on Veganism. *Journal of Health Psychology*, 9(4):599–616, jul 2004. doi: 10.1177/1359105304044046. URL https://doi.org/10.1177%2F1359105304044046.

David Sontag and Dan Roy. Complexity of inference in latent dirichlet allocation. In *Advances in neural information processing systems*, pages 1008–1016, 2011.

Geoffrey M. Stephenson, Bromley H. Kniveton, and Wolfgang Wagner. Social influences on remembering: Intellectual interpersonal and intergroup components. *European Journal of Social Psychology*, 21(6):463–475, nov 1991. doi: 10.1002/ejsp.2420210602. URL https://doi.org/10.1002%2Fejsp.2420210602.

Peng Gang Sun. Analysis of resolution limit in community detection. In *2014 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*. IEEE, aug 2014. doi: 10.1109/fskd.2014.6980932. URL https://doi.org/10.1109%2Ffskd.2014.6980932.

P. Suraj and V. S. Kumari Roshni. Social network analysis in student online discussion forums. In *2015*

IEEE Recent Advances in Intelligent Computational Systems (RAICS). IEEE, dec 2015. doi: 10.1109/raics.2015.7488402. URL https://doi.org/10.1109%2Fraics.2015.7488402.

Eliza Tanner. Chilean Conversations: Internet Forum Participants Debate Augusto Pinochet's Detention. *Journal of Communication*, 51(2):383–403, jun 2001. doi: 10.1111/j.1460-2466.2001.tb02886.x. URL https://doi.org/10.1111%2Fj.1460-2466.2001.tb02886.x.

Alvin Toffler. *The Third Wave*. 1980.

L. M. Verbrugge. The Structure of Adult Friendship Choices. *Social Forces*, 56(2):576–597, dec 1977. doi: 10.1093/sf/56.2.576. URL https://doi.org/10.1093%2Fsf%2F56.2.576.

Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994. doi: 10.1017/cbo9780511815478.002. URL https://doi.org/10.1017%2Fcbo9780511815478.002.

Tom White. *Hadoop: The Definitive Guide, 4th Edition*. O'Reilly Media, Inc., 2011.

Kipling D. Williams, Cassandra L. Govan, Vanessa Croker, Daniel Tynan, Maggie Cruickshank, and Albert Lam. Investigations into differences between social- and cyberostracism. *Group Dynamics: Theory Research, and Practice*, 6(1):65–77, 2002. doi: 10.1037/1089-2699.6.1.65. URL https://doi.org/10.1037%2F1089-2699.6.1.65.

Matei Zaharia, Mosharaf Chowdhury, Michael J Franklin, Scott Shenker, and Ion Stoica. Spark: Cluster computing with working sets. 2009.

Amy X Zhang, Bryan Culbertson, and Praveen Paritosh. Characterizing online discussion using coarse discourse sequences. In *Proceedings of the Eleventh International Conference on Web and Social Media. AAAI Press*, 2017.

Jun Zhang, Mark S. Ackerman, and Lada Adamic. Expertise networks in online communities. In *Proceedings of the 16th international conference on World Wide Web - WWW '07*. ACM Press, 2007. doi: 10.1145/1242572.1242603. URL https://doi.org/10.1145%2F1242572.1242603.

Yulei Zhang, Shuo Zeng, Li Fan, Yan Dang, Catherine A. Larson, and Hsinchun Chen. Dark web forums portal: Searching and analyzing jihadist forums. In *2009 IEEE International Conference on Intelligence and Security Informatics*. IEEE, 2009. doi: 10.1109/isi.2009.5137274. URL https://doi.org/10.1109%2Fisi.2009.5137274.