Centro de Investigación y de Estudios Avanzados
del Instituto Politécnico Nacional

Unidad Zacatenco

Departamento de Computación

Herramienta para el preprocesamiento de tweets con base en búsqueda por tópico

Tesis que presenta

Yareli Licet Andrade Jiménez

para obtener el Grado de

Maestro en Ciencias en Computación

Directora de la Tesis

Dra. Xiaoou Li Zhang

México, Distrito Federal — Marzo, 2015

Resumen

El análisis de redes sociales es un tema que ha cobrado importancia en los últimos años. Esto se debe principalmente al éxito de las redes sociales en línea. Las redes sociales son un medio de comunicación que permite tener disponible gran cantidad de información. Conocer la información que se encuentra implícita en los mensajes compartidos en redes sociales ayuda a averiguar los intereses y opiniones de los usuarios respecto a determinado tema.

Existen diversos tipos de análisis que pueden realizarse en redes sociales. No obstante, en cualquier caso es necesario seguir un proceso de análisis de datos. Este proceso se compone básicamente de cuatro etapas: la obtención de datos, el preprocesamiento de datos, el análisis o minería de datos y la interpretación de los resultados obtenidos. El preprocesamiento es la fase de mayor importancia para la obtención de resultados óptimos.

Como consecuencia al fenómeno de las redes sociales en línea, se han desarrollado diferentes aplicaciones que permiten la exploración y análisis de información generada en éstas. Sin embargo, una de las dificultades que presenta el análisis de redes sociales es la gran cantidad de datos que se requieren analizar. El preprocesamiento es la etapa clave que permite reducir información. A su vez este proceso debe garantizar que la información que se obtiene es suficiente para la etapa de análisis.

El propósito de esta tesis es desarrollar una herramienta para el preprocesamiento de datos de redes sociales, particularmente Twitter. Nosotros planteamos que la base del preprocesamiento sea la búsqueda por temas. Como parte de la solución se desarrolló un módulo para obtener información de esta red social, el cual permite crear corpus con datos actuales y relacionados con un tema específico. Por último, se consideró el algoritmo TF-IDF como ejemplo para analizar los datos obtenidos en la etapa de preprocesamiento y así verificar la funcionalidad de los mismos.

Abstract

The analysis of social networks is a topic that has gained importance in recent years. This is mainly due to the success of online social networks. Social networks are a media that allows lots of information available. Knowing the information that is implicit in shared messages on social networks helps to determine the interests and opinions users regarding certain topic.

There are different types of analysis that can be performed on social networks. However, in any case it is necessary to follow a process of data analysis. This process basically consists of four stages: data collection, data preprocessing, analysis or data mining and interpretation of results. The preprocessing phase is the most important for obtaining optimal results.

As a result the phenomenon of online social networks, different applications that allow the exploration and analysis of information generated in them have been developed. However, one of the difficulties of social network analysis is the amount of data that is necessary to analyze. The preprocessing is the key stage to reduce information. In turn, this process should ensure that the information obtained is sufficient for the analysis stage.

The purpose of this thesis is to develop a tool for data preprocessing of social networks, particularly Twitter. We propounded the subject search as a basis of preprocessing stage. As part of the solution, a module for collect Twitter information has been developed, which allows create corpus with current data and related with an specific topic. Finally, TF-IDF algorithm was considered as an example to analyze obtained data in the preprocessing stage, in order to verify the functionality thereof.

Agradecimientos

Al Centro de Investigación y Estudios Avanzados del I.P.N. (CINVESTAV) por permitirme continuar mis estudios en una institución de alto prestigio.

Agradezco al Consejo Nacional de Ciencia y Tecnología (CONACyT) por el apoyo económico que me brindó como becario durante mis estudios de maestría.

A mi asesora, la Dra. Xiaoou Li, por la orientación y el tiempo dedicado durante el desarrollo de esta tesis.

A mis sinodales, Amilcar Meneses Viveros y Maricela Claudia Bravo Contreras, por su valioso tiempo y las aportaciones a mi trabajo.

A mis compañeros de maestría por su amistad y por compartir esta aventura.

Finalmente dedico este proyecto a mi familia. A mis padres, quienes siempre me han dado su confianza y cariño. A mis hermanos que estuvieron conmigo a lo largo de esta etapa. A David, mi compañero de vida, por su amor incondicional. Gracias totales.

Índice general

Índice de figuras

Índice de tablas

Lista de algoritmos

Capítulo 1

Introducción

1.1. Antecedentes

Gracias a la Web 2.0 en los últimos años se ha tenido un incremento drástico de información. Una de las causas de este fenómeno es la creciente utilización de redes sociales. En ellas se generan día con día inmensidad de datos, convirtiendo este medio en una fuente de información que es consultada a diario por una gran cantidad de usuarios.

Actualmente hay varias redes sociales en línea disponibles con más de 100 millones de usuarios registrados, entre las cuales se encuentran: YouTube, Facebook, Twitter, LinkedIn y MySpace. Este fenómeno es debido al incremento espectacular en el número y la popularidad de las redes sociales en línea, en donde la gente puede descubrir y compartir contenidos. El crecimiento exponencial de las comunidades en línea ha estimulado el interés de varios investigadores en el estudio de redes sociales, tal como el descubrimiento de los intereses de los usuarios. Conocer los intereses de un usuario es de ayuda en diversas tareas, por ejemplo en el desarrollo de publicidad en línea. Otros temas de interés relacionados con el análisis de redes sociales son la detección de comunidades y el desarrollo de tecnologías de recomendación para los usuarios.

Los datos compartidos por los usuarios de redes sociales en línea contienen información que no se puede apreciar a simple vista. Por ejemplo, opiniones acerca de un tema en particular o algún producto, y detección de eventos en determinada zona. Es por ello que existe interés, por parte de empresas y ciertos usuarios, por conocer la información que se encuentra de forma implícita en el contenido generado en diversas redes sociales.

Considerando que las redes sociales en línea son un medio en el que se intercambian infinidad de datos entre usuarios, así como el impacto que hasta la fecha éstas tienen en la sociedad, es claro que el análisis de los datos generados en este medio de comunicación es necesario para lograr explotar la información oculta en ellos.

1.2. Motivación

Dado que en la actualidad las redes sociales son protagonistas en la Web 2.0, hay un sin fin de investigaciones que se realizan para analizar los datos que son generados en este medio de comunicación.

Tomando en cuenta que la cantidad de datos que se generan diariamente es inmensa, es muy complicado poder realizar estudios con datos generados recientemente. Una de las opciones que se tiene para poder realizar, por ejemplo, análisis de opinión, es utilizar recopilaciones de datos existentes para investigación científica; sin embargo, estos conjuntos de datos no son de utilidad para conocer información referente a temas de interés actuales.

En Twitter, particularmente, se envían 500 millones de tweets por día al rededor del mundo, por lo que la cantidad de datos que pueden ser analizados en esta red social es gigantesca. Una característica sobresaliente de Twitter es que los usuarios comparten información acerca de eventos recientes, sobre todo en respuesta a noticias de última, como desastres naturales, información de celebridades o protestas masivas. Es por esta razón que es necesario contar con una herramienta que permita obtener datos en cualquier momento y así efectuar análisis de datos que den resultados inmediatos sobre lo que ocurre en determinada red social.

Considerando las etapas que forman parte del proceso de análisis de datos es relevante destacar que la etapa de preprocesamiento es la más importante. Por tal motivo, y considerando los estudios que hasta la fecha se han realizado, proponemos desarrollar la fase de preprocesamiento para datos de Twitter considerando tópicos específicos.

Aunque actualmente ya se cuenta con otros enfoques de análisis de redes sociales en línea, el alcance de éstos se limita por la información que puede encontrarse en el conjunto de datos que consideran. En este trabajo se planteamos un enfoque por tópicos. Este enfoque permite generalizar la búsqueda, logrando considerar un rango mayor en la etapa de preprocesamiento de datos.

1.3. Planteamiento del problema

Sitios como Blogs, Facebook y Twitter son utilizados para almacenar y compartir contenidos entre usuarios. Dado que en la mayoría de ellos la información es pública, destaca el hecho de que ésta puede ser utilizada para realizar diferentes estudios. Existen distintos tópicos dentro del análisis de redes sociales. Sin embargo, la mayoría de ellos se caracterizan por requerir de una serie de etapas para obtener un resultado. En general el proceso de análisis consta de 4 fases: la obtención de datos, el preprocesamiento de datos, el análisis o minería de datos y la interpretación de los resultados obtenidos (ver Figura 1.1).

La creación de un corpus consiste en la obtención de una muestra de datos sobre los cuales se va a realizar el análisis. Es necesario saber qué datos se necesitan, dónde se pueden

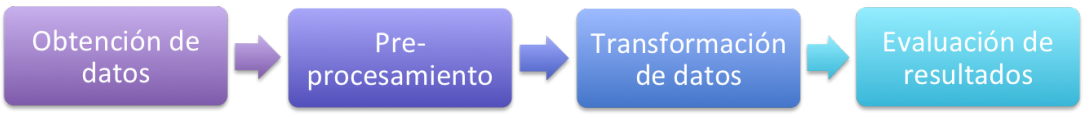

Figura 1.1: Etapas del proceso de análisis de datos.

encontrar y cómo se pueden obtener.

Una vez que se dispone de datos es necesario realizar un preprocesamiento para así eliminar información secundaria para los objetivos de análisis propuestos. En este paso se pueden utilizar distintos métodos de transformación para reducir el número de elementos a ser considerados en la siguiente etapa.

Ya que se tienen los datos preprocesados, el objetivo es decidir la tarea de minería de datos que se va a aplicar en ellos, por ejemplo clasificación o agrupamiento. La elección del algoritmo a utilizar define además el tipo de entradas más apropiadas para lograr el objetivo deseado.

Finalmente una vez que se tienen los resultados es necesario dar una interpretación de los patrones encontrados con el fin de obtener el conocimiento pretendido. En ocasiones, dependiendo de la salida del proceso, es necesario regresar a una de las etapas previas.

Actualmente, las aplicaciones que se enfocan en el análisis de redes sociales en línea utilizan sus propios mecanismos para capturar, almacenar, preprocesar y analizar datos. Particularmente en Twitter los estudios realizados cuentan con enfoques muy específicos (por persona, por ubicación, por tweets recientes), los cuales no permiten generalizar el preprocesamiento de datos.

Nosotros proponemos abordar la fase de preprocesamiento considerando búsqueda de tweets por tema. Esta característica permite contar con una perspectiva más general, haciendo énfasis en la necesidad de nuevos enfoques de análisis de datos que permitan solventar las limitaciones de las soluciones existentes.

En esta tesis se presenta la implementación de la etapa de preprocesamiento de análisis de datos. Particularmente se hace énfasis en ésta etapa planteando considerar datos referentes a un tema en particular. La finalidad de este enfoque es contar con datos a los que sea posible aplicar diferentes algoritmos o técnicas de minería de datos y obtener mejores resultados.

1.4. Justificación

La obtención de información implícita en los contenidos compartidos en redes sociales es necesaria tanto para conocer la opinión de las personas, en este medio, respecto a situaciones actuales como en la toma de decisiones de algunas empresas. La utilidad de los resultados dados por aplicaciones existentes puede mejorar si se cuenta con una herramienta que considere las características de las redes sociales actuales. Particularmente la etapa de preprocesamiento debe considerar las singularidades del conjunto de datos a analizar.

El esfuerzo que requiere cada fase del proceso de análisis es distinto. Gran parte del esfuerzo de este proceso recae sobre en la fase de preprocesamiento, la cual es crucial para tener éxito en la obtención de información implícita, pues es esta etapa en la que se preparan los datos para poder ser utilizados por determinado algoritmo.

Los usuarios de redes sociales en línea de hoy en día constantemente publican información en sus respectivos perfiles, que consiste en datos como la ubicación geográfica, intereses, pasatiempos, opiniones, etc. La información que se encuentra en las redes sociales sirve como base para el intercambio de contenidos y para identificar temas que son de interés para los usuarios. Sin embargo, en la práctica no es viable leer cada una de las entradas de diferentes usuarios para poder identificar similitudes existentes entre ellas.

A partir de la importancia que ha cobrado el estudio de redes sociales ha surgido un gran número de investigaciones enfocadas en este tema. Cada propuesta de solución utiliza mecanismos propios para cada etapa del proceso de análisis. El objetivo de crear aplicaciones que permitan la exploración y análisis de redes sociales es facilitar a los usuarios la toma de decisiones. Una de las problemáticas existentes es que las etapas de obtención de datos y de preprocesamiento varían de acuerdo al tipo de análisis que se va a realizar. Por tal motivo, contar con una solución que permita llevar a cabo la etapa de preprocesamiento con base en temas va a permitir obtener mejores resultados en la fase de análisis.

Este trabajo de tesis podrá ser tomado como base para desarrollar una herramienta que integre todo el proceso de análisis de datos de redes sociales. La implementación del módulo de preprocesamiento tiene como propósito facilitar al usuario la exploración de datos para su utilización en distintos tipos de análisis.

1.5. Contribuciones

El objetivo general de esta tesis es implementar una herramienta de preprocesamiento de datos para análisis de redes sociales tomando como referencia un tema específico.

Para alcanzar el objetivo planteado se llevaron a cabo las siguientes tareas:

- Se analizaron las características de las redes sociales en línea existentes. Como resultado, se llegó a la conclusión de que Twitter es la red social que mejor se adecua para contemplar como caso de estudio.

- Se desarrolló un módulo que se encarga de llevar a cabo la etapa de proprocesamiento de datos. Este modulo permite reducir la información contenida en los tweets, eliminando así texto que no es significativo para la etapa de análisis.

- Se realizó un programa que permita realizar la obtención de datos considerando un tema específico de búsqueda.

- Se implementó el algoritmo TF-IDF que considera como entrada los datos obtenidos en el módulo de preprocesamiento para así probar la funcionalidad de los mismos.

- Se diseñó una herramienta que permite incluir de forma integral la obtención de datos de Twitter mediante su API, el módulo de preprocesamiento y la aplicación del algoritmo TD-IDF.

Finalmente, las aportaciones que se obtuvieron como resultado de esta investigación son:

- Un módulo de preprocesamiento de datos de Twitter con base en búsqueda por tópico. La entrada para la etapa de preprocesamiento es un conjunto de tweets previamente obtenido con la herramienta.

- Una herramienta cuya interfaz permite al usuario buscar y almacenar datos de Twitter. Posteriormente es posible minimizar el conjunto de datos obtenido mediante las tareas del módulo de preprocesamiento que se implementó.

1.6. Estructura del documento

Este documento está estructurado en seis capítulos. Después de haber planteado el problema que se pretende resolver junto con los objetivos del proyecto, el resto de esta tesis se encuentra organizada de la siguiente manera:

En el capítulo 2 se mencionan las características de las redes sociales en línea, así como los principales elementos relacionados con su estudio. Se describen las tareas de recuperación de información de redes sociales, así como el tipo de comunidades existentes en Twitter y sus características. También se describe el proceso de minería de textos que será de utilidad para el desarrollo de la herramienta propuesta.

En el capítulo 3 se detalla todo lo relacionado con Twitter, que es la red social en línea que consideramos como caso de estudio para la implementación de la herramienta

propuesta. Se proporcionan los conceptos básicos, sus características y parte de los trabajos que actualmente existen en relación con esta red social.

En el capítulo 4 se describe la contribución de este trabajo de tesis, la cual consta de la colección de datos recientes a través de la API de Twitter, la creación de una base de datos para el almacenamiento de la información obtenida, la selección de métodos a implementar en la etapa de preprocesamiento de tweets y el algoritmo TF-IDF que se consideró para realizar pruebas con los datos obtenidos. También se detalla el desarrollo de la interfaz de usuario que servirá para integrar las dos primeras etapas del análisis de datos de Twitter.

En el capítulo 5 se presentan los resultados obtenidos en los experimentos, las características de los datos de entrada y la salida obtenida posterior a la fase de preprocesamiento de tweets. Se describen particularmente dos casos que fueron implementados con la finalidad de observar el desempeño de la herramienta desarrollada.

Finalmente en el capítulo 6 se dan las conclusiones de este trabajo. Así mismo se indican las posibles vertientes de trabajo futuro que podrían surgir de esta tesis.

Capítulo 2

Redes sociales en línea

Este capítulo tiene como objetivo describir los temas que se relacionan directamente con esta investigación. En primera instancia se abarca el tema de la Web 2.0, se mencionan sus características y los cambios que han surgido con el paso del tiempo. También se menciona el tema de redes sociales, su importancia y el tipo de análisis que hay. Posteriormente se hace una breve discusión en el área de minería web, particularmente minería de textos y su aplicación en diferentes tipos de problemas. Por último se incluyen las tareas que implica el procesamiento de lenguaje natural, debido a que es un tema indispensable para el análisis de lenguaje escrito.

2.1. La Web 2.0

En años recientes la manera en la que la Web es percibida por los usuarios ha cambiado drásticamente. Se pasó de ser usuarios pasivos que sólo reciben información a participantes activos donde sus contribuciones son importantes, creando así una plataforma de colaboración. Este cambio se debe principalmente a la aparición de nuevos tipos de sitios web que se han enfocado más a la socialización y la colaboración entre los usuarios. Entre los sitios web que han cambiado la percepción de la web podemos mencionar como los mas representativos Facebook y Twitter. Estos dos sitios están directamente relacionados con la Web 2.0 debido a que ambos dependen en gran parte de la información generada por los usuarios.

Cuando se habla de Web 2.0 esta implícito el término "redes sociales", pues esta web apareció a partir de esta forma de interactuar entre los usuarios de la red. Ésta ha facilitado la publicación de contenido sin tener ningún conocimiento técnico. Se puede publicar una fotografía con simplemente dar un clic, o hacer un comentario con sólo redactarlo y hacer clic en publicar. Sin duda alguna esto ha atraído a personas que antes ni siquiera pensaban en acercarse a una computadora. También gracias a este crecimiento muchas empresas han visto la obtención de información de la web como una posible ventaja para ellos.

2.1.1. Concepto de Web 2.0

El concepto Web 2.0 es mencionado en [2] por Darcy DiNucc en 1999, mas tarde el término se popularizó con Tim O'Reilly [1] cuando apareció en una lluvia de ideas entre O'Reilly y MediaLive Internacional. El termino Web 2.0 como muchos otros conceptos importantes, no tiene límites fuertemente establecidos, sino más bien un centro gravitacional. La Web 2.0 se puede visualizar como un conjunto de principios y prácticas que entrelazan un sistema solar de sitios que demuestran algunos o todos esos principios, dependiendo de la distancia que exista entre dichos sitios y el núcleo.

Principios de la Web 2.0

La publicación de anuncios fue el primer servicio web en ser ampliamente desplegado. Cada anuncio se puede considerar como la cooperación entre dos sitios web, entregando de esta forma contenido a los lectores en otras computadoras.
Algunos de los principios y recursos que le dieron el éxito que actualmente tiene la Web 2.0 se listan a continuación:

- El hiperlinkeado es la técnica sobre la cual se fundamente la web. Cada vez que un usuario agrega contenido, éste es ligado a la estructura de un sitio y descubierto por otros usuarios. De esta forma, quienes comparten o hacen uso de dicho contenido, crean asociaciones, tal y como sucede con nuestro cerebro, en donde se forman conexiones orgánicas que alimentan y potencian las actividades.

- Muchos de los grandes buscadores de hoy en día comenzaron como simples catálogos o directorios de links, cuyo contenido podía ser alimentado por los usuarios. En realidad, parte del éxito que carga la mayoría de los sitios web de gran relevancia como lo son Google, Yahoo, eBay, Wikipedia y Amazon, se debe a las técnicas novedosas de estructuración de sus sitios web y la incorporación de los usuarios para enriquecer el contenido que se les proporciona.

- Muchas de las aplicaciones más significativas en la web, hacen uso de bases de datos complejas y especializadas, al grado de llegar a ser motivos de competencia clave para empresas de prestigio.

Diferencias entre Web 2.0 y Web 2.1

Como se mencionó en la sección anterior no existe un consenso general acerca de lo que es Web 2.0, pero existen diferencias que se han destacado para poder afirmar que ésta ha sido la forma en la que la Web evolucionó. La principal diferencia radica en el hecho de que las aplicaciones de Web 1.0 son estáticas y son enfocadas a hacer del usuario sólo un receptor,

Web 1.0	**Web 2.0**
DoubleClick	Google AdSense
Ofoto	Flickr
Akamai	BitTorrent
mp3.com	Napster
Britannica Online	Wikipedia
personal websites	blogging
evite	upcoming.org and EVDB
domain name speculation	search engine optimization
page views	cost per click
screen scraping	web services
publishing	participation
content management systems	wikis
directories (taxonomy)	tagging ("folksonomy")
stickiness	syndication

Tabla 2.1: Web 1.0 y 2.0[1]

mientras que la Web 2.0 se enfoca en hacer al usuario una parte clave, pues es él quien genera la información que otros usuarios van a consultar logrando una retroalimentación.

En la tabla 2.1 se presenta una comparativa más de lo que es Web 1.0 y 2.0, esta comparación es más una clasificación de las aplicaciones que pertenecen a cada una de las generaciones de la Web [1] .

2.1.2. Importancia de la Web 2.0

Ya se mencionaron las ventajas de los sitios que pertenecen a la Web 2.0 pero además de los beneficios sociales tales como conocer a personas con los mismos gustos o intereses, comunicarse con amigos o incluso publicar fotos; el principal interés de la información que generan los usuarios recae en las grandes empresas. Estas empresas diseñan productos o incluso campañas de marketing basadas en lo que los consumidores quieren. Así es como los anuncios que aparecen en el navegador son resultado de un continuo seguimiento acerca de las actividades de los usuarios en la Web.

La nueva forma en la que los usuarios interactúan también ha impactado en la educación pues con la creación de foros, blogs o wikis y el interés de las personas sobre ellos, se ha vuelto un recurso común en lugares donde acceder a los recursos necesarios es posible. También ha permitido que cualquier persona tenga la posibilidad de generar contenido. Existen blogs o páginas de noticias con más visitas que las páginas oficiales de los periódicos.

El concepto Web 2.0 abarca una serie de aplicaciones que proporcionan servicios interactivos en red, proporcionando al usuario el control de sus datos. Entre estas aplicaciones podemos mencionar las redes sociales, blogs, wikis, y la sindicación de contenidos. En la Tabla 2.2 se da una breve descripción de estos servicios.

Servicio	Descripción
Wikis	Son espacios editados y mantenidos por los propios usuarios con la finalidad de compartir conocimientos. La creación y edición de contenidos se basa en la interacción de los propios usuarios, de tal forma que múltiples autores puedan crear, modificar o eliminar contenidos identificando a cada usuario que realiza un cambio. Su principal ventaja es la flexibilidad y la facilidad de elaboración. Es ideal para que pequeños grupos de investigación intercambien ideas.
Foros	Son aplicaciones Web desarrolladas para dar soporte a discusiones o debates entre usuarios referentes a temáticas concretas.
Blogs	Su principal característica es su actualización frecuente. Los autores ingresan cronológicamente entradas breves (artículos, opiniones, sugerencias, enlaces, noticias). Además, otros usuarios pueden generar comentarios y opiniones acerca de las entradas publicadas.
Redes sociales	Son servicios Web de socialización entre personas que comparten entre sí relaciones (de parentesco, personales) o gustos semejantes, o que desean explorar los intereses de otros.
Sindicación de contenidos	El RSS (Rich Site Summary) es parte de la familia de los formatos XML. El objetivo es la distribución masiva de información (noticias) contenida en diferentes sitios. La mayoría de los autores definen un archivo RSS como una descripción estructurada (especie de resumen) de uno o varios sitios Web.

Tabla 2.2: Servicios de la Web 2.0.

Es importante considerar que el que una persona pueda generar su propia información (publicar fotos, videos, opiniones) es un conflicto, pues una vez que esta información es "subida" a un sitio, ésta se convierte en algo público sobre lo cual se pierden lo derechos.

2.2. Redes sociales

Una red social es un conjunto bien definido de actores (individuos, grupos, organizaciones, etc.) que están vinculados unos a otros a través de una o un conjunto de relaciones sociales (amistad, familia, intereses comunes, preferencias, conocimiento, actividades).

El área de redes sociales tiene como objetivo el estudio de las entidades sociales (personas en una organización, llamados actores) y sus interacciones y relaciones [3]. Las interacciones y relaciones se pueden representar con una red o gráfico, donde cada vértice (o nodo) representa un actor y cada enlace representa una relación. Desde la red, es posible estudiar las propiedades de su estructura, así como el rol, posición y prestigio de cada actor social. También se pueden encontrar varios tipos de sub grafos, por ejemplo, las comunidades formadas por grupos de actores. El análisis de redes sociales es útil para la Web, ya que la Web es esencialmente una sociedad virtual, y por lo tanto una red social virtual, donde cada página puede ser considerado como un actor social y cada hipervínculo como una relación. Muchos de los resultados de las redes sociales pueden ser adaptados y ampliados para su uso en el contexto Web. Las ideas de análisis de redes sociales son, decisivos para el éxito de los motores de búsqueda web.

Los principales tipos de análisis para una red social son de centralidad y de prestigio. Estos anásis están muy relacionados con el análisis de hipervínculos y la búsqueda en la Web. Tanto la centralidad como el prestigio son medidas de grado de influencia de un actor en una red social.

Centralidad

Los actores importantes o destacados son aquellos que están vinculados o involucrados con otros actores en gran medida. En el contexto de una organización, una persona con gran cantidad de contactos (links) o comunicación con muchas otras personas de la organización se considera más importante que una persona con relativamente menos contactos. Los enlaces también pueden ser llamados lazos. Un actor central es aquel que forma parte de varios lazos. En la Figura 2.1 se muestra un ejemplo sencillo utilizando un grafo no dirigido. Cada nodo de la red social es un actor y cada enlace indica es sus dos extremos cuáles son los actores que se comunican entre sí. Intuitivamente, vemos que el actor 'a' es el actor más

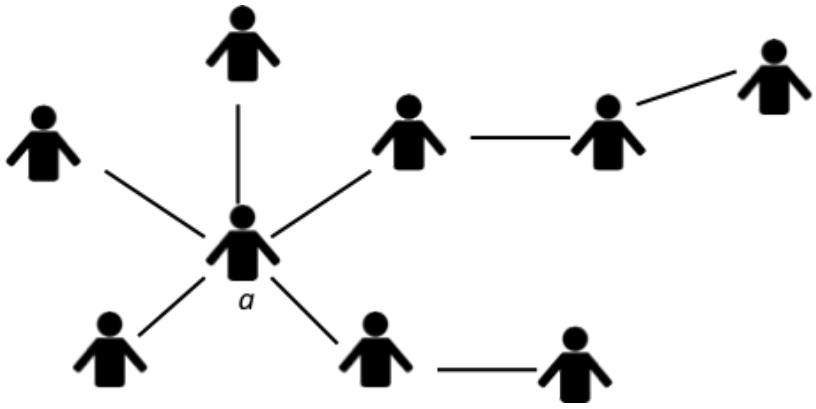

Figura 2.1: Ejemplo de centralidad en una red social.

central porque él puede comunicarse con la mayoría de los otros actores.

Hay diferentes tipos de vínculos o relaciones entre los actores. Por consiguiente hay varios tipos de centralidad, los cuales a su vez son clasificados en grafos dirigidos y no dirigidos.

Prestigio

El prestigio es una medida más concreta de la importancia que tiene un actor de carácter central. Existe un distinción entre los lazos enviados (out-links) y los lazos recibidos (in-links). Un actor prestigioso está definido como aquel que es objeto que tiene gran cantidad de lazos como destinatario. En otras palabras, para calcular el prestigio de un actor, sólo nos fijamos en los lazos (links) dirigidos o dirigidos al actor (in-links). Por lo tanto, el prestigio no se puede calcular a menos que la relación sea direccional o el gráfico sea dirigido.

La principal diferencia entre los conceptos de centralidad y de prestigio es que la centralidad se centra en los out-links, mientras que el prestigio se centra en los in-links. Existen tres medidas de prestigio definidas. La tercer medida de prestigio (es decir, el rango de prestigio) forma la base de la mayoría de los algoritmos de análisis de las páginas web, entre los cuales se encuentran PageRank y HITS.

2.2.1. Análisis de redes sociales en línea

El auge de las redes sociales ha revolucionado la forma de comunicación de las personas, la cantidad de datos que se comparten a diario (imágenes, texto, conversaciones, vídeos) es exorbitante, y las posibilidades a la hora de analizar esos datos para obtener información son igualmente inmensas.

Según estadísticas mostradas en eMarketer [4], aproximadamente el 67.7 % de los millones de usuarios de Internet a nivel mundial usan redes sociales como Facebook, Twitter y LinkedIn, pero existen miles de redes sociales diferentes, e incluso hay instituciones que

cuentan con una red propia. Es por ello que con el paso del tiempo se ha tomado cada vez mayor interés en el estudio y análisis de la información contenida en estos sitios.

Existen diversas áreas enfocadas en el estudio de las redes sociales, dependiendo del enfoque se tenga. Entre las categorías principales se encuentran: minería web, minería de opinión, clasificación y agrupamiento, identificación de entidades, detección temática, análisis de comunidades y minería de grafos. En la siguiente sección se va a abarcar el tema de minería web que se encuentra ampliamente relacionado con esta tesis.

2.3. Descubrimiento de conocimiento en redes sociales

La información en sí misma es útil para las organizaciones cuando se contextualiza, la información asociada a un contexto y a una experiencia se convierte en conocimiento siendo así un recurso intangible que aporta verdadero valor a determinada organización. La generación de conocimiento a partir de información que se encuentra en la web de forma masiva no es una tarea sencilla, ante estos enormes volúmenes de información no estructurada que almacenan algunos repositorios se necesitan sistemas automatizados que permitan extraer conocimiento a partir de ella. En particular las técnicas de minería de datos permiten explorar y extraer conocimiento de colecciones de datos.

Para poder obtener conocimiento de una red social existen tareas básicas que deben realizarse, las cuales se describen a continuación.

- **Recuperación de información:** consiste en extraer de manera automática datos que puedan resultar interesantes para una organización a partir de una consulta realizada.
- **Categorización:** se refiere a asignar a cada elemento obtenido una o varias categorías temáticas a partir de un conjunto de categorías preestablecido.
- **Agrupación:** establece la generación automática de comunidades (grupos de elementos relacionados), por ejemplo, elementos que traten un mismo tema o asunto. A diferencia de lo que ocurre en la categorización, en el proceso de agrupación no existe un conjunto de categorías preestablecido, sino que de acuerdo al algoritmo utilizado se deben generar automáticamente dichas categorías, contribuyendo de esta forma a generar nuevo conocimiento.

2.3.1. Minería web

La *World Wide Web* (Web) es un medio popular e interactivo para difundir información en la actualidad. La web es enorme, diversa y dinámica, por lo tanto aumenta la escalabilidad, datos multimedia y las cuestiones temporales, respectivamente. Debido a esas situaciones,

actualmente hay sobrecarga de información [5]. Cuando un usario busca información en la web se presentan los siguientes problemas [6]:

- Encontrar información relevante.
- Crear nuevo conocimiento no disponible en la web.
- Personalización de la información.
- Aprender acerca de los consumidores o usuarios individuales.

Las técnicas de minería web no solamente son usadas para resolver los problemas de sobrecarga de información. Existen técnicas que se pueden aplicar en diferentes áreas de investigación como bases de datos, recuperación de información y procesamiento de lenguaje natural, principalmente.

La minería web es el uso de técnicas de minería de datos para descubrir y extraer información de documentos y servicios web automáticamente [7]. La extracción de información tiene el objetivo de transformar una colección de documentos, por lo general con la ayuda de un sistema de recuperación de información, en información facilmente digerible y analizable [8]. La extración de información trabaja de la mano con recuperación de información, ya que esta última se encarga de obtener la informacón más importante de un documento, mientras que la primera extrae los documentos importantes de la Web. La calidad de la información y su preprocesamiento es muy importante para este tipo de actividades, por lo tanto se requiere el uso de técnicas confiables para la obtención de información. El conjunto de técnicas de minería de datos que se han desarrollado recientemente para datos relacionales incluyenm relacionales probabilísticos [9], programas de lógica bayesiana [10], clasificadores bayesianos de primer orden [11] y árboles relacionales de probabilidad [12]. En cada uno de estos casos, la estructura y los parámetros de un modelo estadístico se pueden aprender directamente de los datos, lo que facilita el trabajo de los analistas de datos, y mejora en gran medida la fidelidad del modelo resultante. Las técnicas más antiguas incluyen la programación lógica inductiva ([13], [14]) y análisis de redes sociales ([15]).

La noción de red social y los métodos para el análisis de redes sociales han atraido considerablemente el interés y curiosidad de la comunidad científica para analizar la cuestión social y el comportamiento de las personas en décadas recientes. Este interés se debe principalmente al análisis de redes sociales a través de las relaciones entre entidades sociales, y de los patrones e implicaciones de esas relaciones. Desde el punto de vista del análisis de redes sociales el ambiente social puede ser expresado como un patrón, identificando así las relación entre las unidades que interactúan. La presencia de estos patrones en una relación forman una estructura [15].

De acuerdo a [16] la minería de datos web se puede clasificar en tres grupos distintos no disjuntos, dependiendo del tipo de información que se quiera extraer, o de los objetivos:

- Minería del Contenido de la Web [Web Content Mining]: Consiste en extraer información del contenido de los documentos en la web. Se puede clasificar a su vez en:
  - Text Mining: Si los documentos son textuales (planos).
  - Hypertext Mining: Si los documentos contienen enlaces a sí mismos o a otros documentos.
  - Markup Mining: Si los documentos son semi estructurados (con marcas).
  - Multimedia Mining: Para imágenes, audio, vídeo.
- Minería de la Estructura de la Web [Web Structure Mining]: Intenta descubrir un modelo a partir de la tipología de enlaces de la red. Este modelo puede ser útil para clasificar o agrupar documentos.
- Minería del Uso de la Web [Web Usage Mining]: Se intenta extraer información (hábitos, preferencias, etc. de los usuarios o contenidos y relevancia de documentos) a partir de las sesiones y comportamiento de los usuarios navegantes.

2.3.2. Minería de textos

La minería de textos (TM - Text Mining) es un proceso que extrae información útil de texto no estructurado, como los correos electrónicos, periódicos y otros documentos [17]. Su objetivo principal es analizar un conjunto de documentos para recuperar información que no siempre se encuentra explícita en los textos. Considerando que la mayor parte de la información que se genera a diario se almacena como texto, la minería de textos se ha convertido en un área de investigación importante.

La minería de textos surge ante el problema cada vez más apremiante de extraer información automáticamente a partir de masas de textos. Se trata así de extraer información de datos no estructurados: texto plano.

Existen varias aproximaciones a la representación de la información no estructurada [18]:

- "Bag of Words" (bolsa de palabras): Cada palabra constituye una posición de un vector y el valor corresponde con el número de veces que ha aparecido.
- N-gramas o frases: Permite tener en cuenta el orden de las palabras. Trata mejor frases negativas "... excepto ...", "... pero no ...", que tomarían en otro caso las palabras que le siguen como relevantes.

- Representación relacional (primer orden): Permite detectar patrones más complejos (si la palabra X está a la izquierda de la palabra Y en la misma frase).

- Categorías de conceptos: Casi todos se enfrentan con el "vocabulary problem" [19]. Tienen problemas con la sinonimia, la polisemia, los lemas, etc.

Un ejemplo de aplicación basada en minería de textos es la generación automática de índices en documentos. Otras más complicadas consisten en escanear completamente un texto y mostrar un mapa en el que las partes más relacionadas, o los documentos más relacionados, se coloquen cerca unos de otros. En este caso se tratan de analizar las palabras en el contexto en que se encuentren. Aunque aún no se ha avanzado mucho en el área de minería de textos, ya hay productos comerciales que emplean esta tecnología con diferentes propósitos.

La mayoría de los datos se originan en forma digital. Por ejemplo, si se quiere comprar un producto, es un evento que ahora se puede desarrollar de forma electrónica. Dado que tantas transacciones en papel están ahora en formato digital, un gran cantidad de datos están disponibles para su posterior análisis. El concepto de minería de datos consiste en encontrar patrones en un conjunto de datos, y es una respuesta al almacenamiento de grandes volúmenes de datos. La minería de datos ya no es una tecnología emergente en espera de nuevos desarrollos. Aunque su aplicación está lejos de ser universal, las técnicas de minería de datos están muy desarrolladas y para algunas formas de análisis están entrando en una fase de madurez. Desafortunadamente, los métodos de extracción de datos esperan un formato altamente estructurado para los datos. Por consiguiente o se tienen que transformar los datos originales, o los datos deben suministrarse en un formato muy estructurado.

Una de las principales diferencias entre minería de datos y minería de texto es que se consideran dos formatos: números vs. texto. Eso no quiere decir que se trata de dos conceptos distintos. Aunque su composición es muy diferente, muchos de los métodos de aprendizaje para cada área son similares. Esto es debido a que el texto se procesa y se transforma en una representación numérica.

Palabras vacías y lematización

La "bolsa de palabras", una de las posibles representaciones de información no estructurada, tiene como característica principal la generación de una gran cantidad de palabras a partir de un conjunto de documentos dado. Sin embargo no todas las palabras son de utilidad para su análisis, por lo que surge la problemática de reducir el número de elementos.

Para solucionar este inconveniente, existe un método muy utilizado. Este método consiste en crear una lista de palabras comunes que son irrelevantes, conocidas como palabras

vacías (stopwords) [20]. Una vez que se tiene el conjunto de documentos, el primer paso es eliminar todas las apariciones de las palabras contenidas en la lista para posteriormente crear la representación de "bolsa de palabras". Hasta la fecha no existe una lista definitiva de las palabras vacías que deben utilizarse, ya que depende de varios factores, como el tipo de análisis que se requiere hacer, o el idioma de los documentos.

Otra forma muy importante para reducir el número de palabras en la representación es utilizar derivados [20]. Esto se basa en la observación de que las palabras en los documentos con frecuencia tienen muchas variantes morfológicas. El objetivo de esta tarea es reconocer conjuntos de palabras tales como "computing" y "computation", "applied" y "applying", "applies" y "apply". Estas palabras pueden ser consideradas equivalentes debido a que cada par tienen la misma raíz lingüística. Actualmente hay algoritmos que han sido desarrollados para reducir las palabras a su forma mínima, sin embargo no hay un algoritmo estándar para llevar a cabo esta tarea. El uso de palabras raíz es una forma efectiva de reducir el número de palabras en una representación de "bolsa de palabras".

2.3.3. Áreas de aplicación de minería de textos

En esta sección se describen algunos de los métodos y aplicaciones de minería de textos más estudiados y aplicados. La tarea principal consiste en organizar los datos de tal manera que se puedan etiquetar, este proceso se conoce como agrupación. La similitud entre documentos es un característica esencial en la organización de los documentos sin etiqueta. Así mismo la medición de similitud entre distintos documentos es fundamental para la mayoría de las formas de análisis de documentos, especialmente la recuperación de información.

Las aplicaciones mencionadas no incluyen análisis lingüístico, aunque eso no significa que considerar el análisis de la semántica no pueda tener mejoras en el desempeño. Actualmente el tipo de soluciones que tiene predominio son los métodos estadísticos, éxito que se ha tenido por la capacidad cada vez mayor de los recursos informáticos. Es necesario considerar que se requiere del análisis mediante métodos de minería de textos debido a la cantidad de datos digitales con los que se cuenta.

Existen varias técnicas de minería de texto para resolver el problema de la extracción automática de información [21]. A continuación se describe la función que cada una de estas técnicas desempeña en la minería de textos.

- **Clasificación de documentos**

  La categorización de texto es muy utilizada en la clasificación de documentos. Una vez que los datos se transforman en el formato de hoja de cálculo numérico, los métodos de extracción de datos estándar son aplicables. En la Figura 2.2 se ilustra la clasificación de documentos. Los documentos se organizan en carpetas, una carpeta para cada

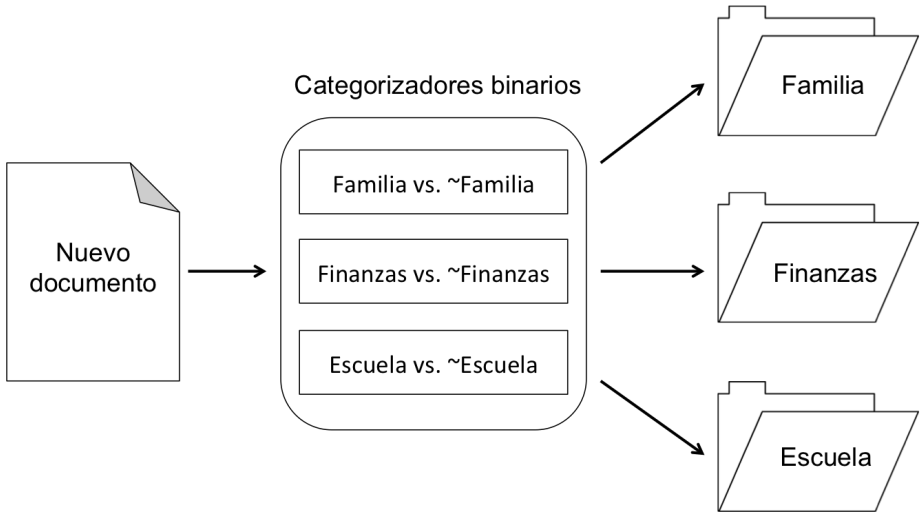

Figura 2.2: Categorización de textos.

tema. Un nuevo documento se presenta, y el objetivo es colocar este documento en las carpetas correspondientes. La aplicación casi siempre es binaria porque un documento por lo general puede aparecer en varias carpetas. Originalmente, este tipo de problema se considera una forma de indexación, similar al índice de un libro. A medida que la cantidad de documentos disponibles en línea es mayor, la aplicabilidad de esta tarea incrementa.

- **Recuperación de información**

  Recuperación de la información es el tema más comúnmente asociado con documentos en línea. La tarea general de recuperación de información se ilustra en la Figura 2.3. Se obtiene una colección de documentos, se dan características de los documentos que se quieren recuperar de la colección, y luego los documentos que coinciden con las especificaciones dadas se presentan como respuestas a la búsqueda. Las características que se utilizan para la recuperación son palabras que ayudan a identificar los documentos almacenados. En un ejemplo típico de la invocación de un motor de búsqueda, se presentan algunas palabras, y estas palabras se hacen coincidir con los documentos almacenados. Los documentos con mayor número de coincidencias se presentan como las respuestas. El proceso puede ser generalizado a un emparejamiento de documentos, donde en lugar de unas pocas palabras, un documento completo se presenta como un conjunto de características. El documento de entrada se empareja a todos los documentos almacenados, y así se tienen la recuperación de los documentos mejor adaptados. Un concepto básico para la recuperación de información es la medición de similitud: se hace una comparación entre los dos documentos, midiendo qué tan similares son los documentos. Para la comparación, incluso un pequeño conjunto de palabras de entrada

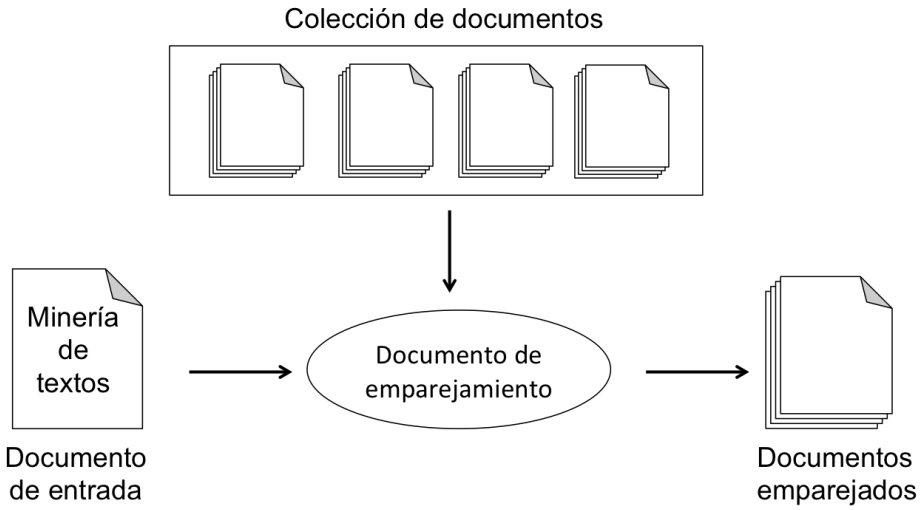

Figura 2.3: Agrupación de documentos.

en un motor de búsqueda puede ser considerado como un documento que puede ser adaptado a otros. Desde esta perspectiva, la similitud de medición se relaciona con los métodos de predicción para el aprendizaje y la clasificación que se llaman métodos del vecino más cercano. El tema más abordado es la similitud de medición y las variaciones de estos métodos, siendo éstos fundamentales para la recuperación de información.

- **Agrupación y organización de documentos**
  En la categorización de texto el objetivo es colocar nuevos documentos en las carpetas correspondientes. Estas carpetas son creadas por una persona con conocimiento de la estructura del documento, alguien que conoce los temas previstos.

  El objetivo general se ilustra en la Figura 2.4. Dada una colección de documentos, se necesita encontrar un conjunto de carpetas de manera que cada una tiene documentos similares. El proceso de agrupamiento es equivalente a la asignación de las etiquetas necesarias para la categorización de texto. Debido a que hay muchas maneras de agrupar los documentos, no es tan eficaz como un proceso de asignación de respuestas (es decir, etiquetas correctas conocidas) a los documentos. Sin embargo, la agrupación puede ser perspicaz. Mediante el estudio de palabras clave que caracterizan a un grupo, se puede conocer algunas de sus características más representativas.

- **Extracción de información**
  La representación de datos se ve en la información en términos de palabras. Esta es una técnica que es sorprendentemente exitosa para muchas aplicaciones. Es necesario definir los atributos que se van a considerar y tenerlos en una base de datos. Por ejemplo, pueden ser variables con valores reales, como el volumen de ventas, o un

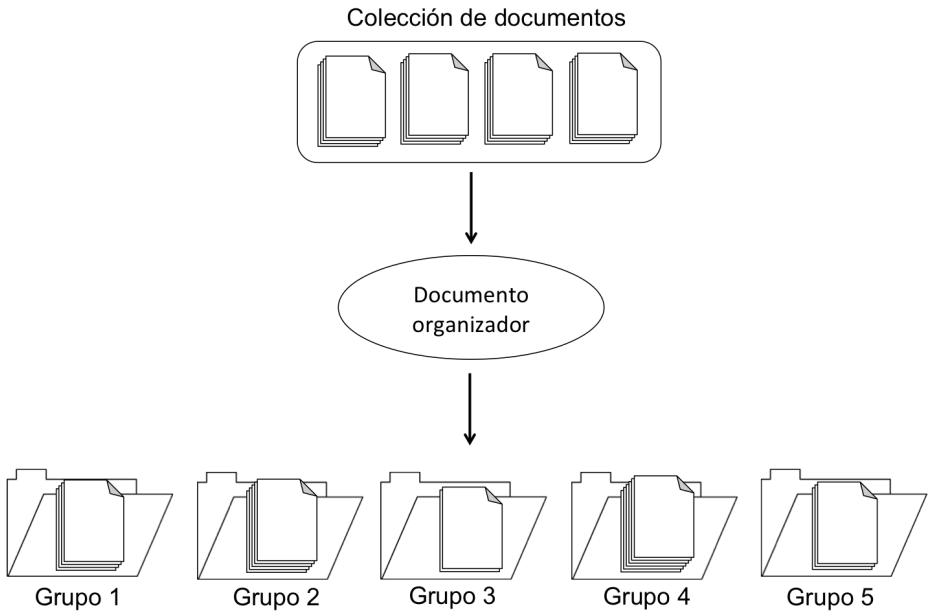

Figura 2.4: Organización de documentos.

código. Posteriormente, para poder extraer información, es necesario lograr que el texto este estructurado, por lo que se emplea una representación que mide la ocurrencia de las palabras. La extracción de información es un subcampo de la minería de texto que intenta mover la minería de texto de la misma forma que el mundo estructurado de minería de datos. En la figura 2.5 se ilustra la tarea de extracción de información. El objetivo es tomar un documento estructurado y rellenar automáticamente los valores de una hoja de cálculo. Una base de datos cuenta con una estructura que considera tablas y campos definidos. Cuando la información no esta estructurada, como la que se encuentra en una colección de documentos, se necesita de un proceso para extraer datos y tenerlos en un formato estructurado. Por ejemplo, se pueden examinar documentos sobre determinadas empresas y extraer los volúmenes de ventas a partir del texto del documento y posteriormente almacenar la información en una base de datos. El atributo que se mide no tiene una posición fija en el texto y no siempre es descrito de la misma manera en diferentes documentos.

Considerando las técnicas antes mencionadas, y de acuerdo a uno de los propósitos que tiene esta tesis, se va a considerar una técnica de agrupación para la extracción de términos relevantes de Twitter. La bolsa de palabras es un modelo que se construye a nivel de una oración, lo cual permite que se considere la frecuencia de los términos en un documento.

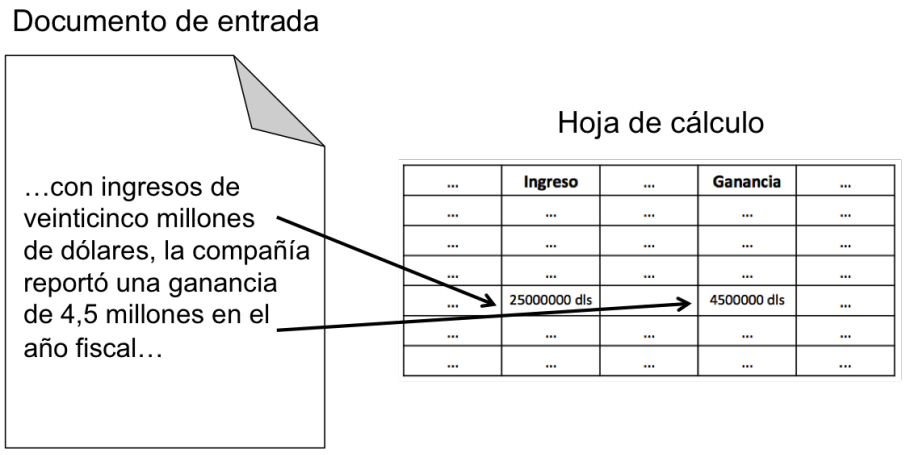

Figura 2.5: Extracción de información.

2.4. Procesamiento de lenguaje natural

El procesamiento de lenguaje natural (NLP - Natural Language Processing) es la función de componentes de software o hardware en un sistema de cómputo que analiza o sintetiza el lenguaje hablado o escrito [22]. NLP tiene diversas aplicaciones, tales como los motores de búsqueda, los traductores y las herramientas para realizar resúmenes automáticamente, todo con la finalidad de facilitar tareas análisis de texto.

Es importante considerar que la minería de textos implica un conjunto de tareas de procesamiento de lenguaje natural que permiten la identificación de entidades (ver Figura 2.6).

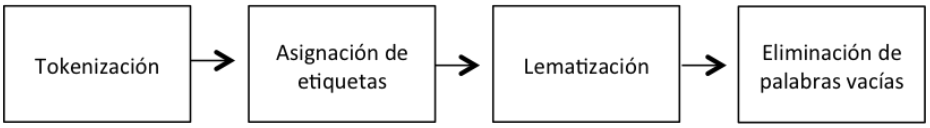

Figura 2.6: Tareas de procesamiento de lenguaje natural.

- **Tokenización**
  Suponiendo que la colección de documentos está en formato XML y se requiere examinar el texto no estructurado para identificar características útiles. El primer paso en el manejo de texto es para romper el flujo de caracteres en palabras o, más precisamente, tokens. Esto es fundamental para su posterior análisis. Sin la identificación de los tokens, es difícil imaginar extraer mayor información sobre el documento. Cada token es una instancia de un tipo, por lo que el número de tokens es mucho mayor que el número de tipos. Propiamente hablando, siempre se debe hacer referencia a la frecuencia de ocurrencia de un tipo. Romper un flujo de caracteres en tokens es

trivial para una persona familiarizada con la estructura del lenguaje. Un programa de computadora, sin embargo, al ser cuestionado lingüísticamente, se encuentra con la tarea más complicada. La razón es que ciertos caracteres que son delimitadores a veces son tokens y a veces no, dependiendo de la aplicación. El espacio de caracteres, tabulación y nueva línea asumimos siempre son delimitadores y no se cuentan como tokens. A menudo son llamados colectivamente los espacios en blanco. Los símbolos " () <>! ?" siempre son delimitadores y también pueden ser tokens dependiendo del entorno. Por ejemplo, un punto, una coma o dos puntos entre los números normalmente no se consideran un delimitador sino más bien parte del número. Cualquier otra coma o dos puntos es un delimitador y puede ser crucial. Un período puede ser parte de una abreviatura (por ejemplo, si tiene una letra mayúscula en ambos lados). También puede ser parte de una abreviatura cuando seguido de un espacio (por ejemplo, Dr.). Sin embargo, algunos de estos representan el termino de las oraciones. Un guión es un terminador y un token si esta precedido o seguido de otro guión. Un guión entre dos números podría ser un símbolo de resta o un separador (por ejemplo, 555 a 1212 como un número de teléfono).

Para obtener las mejores características posibles, siempre se debe personalizar el tokenizador para el trabajo disponible en texto, de lo contrario puede ser necesario añadir después se obtienen los tokens. Se debe tener en cuenta que el proceso de tokenización depende del idioma.

Esta tarea también es conocida como análisis léxico. Su objetivo es segmentar un conjunto de caracteres en unidades con significado llamados tokens. Antes de realizar cualquier análisis lingüístico o de procesar un documento es necesario encontrar y separar cada uno de los elementos que lo conforman. En la Figura 2.7 se muestra un ejemplo de tokenización.

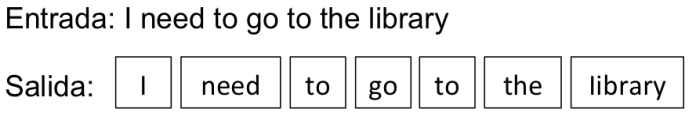

Figura 2.7: Ejemplo de tokenización.

Como se puede observar en el ejemplo anterior, la frase es segmentada en siete tokens, empleando como delimitador de cada token el espacio en blanco.

- **Asignación de etiquetas**

  Esta tarea es efectuada por herramientas que realizan el proceso de asignar partes de la oración u otra clase de marcador léxico a cada palabra en un corpus, en otras

palabras, son sistemas que ayudan en la determinación de la categoría gramatical (por ejemplo, si es verbo, sustantivo, preposición, etc) de cada una de las palabras de un texto o conjunto de ellos. Asimismo, el etiquetado que realizan estas herramientas se aplica de igual forma a signos de puntuación, números, cantidades, entre otros. Dichas herramientas también son conocidas como etiquetadores POS por sus siglas en inglés "Part-of-Speech".

Para llevar el proceso del etiquetado POS es necesario primeramente realizar un análisis morfológico. Un análisis morfológico o análisis estructural es el proceso de descomponer palabras complejas en sus componentes morfológicos (partes significantes de las palabras), dando como salida información sobre la semántica de la palabra y el papel sintáctico que juega en una oración. En otras palabras, un análisis morfológico permite conocer las posibles categorías gramaticales de cada una de las palabras que se encuentran en una oración.

- **Lematización**

  En todo documento escrito en una lengua flexiva, como el español y el italiano, existen múltiples variaciones léxicas de las palabras. Dentro del procesamiento de lenguaje natural es necesario disminuir la cantidad de variaciones léxicas que existan en los documentos a analizar. Para ello se debe obtener el lema o forma canónica, es decir, la base o la forma de diccionario de una palabra.

  El proceso de lematización se lleva a cabo de manera automática por parte de los humanos; cuando queremos buscar las palabras 'encontramos' o 'niñas' en un diccionario las pasamos a 'encontrar' y 'niño', para ello empleamos nuestro conocimiento del mundo. Pero para las computadoras realizar este procedimiento es más complicado, ya que no tienen acceso a este conocimiento, por tanto, es necesario darle una serie de reglas y de recursos.

  Para reducir el número de variaciones léxicas, ya sea por flexiones (caminar - caminamos) o por derivaciones (activar - activación), existen dos herramientas que se emplean en PLN, que son los lematizadores y los truncadores o stemmers. Aunque frecuentemente se confunden ambos términos, cabe aclarar que son dos métodos distintos.

  Los lematizadores son herramientas que emplean diccionarios, al igual que reglas, que buscan obtener el lema de las palabras.

- **Palabras vacías**

  Las palabras funcionales, palabras vacías o stopwords, son las palabras que 'carecen' de significado. Estas palabras son las de mayor frecuencia y las que aportan la menor cantidad de información, entre ellas se encuentran los artículos, las preposiciones, las

conjunciones, entre otras. De las palabras funcionales se crean listas de paro o stoplists. Hay autores que consideran que las palabras funcionales son automáticamente extraídas de un corpus genérico como aquellas con la más alta frecuencia, y posteriormente son validadas por expertos humanos [23]. Sin embargo, existe la posibilidad de agregar algunas otras palabras que se desean eliminar en el procesamiento de lenguaje natural.

El objetivo de emplear stoplists en PLN es reducir la cantidad de datos a analizar. De igual manera disminuye el espacio en memoria o en disco empleado por las herramientas que analizan lenguaje natural.

Capítulo 3

Análisis de Twitter

La experimentación de este trabajo se realizó en Twitter. Este capítulo describe las principales características de esta red social en línea. Además se presenta el estado del arte referente al análisis de datos de Twitter. En primer lugar se da una clasificación del tipo de herramientas existentes. Posteriormente se dan las características principales de aplicaciones que son representativas de cada grupo. Finalmente se hace un análisis comparativo entre ellas.

3.1. Twitter y sus características

Twitter es considerada una red social que le permite a sus usuarios compartir con el mundo información en tiempo real, obteniendo así actualizaciones instantáneas que resultan interesantes para muchas personas. La idea de Twitter es lograr conectar gente con intereses similares. Esta red social consiste en la composición de mensajes cortos denominados tweets.

Uno de los atractivos principales de Twitter es que los mensajes están restringidos en tamaño, permitiéndole a los usuarios identificar rápidamente si algún mensaje es particularmente interesante. Por el lado de la escritura de los mensajes, la restricción de tamaño obliga a estructurar muy bien lo que se quiere decir. La brevedad en los mensajes es probablemente una de las razones por las que Twitter es una de las redes sociales más populares.

A continuación se mencionan las actividades que caracterizan esta red social:

1. Se pueden publicar mensajes hasta de 140 caracteres (incluidos videos, enlaces y fotos).

2. Tiene una funcionalidad denominada “siguiendo”, la cual consiste en suscribirse a una cuenta de Twitter específica (obtener seguidores o followers). Esta actividad permite ver inmediatamente los mensajes que los usuarios a los que se siguen publican en su cuenta.

3. Es posible enviar mensajes directos para mantener conversaciones privadas entre los

usuarios que se siguen mutuamente. Los mensajes tienen un límite de 140 caracteres y pueden contener texto, etiquetas, enlaces, fotos y videos.

4. Cada usuario tiene la posibilidad de crear una lista de grupos de otros usuarios de Twitter por tema o interés (por ejemplo, una lista de amigos, celebridades, atletas). Estas listas también incluyen una cronología de tweets de usuarios específicos que fueron agregados a la lista.

5. Compartir el tweet de un usuario con todos los seguidores se denomina "retweet". Los retweets son utilizados generalmente para compartir noticias o descubrimientos importantes en Twitter, manteniendo su atribución original.

3.1.1. Elementos de un tweet

Los tweets además de estar limitados en tamaño, se caracterizan por constar de diversos elementos. Pueden incluir nombres de @usuario, #etiquetas e hipervínculos [24]. En la Figura 3.1 podemos observar ejemplos de tweets. A continuación se describen cada uno de estos elementos:

Usuarios @: en un tweet se puede enlazar a cualquier otro usuario de Twitter, independientemente de que se tenga o no en la lista de seguidores. Esta mención se notifica al usuario en cuestión.

Etiquetas #: una etiqueta, también llamada hashtag, es cualquier palabra o frase precedida directamente por el símbolo # y su finalidad es crear temáticas. De esta manera cualquier usuario que publique un tweet incluyendo una etiqueta será encontrada fácilmente. Además, si un usuario hace clic en una palabra con etiqueta, es posible visualizar todos los tweets que incluyen esa palabra clave o tema.

Hipervínculos: otro de los elementos que puede ser añadido en un tweet son los enlaces a otros sitios web, artículos, fotos y videos.

Un tweet tiene características específicas que lo identifican como tal (ver Figura 3.2), estas son:

Nombre de la cuenta de Twitter: Este dato proporciona el nombre del usuario, pero no necesariamente debe utilizarse el nombre verdadero, también puede ser utilizada alguna otra identificación.

Nombre de usuario: El @nombredeusuario es la identidad única en Twitter. A partir del nombre de usuario se crea una URL correspondiente al perfil del usuario en Twitter, por ejemplo: twitter.com/nombredeusuario. El signo @ también se usa para mencionar a usuarios en tweets, por ejemplo: ¡Comenzando en @Twitter!. Es importante recalcar que el nombre de cuenta y el nombre de usuario no necesariamente deben ser el mismo.

Figura 3.1: Ejemplos de tweets.

Foto de perfil: Imagen personal que se carga en el perfil de twitter a través de la pestaña "Configuración".

Marca de tiempo del tweet/fecha: Esto indica cuándo se envió un tweet. Al crear un tweet se genera un enlace permanente el cual puede consultarse en cualquier instante.

Texto del tweet: Cada tweet debe contener como máximo 140 caracteres. El tamaño ideal para una gran idea, un titular o una observación oportuna.

Figura 3.2: Elementos de un tweet.

3.1.2. Consultas por nombre de usuario

Existen diferentes maneras de realizar consultas en Twitter, las cuales varían en función del tipo de información que se quiere obtener, ya que el alcance de los tweets de una cuenta depende del tipo de consulta. A continuación se describen las cuatro diferentes consultas que se pueden ejecutar:

Tweets hacia, desde y sobre una cuenta "ejemplo". Ejecutar un consulta con un nombre de usuario, pero sin incluir el símbolo '@', devuelve todas las menciones de dicha cuenta de Twitter (incluyendo retweets y respuestas), así como todos los tweets dicha cuenta. Este es el conjunto más completo de estadísticas de alcance para una cuenta específica de Twitter.

Tweets desde y alrededor de una cuenta "@ejemplo". Ejecutar una consulta con un nombre de usuario incluyendo el símbolo '@', devuelve todas las menciones de la cuenta, pero no cualquier tweet de dicha cuenta. Esta consulta permite saber cuántas personas están mencionando cierta cuenta de Twitter, y las formas en que están hablando de ello (incluye todos los retweets, respuestas y menciones).

Tweets de una cuenta "a: ejemplo". Se requiere ejecutar la consulta incluyendo el operador 'a:' y un nombre de usuario, sin utilizar el símbolo @. De este informe sólo se obtendrán respuestas directas a dicha cuenta (en donde el nombre de usuario es la primera palabra en el Tweet). Esta consulta es útil para saber qué es lo que la gente comenta a esa cuenta.

Tweets de una cuenta "de: ejemplo". Se requiere ejecutar una consulta con el operador 'de:' y un nombre de usuario, sin utilizar el símbolo @. Esta consulta sólo obtiene los tweets de esa cuenta. Es útil para medir el alcance de una cuenta individual de Twitter, y para averiguar los tipos de tweets que el usuario de la cuenta publica.

3.1.3. Modelos de conversación

La estructura de los grupos existentes en las redes sociales puede tomar diversas formas de acuerdo a su comportamiento. Particularmente en Twitter hay por lo menos seis estructuras distintas de comunidades que se forman en función del tema que se discute, las fuentes de información que se citan, las redes sociales de las personas que hablan sobre el tema y los líderes de la conversación. Cada uno tiene una forma y estructura social diferente: dividido, unificado, fragmentada, agrupado, y con centro de actividad y radio hacia el interior o hacia el exterior.

Con base en la realización de un análisis de miles de representaciones de Twitter por un centro de investigación [25], se identifican seis diferentes tipos de conjuntos en esta red social.

- **Conjuntos polarizados:** Este tipo de estructuras cuentan con dos grupos grandes y densos que tienen poca conexión entre ellos. Los temas que se discuten son de gran interés social, por ejemplo la política. De hecho, por lo general hay poca conversación entre estos grupos a pesar del hecho de que ellos se centran en el mismo tema. Los grupos polarizados en Twitter no están discutiendo entre ellos, de hecho no se toman en cuenta unos a otros, lo que hacen es hacer referencia a diferentes recursos de la Web, así como el uso de diferentes hashtags.

- **Conjuntos estrechos:** Este tipo de discusiones se caracterizan porque las personas están altamente interconectadas con pocos participantes aislados. Muchas conferencias, temas profesionales, grupos de aficionados y otros temas que atraen a determinadas comunidades hacen que se formen este tipo de estructuras.

- **Conjuntos de marcas:** Cuando productos reconocidos, servicios o temas populares (por ejemplo: celebridades) se discuten en Twitter, a menudo hay comentarios de muchos participantes desconectados. Es así como temas populares pueden atraer a grandes poblaciones fragmentadas de Twitter, en donde usuarios "aislados" se enfocan en comentar a cerca de un mismo tópico, pero sin conversar entre ellos. Cuanto mayor sea la población que habla de una marca, es menos probable que los participantes estén conectados el uno al otro.

- **Conjuntos de comunidades:** Existen algunos temas que son populares y pueden desarrollar múltiples grupos más pequeños, cada uno con su propia audiencia, personas influyentes, y fuentes de información. Estos grupos de comunidades parecen comercios con múltiples centros de actividad. Noticias globales suelen atraer la cobertura de muchos medios, cada uno con su propio seguimiento; lo cual crea una colección de grupos medianos y un buen número de aislamientos.

- **Redes de difusión:** Comentarios de Twitter en torno a noticias de última hora y su salida de medios conocidos y expertos, tienen un centro distintivo y una estructura en la que la mayoría de los usuarios repiten lo que los diarios y medios de comunicación destacados publican en Twitter. Los miembros de este tipo de audiencias a menudo están conectados únicamente a la fuente de noticias, sin necesidad estar conectados entre ellos. En algunos casos existen subgrupos más pequeños de personas que sí se encuentran conectadas y que discuten las noticias unos con otros.

- **Redes de soporte:** Las quejas de clientes para un negocio importante son a menudo manejados por una cuenta de servicio de Twitter, la cual intenta resolver y gestionar problemas de los clientes en torno a sus productos y servicios, generando así una estructura contraria a las redes de difusión. En las redes de soporte el centro del grupo da respuestas a los múltiples usuarios que se encuentran desconectados entre ellos, lo que genera radios hacia afuera. En contraste, en las redes de difusión, el centro de información es considerado mediante retweets por muchas personas desconectadas, creando así radios interiores.

3.2. Estado del arte

Considerando la cantidad de estudios relacionados con el análisis de datos generados en redes sociales, a continuación se describen algunos trabajos de investigación relacionados con Twitter. En particular se mencionan tres tareas en las que se pudieron ubicar dichos trabajos, de acuerdo al tipo problema que resuelven.

1. Creación de Corpus

El auge de los medios sociales y otras formas de contenido generado por los usuarios han creado la demanda de búsqueda en tiempo real: en contra de una corriente de alta velocidad de documentos entrantes, los usuarios desean una lista de resultados relevantes en el momento en que se emita la consulta.

La red de información en tiempo real de Twitter es uno de los temas de investigación para las tareas de recuperación de información, tales como búsqueda en tiempo real. Sin embargo, es importante considerar las restricciones impuestas por los términos de servicio de Twitter.

En el contexto de búsqueda en tiempo real de los tweets, el artículo [26] describe una arquitectura de recuperación de datos en dos etapas, la primera se encarga de la generación de candidatos, y la segunda etapa consiste en revisiones manuales para obtener la salida final.

En [27] se detalla una nueva metodología para la creación y difusión de corpus Twitter, desarrollado a través de la colaboración entre la Conferencia de texto Recuperación (TREC) y Twitter. Además, se analiza si este enfoque de distribución sigue siendo robusto con el tiempo.

2. Localización de eventos

La mayoría de los enfoques que tienen como objetivo extraer información de eventos de fuentes que suelen utilizar el contexto temporal de los textos. Sin embargo, aprovechar la información de la ubicación de los mensajes georreferenciados, también es importante para detectar eventos localizados, tales como eventos públicos o situaciones de emergencia.

En [28] se presenta un nuevo marco para detectar eventos localizados en tiempo real, a partir de una cuenta de Twitter, y realizar un seguimiento de la evolución de este tipo de eventos en el tiempo. Para ello, las características espacio temporales de las palabras clave están siendo extraídos continuamente para identificar candidatos significativas para las descripciones de eventos. Para determinar los eventos más importantes en determinado plazo se introduce un sistema de puntuación para los eventos. Se demuestra la funcionalidad de este sistema, llamado Even Tweet, usando una corriente de tweets de Europa durante el

Campeonato de Europa de la UEFA de 2012.

3. Recuperación de información
Uno de los temas principales en esta categoría dentro de redes sociales es la detección de eventos en Twitter. Ya que esta red social está emergiendo rápidamente en los últimos años, parte de los usuarios están utilizando Twitter para reportar los eventos de la vida real.

Existen investigaciones que se centran en la detección de eventos mediante el análisis de los textos publicados en Twitter. Aunque la detección de eventos ha sido durante mucho tiempo un tema de investigación, las características de Twitter hacen que no sea una tarea trivial. Simplemente la existencia de Tweets que tienen texto sin sentido.

Los algoritmos de detección de eventos existentes se pueden clasificar en dos categorías, los que detectan eventos con base en la distancia semántica entre tweets, y los que descubren los acontecimientos mediante la distribución de palabras.

En [29] se describe un trabajo que aborda este tema. EDCoW (Detección de eventos con agrupación de señales basado en wavelets) construye señales para palabras individuales mediante la aplicación de análisis de las señales basadas en la frecuencia de las palabras. Se eliminan las palabras triviales, examinado sus correspondientes autocorrelaciones de señal. Las palabras restantes se agrupan para formar los eventos con una técnica de particionamiento gráfico basado en la modularidad.

Este trabajo se enfoca en la recuperación de información. Es por ello que a continuación se hace un análisis de las soluciones existentes para este problema en particular.

3.2.1. Tipos de análisis de Twitter

Actualmente existen muchas herramientas que permiten realizar el análisis de tweets de formas variadas. Sin embargo, es importante considerar que cada una de estas aplicaciones tiene ciertas ventajas una sobre otra.

En primera instancia es posible realizar una clasificación considerando que el uso algunas conlleva un costo. Entonces es posible contemplar dos grupos, herramientas comerciales y herramientas libres.

Herramientas comerciales
El primer grupo tiene como característica principal el requerir de un pago mensual para poder hacer uso de estas aplicaciones. El tipo de pago varia dependiendo de cada aplicación, éste puede ser por la cantidad de reportes obtenidos o por periodos de tiempo, ya se mensual o anual. Así mismo, algunas de estas herramientas cuentan con un periodo de prueba para conocer su funcionamiento.

Las herramientas comerciales en su mayoría son de tipo Web, por ejemplo: Twitter Ar-

chivist [refnum1] y Topsy [refnum2]. Cabe señalar que en algunas de las aplicaciones de este tipo los datos que son considerados para realizar el análisis de información sólo se pueden visualizar en la interfaz al momento de efectuar la solicitud del servicio, y no siempre están disponibles para su almacenamiento. Los resultados que son obtenidos por estas herramientas se pueden visualizar en la misma interfaz y hay algunos casos en los que es posible obtener un reporte y almacenarlo.

Herramientas libres

Este segundo grupo contempla todas las aplicaciones que son de uso libre, es decir, no se requiere realizar ningún pago para poder utilizarla. Asimismo es importante resaltar que se tienen ciertas limitaciones en este tipo de aplicaciones, por ejemplo, la cantidad de tweets a analizar o los periodos de tiempos considerados para obtener los tweets.

Además de considerar si las herramientas son gratuitas o no, también se puede hacer una clasificación más específica. De acuerdo a los resultados obtenidos por las aplicaciones existentes es evidente que hay diferentes tipos de análisis que pueden realizarse a los datos de Twitter. A continuación se describe cada una de las categorías definidas a partir de este concepto.

- **Análisis estadístico**

  Hay aplicaciones que tienen como objetivo obtener cifras representativas de Twitter. Es posible obtener cantidades relacionadas a un usuario de Twitter en particular, por ejemplo, cantidad de seguidores y número de tweets publicados. De igual forma puede hacerse el conteo de los tweets que incluyen determinado hashtag y los usuarios que más lo utilizan.

  En general este tipo de aplicaciones permite monitorear la actividad ya sea de usuarios o de temas que son mencionados en Twitter, contando así con cifras que permitan visualizar la actividad que se da en determinado en esta red social y el impacto que ésta tiene.

- **Análisis de sentimientos**

  El análisis de sentimientos es un proceso que intenta predecir el tipo de sentimiento que los usuarios expresan a través de información obtenida de los tweets. El objetivo de las herramientas que realizan este tipo de análisis es determinar el tipo de publicaciones realizadas por uno o varios usuarios de Twitter y así conocer la opinión que se tiene sobre un producto, noticia o figura pública en particular.

  La salida general que pueden obtenerse por este tipo de aplicaciones es saber si la opinión dada en un conjunto de tweets es positiva, negativa o neutra.

- **Análisis de tópicos y/o hashtags**

  Hay herramientas que permiten hacer el análisis de tweets basando la obtención de tweets en la búsqueda de palabras clave en esta red social. Una vez obtenido el conjunto de datos se obtiene información como: los hashtags relacionados con el término definido, nombres de usuario que utilizan ese término y las palabras más mencionadas.

  Los resultados dados por este tipo de herramientas en general son listas de nombres de usuario, de hashtags o de términos clave. También hay aplicaciones que muestran los tweets que fueron encontrados y en los cuales se realiza el análisis para la obtención de resultados.

3.2.2. Herramientas existentes

Como se describe en la sección anterior, una de las posibles clasificaciones que puede hacerse de las herramientas que se enfocan en el análisis de datos de Twitter es de acuerdo al tipo de análisis que éstas realizan. A continuación se mencionan algunas de las aplicaciones más representantes para la obtención y/o análisis de datos de esta red social.

Twitter Analytics: Architecture, Tools and Analysis

Esta herramienta se encarga de estudiar el comportamiento temporal de los mensajes que llegan a una red social [30]. En la documentación encontrada se describe un estudio realizado específicamente a los tweets y retweets enviados al presidente Barack Obama en Twitter. Se consideran los tiempos entre llegadas entre los tweets, el número de retweets y las coordenadas espaciales (latitud, longitud) de los usuarios que enviaron los tweets. El modelado del proceso de llegada de tweets en Twitter puede ser aplicado para predecir el comportamiento cotidiano del usuario en las redes sociales.

Este equipo de trabajo desarrolló una arquitectura de software que utiliza una interfaz de programación de aplicaciones de Twitter (API) para obtener los tweets enviados a usuarios específicos. Posteriormente, se extraen los identificadores de usuario y las marcas de hora exactas de los tweets. Se utilizan los datos obtenidos para caracterizar los tiempos entre llegadas entre tweets y el número de retweets. Los estudios realizados en esta investigación indican que el proceso de llegada de nuevos tweets a un usuario puede ser modelado como un proceso de Poisson, mientras que el número de retweets sigue una distribución geométrica.

La arquitectura de la recopilación de datos funciona de manera independiente al sistema operativo. Los resultados obtenidos en esta investigación se pueden aplicar para estudiar las correlaciones entre los patrones de comportamiento de los usuarios y sus ubicaciones.

Twitter Archivist

Tweet Archivist es una aplicación que puede guardar tweets antes que desaparezcan [31]. Los tweets obtenidos son referentes a un usuario en particular, un hashtag o un término cualquiera. Tweet Archivist no tiene acceso a todos los tweets que se han publicado, pero una vez que éstos han sido guardados en un archivo se tiene la seguridad de que no se perderá ningún tweet.

La característica principal de Tweet Archivist es que genera 10 visualizaciones basadas en cada archivo de tweets, previendo así tendencias y comportamientos. La información que obtiene son: usuarios, palabras, URLS, origen del tweet, idioma, número de tweets, menciones de usuarios, hashtags, imágenes y número de seguidores. Es importante mencionar que la aplicación es web y tiene un costo mensual.

Topsy

Topsy es un motor de búsqueda en tiempo real impulsado por la Web social [32]. A diferencia de los motores de búsqueda tradicionales, Topsy clasifica y posiciona los resultados de una búsqueda basados en las conversaciones más influyentes que millones de personas están teniendo todos los días sobre cada término o tema específico consultado.

Topsy fue diseñado para ayudar a dar sentido a los miles de millones de tweets que son publicados en Twitter. Esta herramienta en su versión básica es gratuita. Pero también existe la versión Pro que cuenta con funciones más sofisticadas disponibles por una cuota anual.

Topsy cuenta con una base de datos que contiene todas las publicaciones de Twitter que incluye desde el primer tweet del 2006, lo que implica alrededor de 425 millones de publicaciones. Topsy ordena sus resultados para buscar, analizar y extraer ideas de conversaciones y las tendencias en los sitios web públicos sociales como Twitter y Google+.

La búsqueda Topsy es más poderosa que la búsqueda estándar y avanzada de Twitter. Se pueden buscar tweets haciendo referencia a una persona así como rastrear popularidad de una palabra clave o menciones de un dominio a través del tiempo. Las actividades que Topsy permite son:

- Muestra todas las acciones referentes a un término específico.
- Encuentra influyentes sobre un tema específico.
- Buscar todo el contenido de una persona.
- Buscar todos los tweets de un usuario y los enlaces que se están compartiendo.

- Encuentra resultados de sentimiento de una marca y un seguimiento en el tiempo.
- Lleva a cabo análisis detallados de seguimiento de palabras clave.

TwiCube: A Real-Time Twitter Off-Line Community Analysis Tool

TwiCube es una herramienta en línea que emplea un novedoso algoritmo capaz de identificar la comunidad social de la vida real de un usuario, lo que sería una comunidad fuera de línea del usuario, sólo a partir de la relación de estructura entre los seguidores y followers del usuario a examinar [33]. Con base en la comunidad fuera de línea identificada, TwiCube proporciona un resumen de los intereses del usuario, hábitos y análisis de popularidad.

TwiCube es una herramienta en línea para el análisis de la comunidad fuera de línea. El proceso funcionamiento de esta herramienta inicia con un bucle de ejecución de TwiCube que se desencadena por una consulta para un determinado usuario de destino y culmina con la visualización de la red de amigos fuera de línea y estadísticas relacionadas. Se utilizan dos recursos externos: Twitter y FreeBase. Sin embargo es importante mencionar que no se lograron obtener resultados al realizar diferentes pruebas.

TwitterStand: News in Tweets

Este sistema surgió como resultado de una investigación acerca del uso de Twitter [34]. TwitterStand permite el procesamiento de noticias a partir de tweets. La actividad clave es la captura de los tweets que corresponden a las noticias de última hora. El resultado es análogo a un servicio de noticias distribuido con la diferencia de que las identidades de los contribuyentes no se conocen. Por otra parte, una característica importante es que los tweets no se encuentran organizados, sino que se producen conforme las noticias van sucediendo.

Esta herramienta aborda tres tareas principalmente: quitar el ruido existente en los tweets obtenidos, la determinación de grupos con ciertos intereses y la determinación de los lugares relacionados con los tweets.

Un aspecto interesante de TwitterStand es la agrupación de los tweets, la cual toma en cuenta la colección de los hashtags correspondientes a un tema en particular (noticia).

3.2.3. Análisis comparativo de herramientas existentes

La Tabla 3.3 muestra una comparación de las aplicaciones previamente descritas considerando la fuente de datos, el preprocesamiento y el resultado, ya que son elementos importantes de la herramienta que se propone en esta tesis.

Herramienta	Fuente de datos	Pre procesamiento	Salida
Twitter Analytics	Casos específicos (tweets y retweets enviados al presidente Barack Obama).	Modelado del proceso de llegada de tweets.	Comportamiento de llegadas de los tweets a un persona en particular.
TwitterStand	Tweets recientes (obtenidos al momento de realizar el análisis)	Eliminar el ruido existente en los tweets, determinación de grupos con ciertos intereses y la obtención de lugares relacionados con los tweets.	Hashtags relacionados y visualización de los lugares geográficos en que se encuentran las personas que publicaron.
TwiCube	Consulta información sobre determinado usuario.	Analiza la estructura de seguidores y usuarios seguidos.	Red de amigos fuera de línea y estadísticas relacionadas.
Twitter Archivist	Tweets referentes a un usuario en particular, un hashtag o un término cualquiera.	No se menciona (herramienta comercial)	Usuarios, palabras, URLS, origen del tweet, idioma, número de tweets, menciones de usuarios, hashtags, imágenes y número de seguidores.

Figura 3.3: Herramientas representativas de análisis de datos de Twitter.

Las herramientas existentes obtienen un conjunto de datos los cuales son contemplados para el funcionamiento de la aplicación. La fuente de datos se refiere a los datos (tweets) que son considerados por cada aplicación para que posteriormente se realice el análisis de los mismos.

Twitter Analytics considera sólo los siguientes datos de cada tweet: identificadores de usuario, marcas de tiempo exactas, y coordenadas espaciales. Pos su parte, TwitterStand considera como fuente de datos los tweets obtenidos en el momento en que se realiza el análisis. TwiCube contempla información obtenida del perfil de usuario que se quiera analizar. Finalmente Twitter Archivist considera como parámetro de búsqueda de datos un usuario en particular, un hashtag o un término definido por el usuario. Como puede observarse, la herramienta más general en cuanto a la colección de datos es Twitter Arquivist, ya que cuenta con 3 parámetros de búsqueda distintos.

En cuanto al pre procesamiento que realiza cada herramienta a los datos almacenados difieren entre sí, esto conforma al tipo de resultado que desea alcanzarse. Twitter Analytics se encarga de realizar un modelo de tiempo de llegada de los tweets considerando el tiempo de llegada de cada texto. Mientras tanto TwitterStand lleva a cabo un procedimiento más completo, cuenta con una etapa de eliminación de ruido de los textos, la detección de grupos que compartan intereses y la obtención de los lugares en que fueron publicados los tweets. Por su parte TwiCube sólo se encarga de analizar la estructura de los usuarios con los que se interactúa. Referente a Twitter Archivist, a pesar de que es la herramienta más completa, no

proporciona información a cerca del procesamiento que se da a los datos para poder obtener los resultados señalados. Considerando las herramientas descritas, TwitterStand es la que aparentemente cuenta con el pre procesamiento más completo, pues considera una etapa de limpieza de tweets, que sin duda alguna es indispensable en este tipo de datos.

Finalmente, la salida hace referencia al tipo de resultados que se obtienen en cada herramienta, ya que cada una varía de acuerdo al tipo de análisis que se realiza a los datos. Twitter Analytics muestra cifras que permiten conocer el comportamiento de los tweets que llegan a un usuario en particular (frecuencia de llagada). TwitterStand da como resultado los hashtags relacionados con el conjunto de datos, así como la visualización de los lugares geográficos en que se encuentran las personas que publicaron dicha información. TwiCube permite crear una red de amigos fuera de línea, así como las estadísticas relacionadas con dicha red. Por último, Twitter Archivist da como resultado hasta 10 distintas salidas: usuarios, palabras, URLS, origen del tweet, idioma, número de tweets, menciones de usuarios, hashtags, imágenes y número de seguidores; la salida varia dependiendo del parámetro de búsqueda seleccionado. Tomando en cuenta el tipo de resultados que arroja cada herramienta, se puede considerar a Twitter Archivist como la que tiene mayor variedad en cuanto al tipo de representaciones de información dadas como salida.

Capítulo 4

Herramienta propuesta: Twitter AT

En este capítulo se describe el desarrollo de la herramienta *Twitter AT*. Se describen los módulos creados para la implementación del preprocesamiento de datos de Twitter. En primera instancia se explica el módulo de recolección de tweets. Posteriormente se especifican los métodos aplicados en la etapa de preprocesamiento de tweets. Finalmente se presenta el algoritmo implementado en el módulo de análisis de datos.

4.1. Descripción general del sistema

El objetivo principal de este trabajo de tesis es implementar una herramienta que permita realizar el preprocesamiento de datos de redes sociales en línea, particularmente Twitter. Debido a que la etapa de preprocesamiento de datos requiere de diferentes tareas para obtener resultados óptimos, es necesario integrar éstas en un módulo.

Para poder realizar pruebas del funcionamiento del módulo de preprocesamiento de tweets es necesario incluir dos módulos más como parte de la solución: Un módulo de obtención de datos y un módulo de análisis de datos.

El módulo de obtención de datos permite recolectar tweets con base en tópicos definidos por el usuario. Esta función ayuda a obtener información que va a ser la entrada para la etapa de preprocesamiento de datos.

El módulo de análisis de datos tiene como función principal poder verificar la utilidad de los datos de salida de la etapa de preprocesamiento. Para poder hacer pruebas se seleccionó el algoritmo TF-IDF que permite la ponderación de términos de un documento mediante la asignación de pesos.

Por último, otro aspecto significativo para alcanzar el objetivo definido es el almacenamiento de información. La forma en que se accede a los datos es clave en el tiempo de ejecución de las tareas de cada módulo. Para solventar esta dificultad se creó una base de datos que posibilita la interacción de cada módulo con los datos que precisa manipular. Esta funcionalidad de la herramienta permite a la vez crear múltiples corpus de tweets.

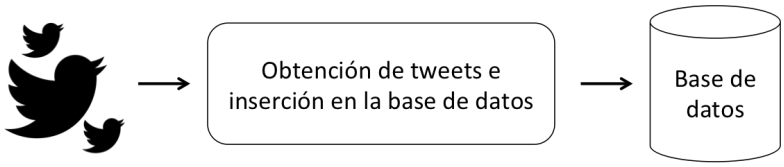

Figura 4.1: Fase de obtención de datos.

Características

Twitter AT integra la obtención de información de Twitter, limpieza de textos, reducción de información y el algoritmo TF-IDF. Para implementar el sistema fue necesario dividir el problema en distintas fases, de tal manera que se pueden distinguir tres módulos principales, cada uno de ellos dedicado a resolver una tarea específica.

Los módulos que componen la herramienta son: obtención de datos, preprocesamiento de tweets y análisis de datos.

- **Obtención de datos**
  Este módulo se encarga de la obtención de datos generados por usuarios de Twitter (ver Figura 4.1). Estos datos son preprocesados en la siguiente etapa. Para resolver esta tarea se desarrolló un módulo de recolección de datos de Twitter. Este módulo se encarga de crear corpus de tweets, con base en temas específicos, mediante peticiones al servidor de esta red social. La información obtenida se almacena posteriormente en la base de datos de la aplicación.

- **Preprocesamiento de tweets**
  La segundo módulo corresponde al preprocesamiento del corpus de tweets recolectados. El preprocesamiento consta de una serie de pasos para lograr eliminar información innecesaria y así reducir la cantidad de datos a analizar en la última etapa.

- **Análisis de datos**
  Este módulo incluye la aplicación del algoritmo TF-IDF en el conjunto de datos. Con esta implementación *Twitter AT* obtiene como salida los pesos de cada palabra de los tweets analizados. Este análisis permite encontrar los términos más relacionados con determinado tópico en Twitter.

En la Figura 4.2 se listan las funciones de cada módulo. Éstas son detalladas en la sección 4.4.

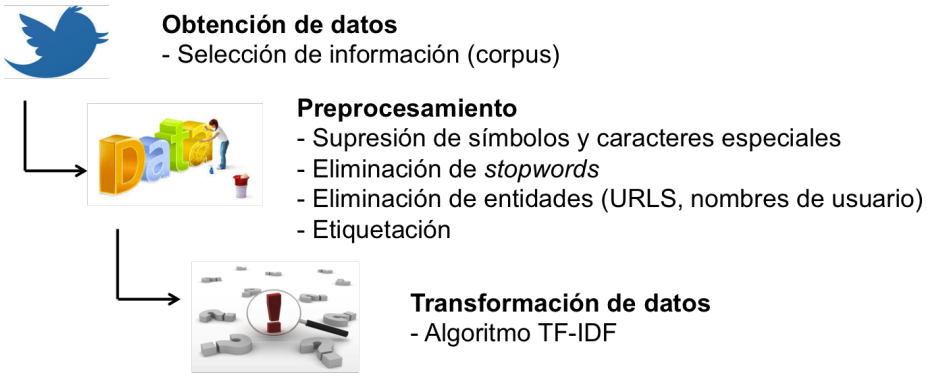

Figura 4.2: Metodología del proyecto (diagrama general).

4.2. Arquitectura

La arquitectura de *Twitter AT* consta de tres capas. Se decidió desarrollar la herramienta con este principio para que en un futuro sea fácil incluir nuevas funciones en los diferentes módulos. A continuación se mencionan las características de cada capa:

- **Capa de presentación:** Incluye las bibliotecas de Java necesarias para la creación de la interfaz de la herramienta.
- **Capa de negocios:** Consta de múltiples bibliotecas que se requieren para la implementación de cada una de las etapas de procesamiento de los datos. Estas bibliotecas son: Stanford CoreNLP, Stanford POS Tagger y Twitter4J.
- **Capa de datos:** Esta capa permite mantener los datos intactos mientras se están realizando las diferentes operaciones de la etapa de preprocesamiento. Los resultados finales se insertan en la base de datos.

Adicionalmente es necesaria la conexión a internet para poder obtener datos del servidor de Twitter. Para realizar esta operación se requiere de la comunicación con la API de esta red social.

En la Figura 4.3 se muestra una diagrama de la arquitectura de la herramienta.

4.2.1. Biblioteca Twitter4j y API Twitter

Twitter tiene la API REST (REpresentational State Transfer), que mediante un mecanismo de autenticación basado en el estándar OAuth permite leer y escribir datos de los diferentes objetos de Twitter (tweets, usuarios y lugares).

Para realizar la petición al servidor de Twitter, se hace uso de una de las APIS proporcionadas por Twitter: REST API. Esta API permite el acceso al core de los datos de

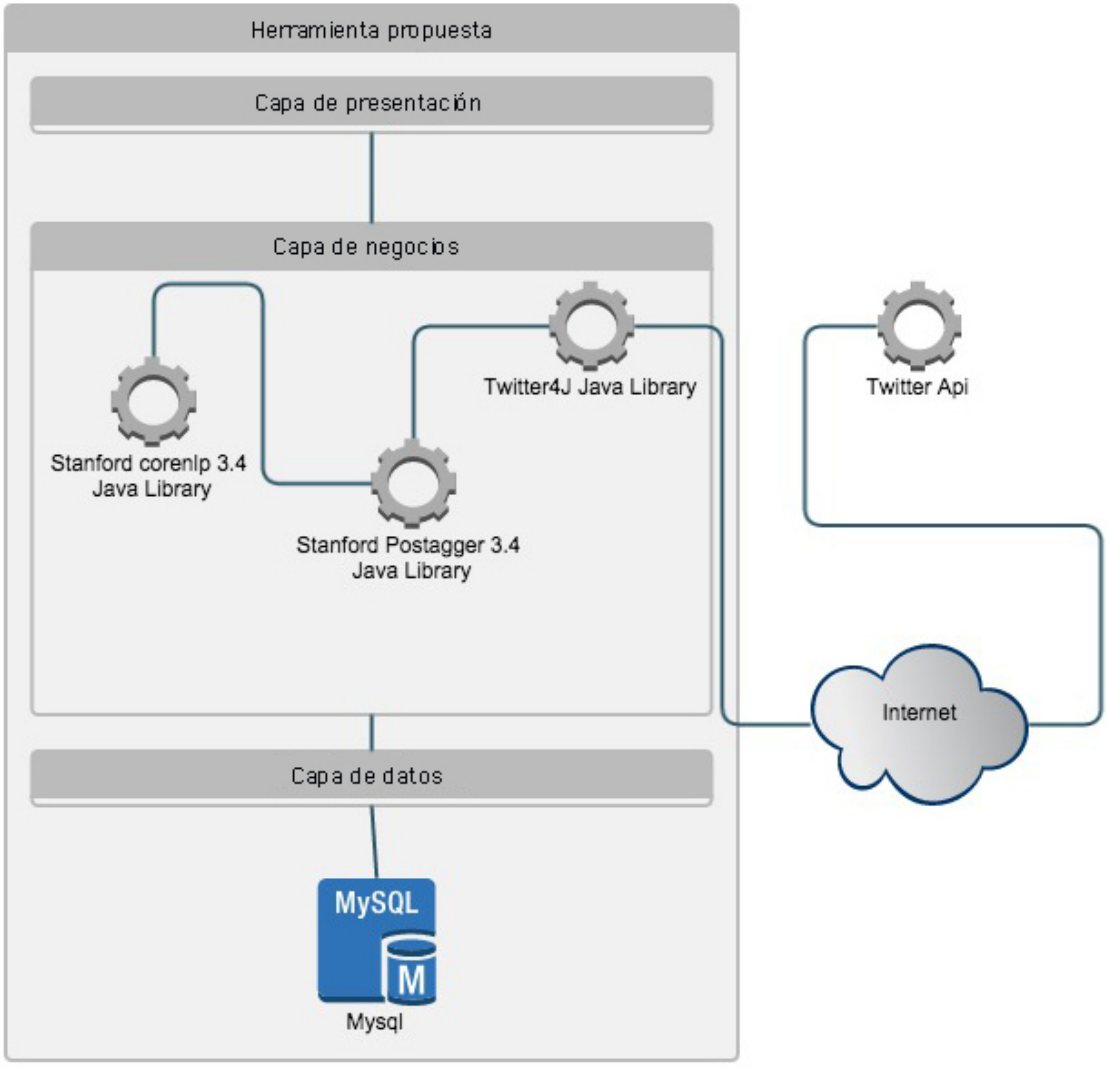

Figura 4.3: Arquitectura del sistema.

Twitter. Todas las operaciones que son posibles realizar vía Web son posibles realizarlas desde la API. La autenticación es necesaria de acuerdo al tipo de operación que se realice. Los formatos que soporta esta API son: XML, JSON y RSS.

La obtención de tweets se realiza a través de la API de Twitter, la cual permite realizar consultas mediante diferentes opciones de búsqueda. Para acceder a la API de Twitter a través de la herramienta creada es necesario registrar y autenticar la aplicación con los servidores de Twitter. Al momento de realizar las peticiones es necesario autenticarse tanto a nivel de aplicación, como a nivel de usuario. A continuación se muestra el pseudocódigo para la etapa de autenticación.

Algoritmo 1 Autenticación en Twitter

Entrada: Tokens de acceso OAuth generados en el sitio de Twitter (https://apps.twitter.com): Consumer Key (API Key), Consumer Secret (API Secret), Access Token y Access Token Secret.

Salida: Se registra un mecanismo para enviar peticiones seguras y autorizadas al API de Twitter mediante el protocolo OAuth (http://oauth.net).

1: Obtener una instancia de Twitter Factory (proporcionada por la biblioteca Twitter4J).
2: Registrar un OAuthConsumer en la instancia de Twitter Factory, con los valores Consumer Key y Consumer Secret.
3: Crear un AccessToken mediante los valores: Acces Token y Access Token Secret.
4: Registrar el OAuthAccessToken, generado en el paso anterior, en la instancia de Twitter Factory.

Twitter4J es una de las bibliotecas Open Source que implementa la funcionalidad del API REST de Twitter y está publicada bajo la licencia Apache 2.0. Esta biblioteca permite integrar aplicaciones Java con el servicio de Twitter.

4.2.2. Bibliotecas Stanford Core NLP y Stanford POS Tagger

La información que se obtiene mediante la biblioteca Twitter4j se procesa utilizando las bibliotecas que proporciona "The Natural Language Processing Group" de la Universidad de Stanford [35]. En principio se utiliza la biblioteca Core NLP, la cual proporciona un algoritmo para para poder obtener las raíces de las palabras que conforman un tweet.

Una vez que se encuentra la raíz de cada palabra se utiliza la biblioteca POS Tagger para clasificar las palabras de acuerdo a sus categorías gramaticales; por ejemplo sustantivo, verbo y adjetivo. Este paso permite reducir la cantidad de palabras identificadas en el conjunto total, evitando considerar palabras similares como elementos independientes. Esta biblioteca utiliza diferentes modelos de lenguaje para poder realizar la clasificación; para nuestra aplicación se utiliza el modelo "English Twitter POS Tagger".

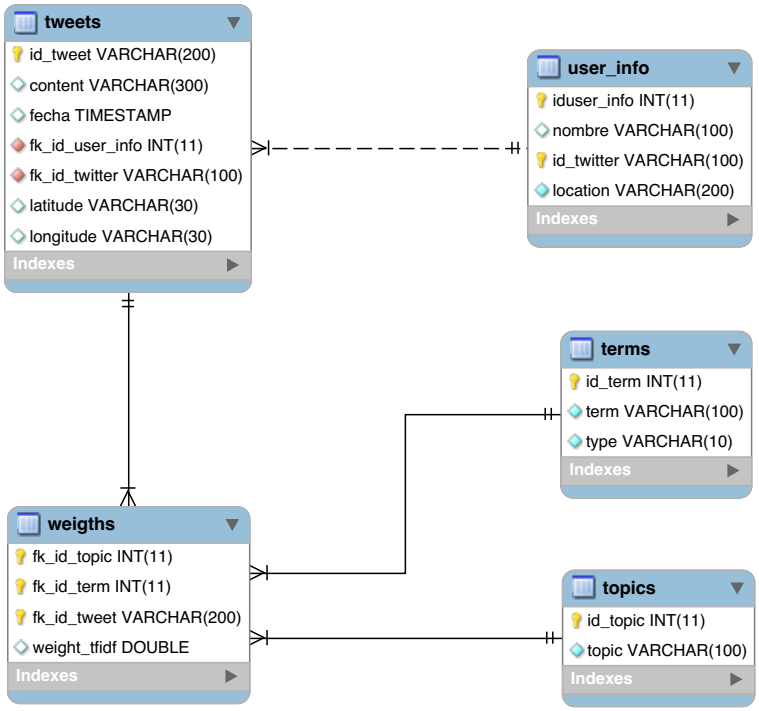

Figura 4.4: Diagrama entidad relación.

4.3. Base de datos

Después de ejecutar las tareas de preprocesamiento sobre el texto de los tweets recopilados, los resultados obtenidos se almacenan en la base de datos. Para llevar a cabo esta tarea, fue necesario diseñar una base de datos cuya estructura permitiera acceder a los datos de forma eficiente. Los principales beneficios de nuestra base de datos son:

- Es posible almacenar los tweets obtenidos haciendo referencia a tópicos específicos. Esto permite ir aumentando el tamaño de los corpus en cualquier momento.
- Ayuda a tener control de los datos que son guardados. Logrando evitar redundancia al momento de obtener los tweets.
- Facilita el acceso a los datos para optimizar los procesos de cada módulo de la herramienta.

En la Figura 4.4 se muestra el modelo entidad relación de nuestra base de datos. Como se puede observar en el diagrama entidad relación, la base de datos está conformada por 5 tablas.

La estructura de cada tabla se describe a continuación.

Tabla de Tweets

Almacena la información de los tweets recolectados. La tabla consta de 7 campos: 'id_tweet' es el identificador que asignamos a cada registro nuevo en la base de datos, con el fin de evitar duplicidad de información, 'content' almacena la cadena de texto de los tweets, 'fecha' contiene el día en que se creo el tweet, 'longitud' y 'latitud' guardan las coordenadas de ubicación en que el usuario se encontraba cuando publicó en Twitter, 'id_twitter' es el identificador único que cada tweet tiene asignado desde el momento en que se crea y finalmente 'id_user_info' se utiliza para ligar la información de la esta tabla con la de usuarios. Los identificadores son datos numéricos, mientras que el texto del tweet es tipo varchar y tiene un límite de 140 caracteres. Ver Tabla 4.1.

Campo	**Descripción**
id_tweet	Identificador único del tweet.
content	Texto del tweet.
fecha	Fecha de creación del tweet.
id_user_info	Identificador único del usuario.
id_twitter	Identificador único del usuario en twitter.
latitud	Latitud en que fue creado el tweet.
longitud	Longitud en que fue creado el tweet.

Tabla 4.1: Descripción de la tabla "tweets".

Tabla de usuarios

Almacena los datos principales de los usuarios. La tabla consta de 4 campos: 'id_user_info' es el identificador único, 'nombre' contiene el nombre de usuario que se tiene en Twitter, 'location' almacena el país de origen del usuario al que corresponde el tweet e 'id_twitter' permite ligar esta tabla con la de tweets. Ver Tabla 4.2.

Campo	**Descripción**
id_user_info	Identificador único del usuario.
nombre	Nombre de la cuenta de usuario.
id_twitter	Identificador único del usuario en twitter.
location	País de origen del usuario.

Tabla 4.2: Descripción de la tabla "user_info".

Tabla de temas

Guarda cada uno de los temas que son buscados con la herramienta de análisis. Esta tabla consta de 2 campos: 'id_topic' es un identificador único para cada entrada y 'topic' contiene la descripción del tema que se busca en la herramienta. La finalidad de esta tabla es que cuando se realicen múltiples búsquedas sobre un mismo tema lo tweets se consideren como elementos extras a cada uno de los corpus ya existentes. Ver Tabla 4.3.

Campo	Descripción
id_topic	Identificador único del tema.
topic	Descripción del tema.

Tabla 4.3: Descripción de la tabla "topics".

Tabla de términos

Contiene los diferentes temas encontrados posterior al análisis realizado a los tweets. Consta de 3 campos: 'id_term' es el identificador único de cada registro, 'term' almacena cada uno de los tokens obtenidos del tweet y 'type' contiene la categoría gramatical correspondiente a cada token. Guardar esta información permite realizar consultas más específicas referentes al contenido de los tweets y que son de utilidad para el análisis de esta red social. Ver Tabla 4.4.

Campo	Descripción
id_term	Identificador único del término.
term	Descripción del término.
type	Categoría gramatical del término.

Tabla 4.4: Descripción de la tabla "terms".

Tabla de pesos

Guarda los pesos asignados, por el algoritmo TF-IDF, a cada uno de los elementos de la tabla de términos. Esta tabla consta de 4 campos: 'id_topic' que es el identificador único de los registros, 'id_term' contiende el identificador del tema correspondiente a dicho termino para poder relacionarlo con la tabla de temas, 'id_tweet' contiene el identificador del tweet asignado en la base de datos y 'weight_tfidf' contiene el valor de cada palabra de tweet obtenido mediante el algoritmo TF-IDF Ver Tabla 4.5.

Campo	Descripción
id_topic	Identificador único del tópico.
id_term	Identificador único del término.
id_tweet	Identificador único del tweet.
weight_tfidf	Peso obtenido por el algoritmo TF-IDF al tópico.

Tabla 4.5: Descripción de la tabla "weights".

4.4. Módulos de Twitter AT

4.4.1. Obtención de datos

El módulo de recolección de datos consiste en un programa que toma como entrada el tópico al que hacen referencia los tweets, así como la configuración de la búsqueda a realizar.

Consulta de tweets

En primer lugar se requiere realizar el proceso de autenticación en el servidor de Twitter. Posteriormente para realizar la obtención de tweets con base en temas, mediante la API REST de Twitter, es necesario configurar los parámetros de consulta.

Twitter AT cuenta con una configuración predeterminada. Sin embargo, el usuario puede modificar la configuración. Los parámetros que es posible variar en el módulo de búsqueda de datos son: tópico, cantidad de tweets, idioma de los tweets y fecha de creación. A continuación se describe cada uno:

- Tópico: Tema que hace referencia a los tweets que se van a obtener.
- Cantidad de tweets: Ya que existe un límite de peticiones cada determinado tiempo, se tienen cifras específicas para el número de tweets a descargar a la vez: 100, 200, 300... 1000.
- Idioma de los tweets: este trabajo se enfoca únicamente al preprocesamiento de textos en inglés.
- Fecha de creación: De acuerdo a las restricciones del API de Twitter, es posible obtener tweets de hasta 7 días previos al día de la petición.

Para llevar a cabo la tarea de obtener tweets nuestro algoritmo de búsqueda considera como entrada 4 parámetros: cantidad, tópico definido por el usuario, idioma y fecha de creación de los tweets. Los últimos 3 permiten reducir el espacio de búsqueda.

La función de búsqueda de tweets es importante, debido a que evita almacenar información duplicada. Para realizar esta función se requiere considerar el identificador de los tweets. Esta característica es un primer filtro para cuando se quiere incrementar el número de tweets para un mismo tópico, evitando así que se guarde más de una vez el mismo elemento.

El algoritmo 2 muestra el pseudocódigo que realiza la función de búsqueda en nuestra herramienta.

Consideraciones

- Los valores permitidos para obtener cierta cantidad de tweets se establecieron para optimizar el tiempo de ejecución. El número de tweets óptimo esta entre 100 y 1000. Hay ocasiones en que no hay elementos nuevos, y realizar menos peticiones a Twitter es importante para no aumentar el tiempo de búsqueda.

Algoritmo 2 Obtención de tweets

Entrada: Cantidad, tópico, idioma y fecha de creación de los tweets.
Salida: Lista de tweets que coinciden con los parámetros de búsqueda.

1: Crear un objeto *twitterQuery* para enviar las peticiones a Twitter.
2: Crear una lista *tweetsEncontrados* para almacenar los tweets obtenidos.
3: Crear una variable numérica $faltan = 0$, para almacenar el número de tweets faltantes.
4: Establecer el valor de la propiedad *lang* del objeto *twitterQuery* al valor de entrada *idioma*.
5: Establecer el valor de la propiedad *since* del objeto *twitterQuery* al valor de entrada $fecha$.
6: **Mientras** $n < numeroTweetsBuscados$ **Hacer**
7: Asignar a $faltan$ el valor de $numeroTweetsBuscados$ menos el valor de la propiedad $size$ de la lista *tweetsEncontrados*.
8: **Si** $faltan < 100$ **Entonces**
9: Establecer la propiedad *count* del objeto *twitterQuery* al valor almacenado en $faltan$.
10: **Sino**
11: Establecer la propiedad *count* del objeto *twitterQuery* en 100.
12: **Fin Si**
13: Enviar una petición segura y autorizada a Twitter mediante el método *search* del objeto *twitterQuery* y almacenar el resultado en *QueryResult*.
14: **Para todo** *tweet* encontrado *QueryResult* **Hacer**
15: **Si** *tweetId* != al valor de la propiedad *id* del objeto *tweet* **Entonces**
16: Agregar el objeto *tweet* en la lista *tweetsEncontrados*.
17: **Fin Si**
18: **Fin Para**
19: **Fin Mientras**
20: **Retornar** *tweetsEncontrados*

- Hay situaciones en que la cantidad de tweets solicitados es mayor a la que se logra obtener. En esos casos, se indica al usuario el número de elementos que fue posible encontrar.

- Una vez que se tiene la lista de tweets encontrados, éstos son almacenados de forma automática en nuestra base de datos. Estos datos son indispensables para posteriormente comenzar la etapa de preprocesamiento.

Al hacer una petición de nuestra búsqueda al servidor, éste regresa una respuesta en formato JSON, del cual con la ayuda de la librería twitter4j se pueden obtener datos específicos de todo el conjunto de tweets. Los datos que nuestra herramienta almacena son:

- ID de tweet: Identificador único del tweet.

- Texto del tweet: Es cada uno de los mensajes de un máximo de 140 caracteres que se pueden enviar a través del servicio de Twitter.

- ID de usuario: Identificador único del usuario que twitteó.

- Fecha: Fecha en que fue publicado el tweet.

- Idioma: Idioma que el usuario establece al crear una cuenta en Twitter.

4.4.2. Preprocesamiento de tweets

La etapa de preprocesamiento de tweets que se propone en este trabajo consta de dos partes, el proceso de limpieza y el proceso de lematización y etiquetación.

Limpieza de tweets

Para poder realizar la extracción de términos relacionados con un tópico en particular, en primer lugar es necesario hacer una limpieza a los tweets obtenidos. Para llevar a cabo esta tarea se utilizan como base las bibliotecas de NLP Stanford, las cuales permiten realizar múltiples operaciones en textos. El programa desarrollado permite realizar una serie de pasos los cuales se muestran en la Figura 4.5.

- **Quitar símbolos y caracteres especiales.**
  Para quitar los caracteres especiales se realizó una expresión regular que elimina cadenas de un carácter y cadenas que contengan símbolos especiales

- **Eliminar entidades del texto de un tweet.**
  Una de las bibliotecas de NLP Stanford está diseñada especialmente para trabajar con

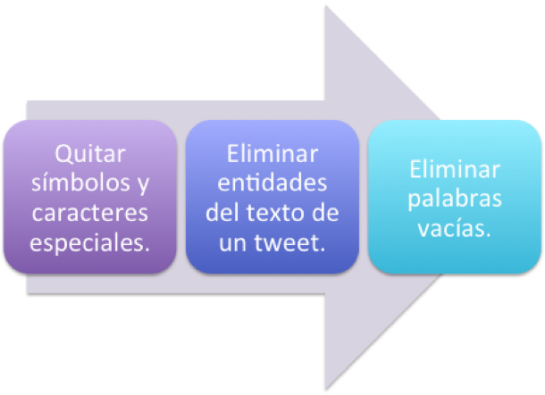

Figura 4.5: Proceso de limpieza de tweets.

librería de twitter de java, por lo que es fácil identificar elementos básicos de un tweet. En este caso, las entidades que se eliminan son URL's y nombres de usuario.

- **Eliminar palabras vacías (stopwords): preposiciones, conjunciones y disyunciones, palabras utilizadas con mucha frecuencia.**
  NLP Stanford también cuenta con una lista de palabras vacías predefinida. Una vez que se llega a esta parte de la limpieza del tweet se realiza una búsqueda de las palabras vacías contenidas en el texto del tweet y se eliminan de la cadena.

A continuación se muestra el flujo que sigue la función encargada de la limpieza de tweets:

- Eliminar los caracteres especiales. Para eliminarlos se utiliza la siguiente expresión regular: ("[^a-zA-Z0-9\\s/:@]", " "), la cual indica que cualquier símbolo que no se encuentre dentro del grupo definido se elimine.
- Eliminar los espacios en blanco.
- Eliminar las palabras vacías (stopwords), por ejemplo preposiciones, conjunciones y disyunciones, las cuales son utilizadas con mucha frecuencia.

Etiquetación y Lematización

Una vez que al conjunto de tweets ha pasado por el proceso de limpieza se procede a la tokenización de la cadena, se continua con la etiquetación de cada una de las palabras, y finalmente se aplica el algoritmo TFIDF para así identificar los términos claves referentes al tópico de análisis.

A continuación se muestra el flujo que sigue la función encargada de la etiquetación y lematización de tweets:

- Se etiquetan las palabras de acuerdo a la categoría correspondiente: adjetivos, verbos, nombres propios.
- Se hace un recorrido de los arreglos que contienen las palabras etiquetadas para eliminar las palabras con etiquetas de Twitter, como URL's y nombres de usuarios.
- Se tokenizan los tweets para obtener una lista de palabras correspondiente a cada uno.
- Se hace un recorrido de la lista de palabras de cada tweet para lematizar las palabras en caso de ser posible.

Asignación de etiquetas

Para etiquetar el contenido de cada tweet se utiliza la librería POS Tagger de Stanford NLP, que permite identificar cada entidad del texto en cuestión. Las principales categorías que se obtienen a partir de la etiquetación son:

- Adjetivos
- Verbos (identificando el tiempo de conjugación)
- Nombres propios

En la tabla 4.6 se muestra una lista de algunas de las etiquetas incluidas en la librería POS Tagger utilizada.

Lematización

Para lograr reducir el número de elementos de un tweet es necesario aplicar la tarea de lematización en el texto.

Dada la cadena de texto correspondiente a cada tweet del conjunto de datos, la lematización permite obtener la raíz de cada palabra. El objetivo es identificar palabras similares y así evitar redundancia en la etapa de análisis.

Por último *Twitter AT* necesita aplicar el método de tokenización a cada uno de los tweets antes de guardar el resultado de la etapa de preprocesamiento en la base de datos.

4.4.3. Transformación de datos

Twitter AT cuenta con la implementación del algoritmo TF-IDF en la etapa de transformación de datos. Se eligió este algoritmo para realizar una comparación entre el funcionamiento de *Twitter AT* y una herramienta ya existente. A continuación se describe el funcionamiento de este algoritmo.

Etiqueta(Tag)	**Descripción**
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective comparative
JJS	Adjective superlative
LS	List ítem marker
MD	Modal
NN	Noun singular or mass
NNS	Noun plural
NNP	Proper noun singular
NNPS	Proper noun plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

Tabla 4.6: Lista de etiquetas que utiliza "POS Tagger"

Algoritmo TF-IDF

Uno de los métodos empleados para la creación de listas de palabras clave para los sistemas de búsqueda de información es el algoritmo TF–IDF (Term Frecuency – Inverse Document Frecuency). Este algoritmo genera listas de palabras clave con un peso que indica qué tan relevante es una palabra con respecto al documento seleccionado y al corpus en general.

Por lo tanto el algoritmo TF–IDF puede ser empleado para determinar la frecuencia de ocurrencia de cada término en la colección de tweets recuperados a partir de una palabra clave. Por medio de este algoritmo se determina qué tan relevantes son los términos relacionados con la palabra clave, estableciendo un ranking entre los mismos.

El coeficiente TF–IDF es una ponderación usada a menudo en tareas de recuperación de información y minería de texto. El coeficiente es una medida estadística usada para evaluar qué tan importante es una palabra respecto a un documento perteneciente a una colección de documentos [36]. En nuestro caso, los tweets son considerados como los documentos.

La frecuencia de un término (TF) en un documento dado es el número de veces que el término aparece en ese documento. Este valor es usualmente normalizado para evitar que documentos extensos tengan ventaja. De esta forma, la importancia del término t_i en el documento d_j está dada por:

$$TF_i, j = \frac{n_{i,j}}{\sum_K n_{k,j}},$$

donde $n_{i,j}$ es el número de ocurrencias de término considerado en el documento d_j, y el denominador es el número de ocurrencias de todos los términos en el documento d_j.

La frecuencia inversa de los documentos (IDF) es una medida de la importancia general del término y se calcula mediante:

$$IDF_i = \frac{log\,|D|}{|dj : t_i \in d_j|},$$

donde el numerador es el número total de documentos en el cuerpo y el denominador es el número de documentos donde el término t_i aparece (i.e., $n_{i,j} \neq 0$). Así, el coeficiente TFIDF para el término t_i en el documento d_j es:

$$TFIDF_{i,j} = TF_{i,j} \times IDF_i.$$

Un valor TFIDF alto es alcanzado por un término con alta frecuencia en el documento considerado, pero baja frecuencia en la colección total de documentos. De esta manera, el coeficiente tiende a filtrar términos comunes.

Finalmente las palabras que tienen un valor TFIDF alto son las que tienen alta frecuencia en un determinado tweet pero baja frecuencia en la colección de total de tweets, dejando

con cifras menores a palabras que son mencionadas en todo el corpus, y por la cual éstas no dan información importante en este análisis.

TF-IDF se utiliza tradicionalmente en la recuperación de documentos, pero actualmente se realizan estudios de su eficiencia en aplicaciones de selección de características, las cuales demuestran que las características ponderadas resultantes muestran un mejor rendimiento que cuando se utiliza por ejemplo una máquina de soporte vectorial (SVM).

Así, una vez que se concluye la fase de preprocesamiento de los tweets se realiza una consulta de los términos obtenidos con base en el peso de cada palabra, para poder visualizar los términos que representan un tema en particular.

En la base de datos existe una tabla llamada pesos, la cual guarda el resultado obtenido por el algoritmo TFIDF en el conjunto de tweets. Al realizar una consulta en nuestra base de datos, es posible obtener una lista en orden de relevancia de los términos que están relacionados con el tópico en cuestión.

4.4.4. Diagrama general de la herramienta

En general, la implementación de *Twitter AT* consta de dos funciones principales. La búsqueda de tweets y el preprocesamiento de los mismos. Estas dos funciones permiten contar con una mejor estructura que facilita el uso de los datos para implementar distintos métodos de análisis.

Como se indicó previamente, el propósito de este trabajo de tesis es desarrollar una herramienta para el preprocesamiento de redes sociales, particularmente Twitter. A continuación se describe el procedimiento que lleva nuestra herramienta para alcanzar el objetivo planteado.

En primera instancia, se obtienen conjuntos de tweets con base en búsqueda por tópico para poder conformar la entrada que va a ser considerada por el módulo de preprocesamiento. Una vez que se tienen almacenados los tweets, es posible comenzar la etapa de preprocesamiento. Las tareas principales que se realizan en este módulo son: eliminar símbolos especiales, eliminar palabras vacías, etiquetar y lematizar el texto. Cada que se realizan estas tareas, el resultado es guardado en la base de datos. Finalmente, la estructura de los datos almacenados brinda una base para poder implementar algoritmos que permitan el análisis de redes sociales. Como ejemplo, en este trabajo se consideró la implementación del algoritmo TF-IDF.

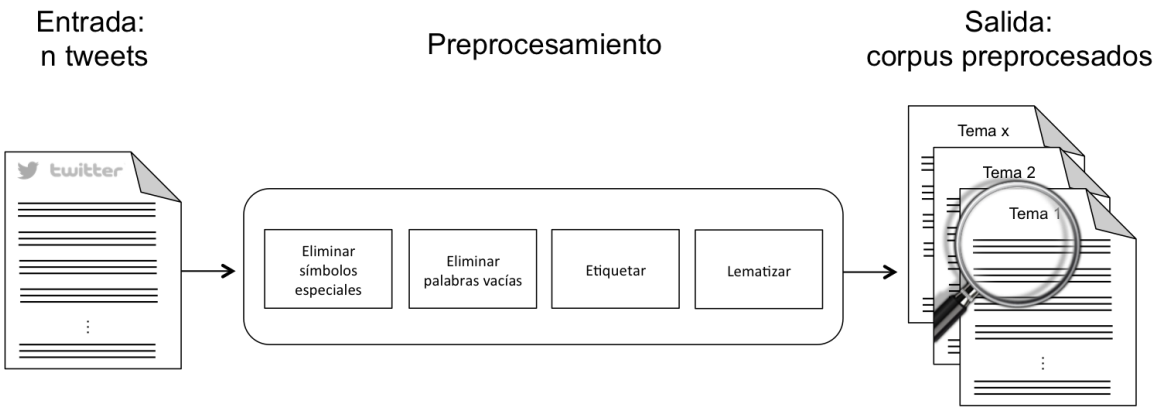

Figura 4.6: Diagrama general de la herramienta.

Capítulo 5

Implementación

Este capítulo describe el funcionamiento de *Twitter AT*. Además muestra los resultados obtenidos en la implementación de dos casos de análisis de un conjunto de tweets relacionados con un tema específico. Posteriormente se realiza una comparación de *Twitter AT* con las herramientas ya existentes.

5.1. Interfaz de usuario

La interfaz de usuario de *Twitter AT* permite realizar la búsqueda de tweets de acuerdo a un tema en particular, así como las tareas del módulo de preprocesamiento de datos.

En la Figura 5.1 se muestra la interfaz general que tiene *Twitter AT*. A continuación se describen los elementos principales que la conforman:

1. En el cuadro de texto "Topic" se introduce el tópico que se va a considerar para la creación del corpus.

2. El botón "Search" permite iniciar la función de búsqueda que la herramienta realiza de acuerdo a la configuración establecida. Los tweets que se logran obtener se muestran en una tabla.

3. El icono del engrane permite hacer modificaciones en los parámetros predefinidos, por ejemplo el número de tweets que se va a obtener.

4. La columna "Id" es un número consecutivo de los registros, lo cual permite identificar la cantidad de tweets que se encontró.

5. La columna "User" muestra el nombre del usuario al que pertenece cada tweet.

6. La columna "Text" contiene el texto original de cada uno de los tweets obtenidos.

7. La columna "Created at" indica la fecha en que el tweet fue publicado.

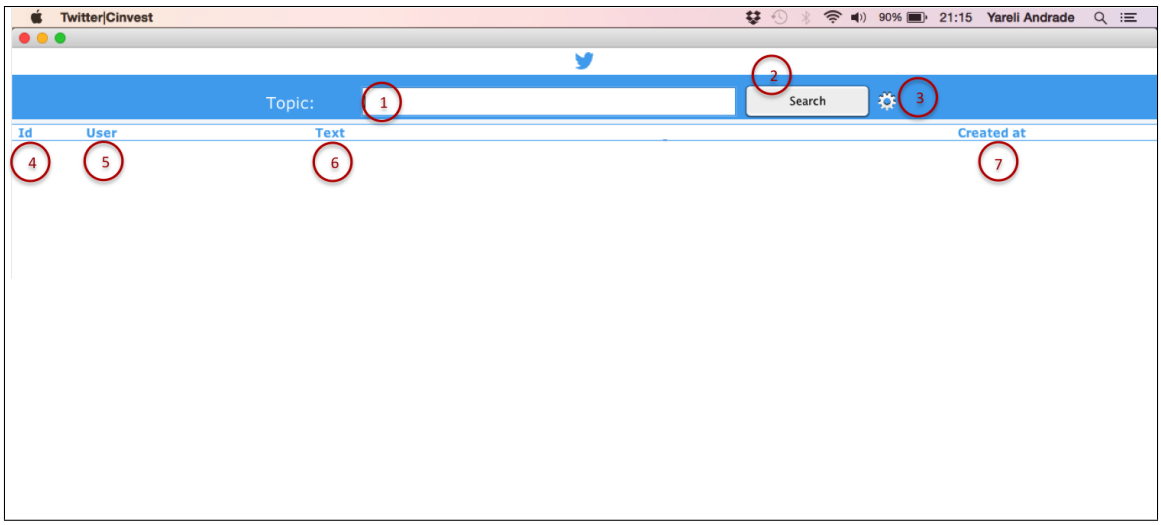

Figura 5.1: Interfaz general de la herramienta.

Configuración de búsqueda

En primera instancia, *Twitter AT* cuenta con un apartado para especificar el tema que se va a tomar en cuenta para hacer la petición a la API de Twitter.

En la Figura 5.2 se puede observar la vista que corresponde a la configuración de la herramienta.

A continuación se indican los parámetros que pueden ser modificados en la configuración de *Twitter AT*:

1. Numero de tweets: consta de un combo que abarca un rango de 100 hasta 10000, considerando múltiplos de 100. Este dato indica la cantidad de tweets que van a obtenerse.

2. Lenguaje: este dato específica el lenguaje que deben tener los tweets que se van a obtener.

3. Fecha: indica de qué fecha se requiere que sean los tweets que se están solicitando. Se tiene un rango establecido, el cual va de 7 días anteriores con límite de la fecha actual. El valor asignado por default es el día en que se realiza la búsqueda. Ver Figura 5.3.

5.2. Obtención de datos

Para la etpa de pruebas s coleccionaron datos acerca de cinco temas distintos. A continuación se listan los detalles de esta etapa.

- **Temas:** *Immigration Reform*, *Ebola*, *iPhone*, *Barack Obama* y *Mexico*

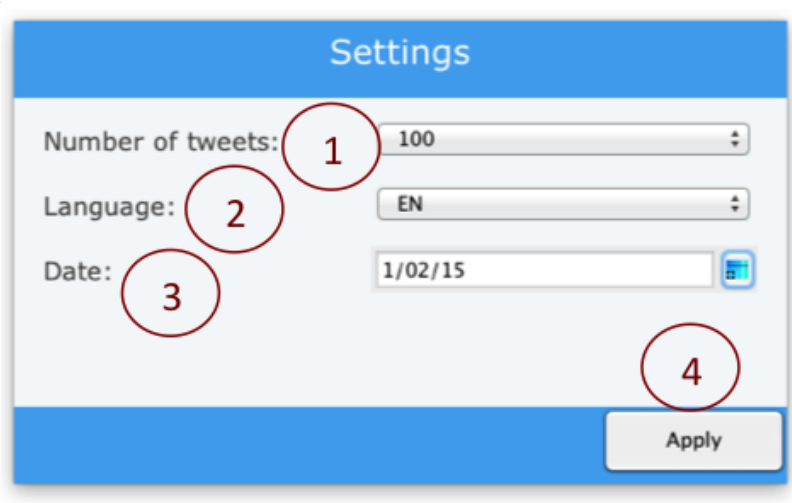

Figura 5.2: Configuración de parámetros de búsqueda.

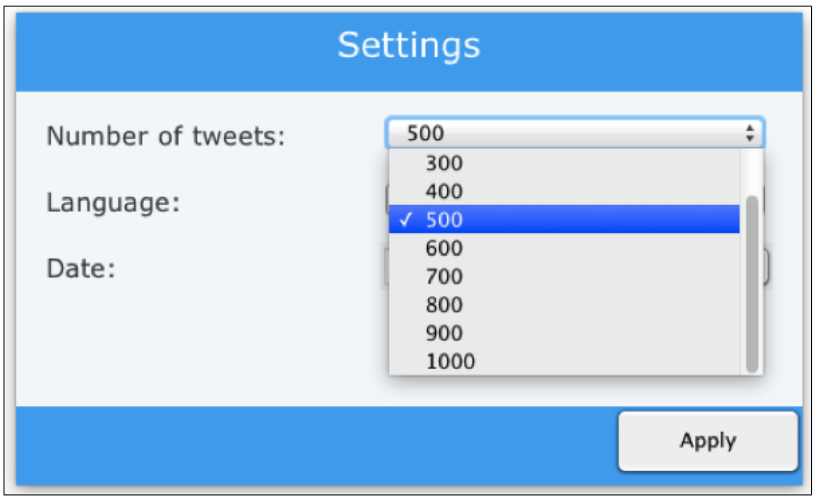

Figura 5.3: Configuración de búsqueda.

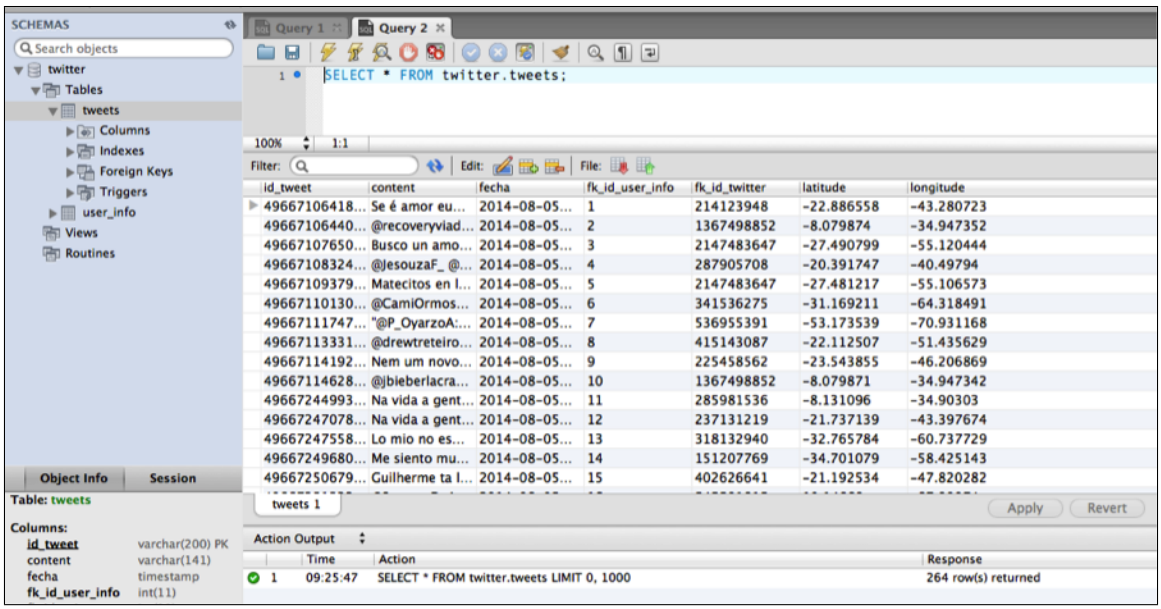

Figura 5.4: Ejemplo de datos almacenados en nuestra BD.

- **Tiempo de colección de datos:**
  un mes - 1000 cada 4 días para cada tópico

- **Cantidad de datos:** 8000 tweets por tema

Una vez que se ha realizado la consulta en Twitter, los datos son almacenados de forma automática en una base de datos (ver Figura 5.4), para posteriormente poder preprocesar dicha información.

En la Figura 5.5 se muestra un ejemplo de los datos que la interfaz de usuario muestra después de realizar el preprocesamiento, a partir de un tema propuesto por el usuario.

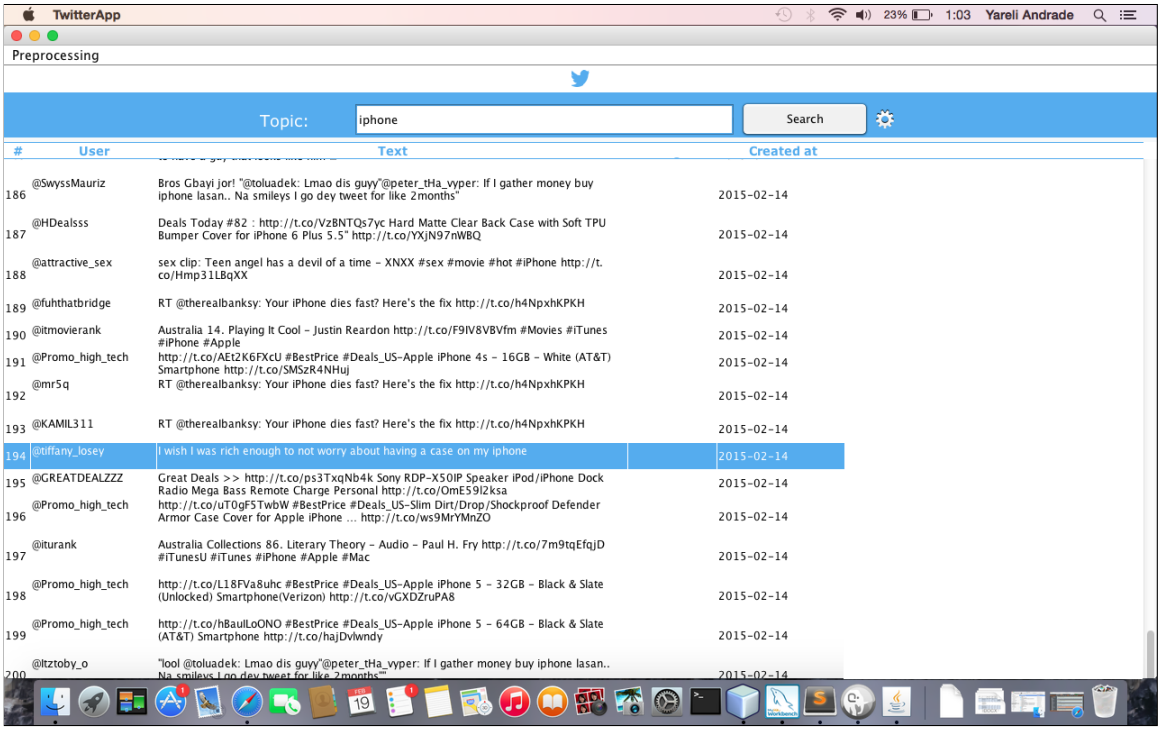

Figura 5.5: Resultados obtenidos en la interfaz de usuario.

5.3. Preprocesamiento de tweets

Como ya se ha descrito en el capítulo anterior, lo primero que hace *Twitter AT* es obtener un conjunto de datos. Es importante señalar que los tweets se obtienen en formato JSON, el cual contiene demasiada información relacionada con cada tweet. Entones es necesario de todo ese documento, sólo extraer los elementos que son de interés para nuestro sistema. En la Figura 5.6 se puede observar un ejemplo de la respuesta que el servidor de Twitter regresa al hacer una petición.

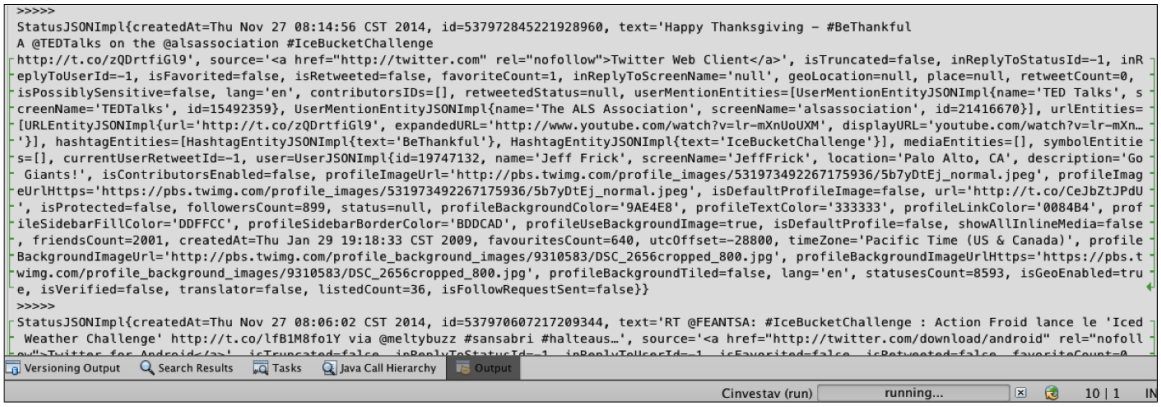

Figura 5.6: Respuesta dada al hacer una petición mediante la API de Twitter.

Una vez que *Twitter AT* hace las peticiones al servidor de Twitter, se obtienen varios datos para cada tweet. El dato principal es el texto. En la Figura 5.7 se muestra el ejemplo de un tweet referente al tema 'ebola'.

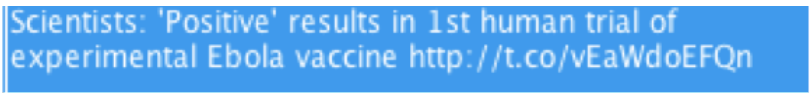

Figura 5.7: Tweet original.

Ya que se tiene el conjunto de tweets, comienza el preprocesamiento de los textos, que consiste en homogeneizar cada uno de los elementos de la colección de datos con la que se va a trabajar. Esto implica eliminar caracteres especiales, cantidades numéricas, palabras vacías (stopwords) y espacios en blanco. En la Figura 5.8 se visualiza la salida que da *Twitter AT* después de estas tareas de preprocesamiento.

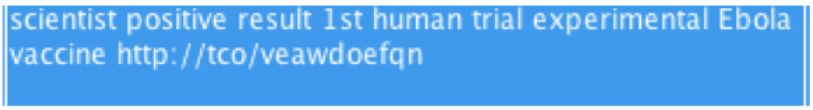

Figura 5.8: Tweet después del proceso de limpieza.

Posteriormente *Twitter AT* realiza el proceso de asignar etiquetas a cada uno de los tokens del tweet. Se eliminan determinadas clases, como URLS y nombres de usuario. En la Figura 5.9 se puede observar el resultado para un tweet posterior a esta etapa de preprocesamiento.

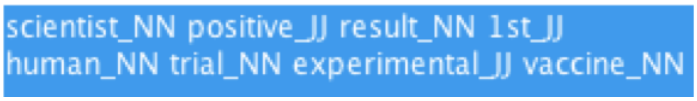

Figura 5.9: Términos clave con su categoría gramatical.

Finalmente, una vez que ya se tienen los tweets preprocesados, continua la implementación del algoritmo TF-IDF para la asignación de pesos a cada una de las palabras. Al ejecutarse el algoritmo se contemplan los elementos que se tienen hasta el momento en el conjunto de datos.

5.4. Transformación de datos

Para poder realizar pruebas con los resultados obtenidos en la etapa de preprocesamiento, se consideraron dos conjuntos de datos para calcular los pesos de palabras con el algoritmo TF-IDF. A continuación se describen los resultados obtenidos en los experimentos que se realizaron.

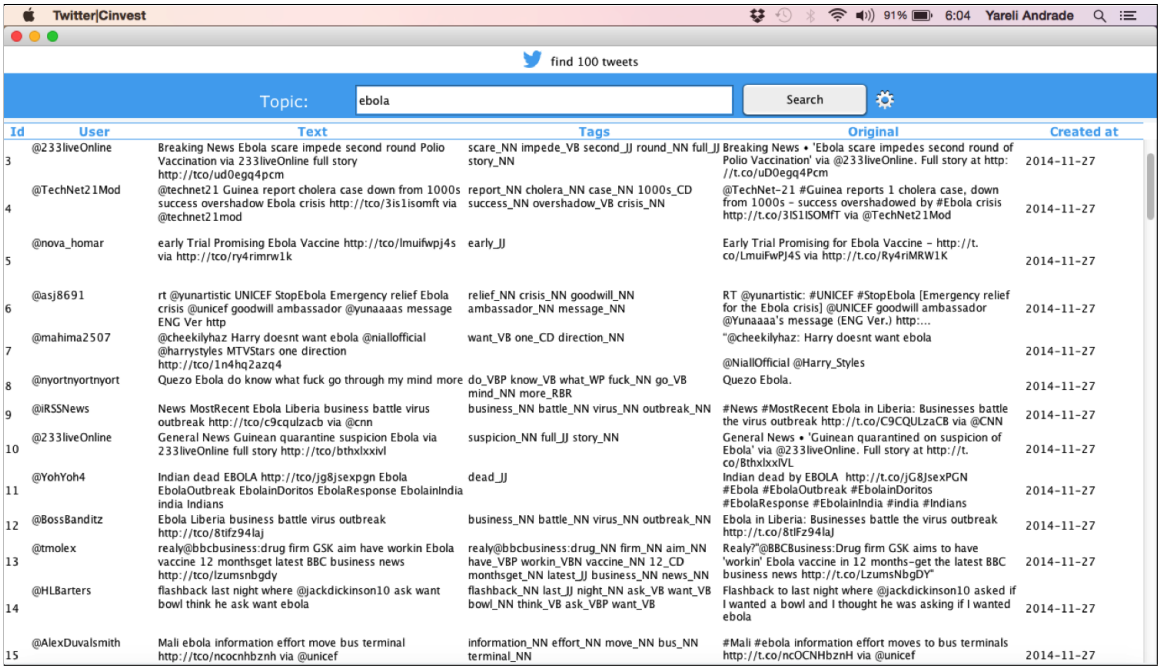

Figura 5.10: Búsqueda de tweets relacionados con el tema "Ebola".

Tópico: Ebola

El primer tópico seleccionado fue el tema del Ebola, un virus que causa una enfermedad en el ser humano, y el cual fue "trend topic" en Twitter. En la Figura 5.10 se muestra parte de los resultados obtenidos al realizar la búsqueda de tweets con el tópico definido.

Como ya se mencionó en el capítulo anterior, una vez que se lleva a cabo el preprocesamiento de los datos obtenidos, estos datos son guardados en la base de datos. Con la finalidad de tener una mejor visibilidad de los resultados obtenidos, después de realizar la consulta y obtener los términos más importantes, se decidió representarlos mediante una gráfica. Puede observarse que en la gráfica también se encuentran los porcentajes correspondientes a cada término, considerando como 100 % la totalidad de términos que se encuentran relacionados con el tópico en la base de datos.

Para este caso en particular, la Figura 5.11 muestra la gráfica de los resultados para el tema "Ebola". Considerando la información obtenida a partir de *Twitter AT*, se puede observar que los términos de mayor interés relacionadas con este tema son "die" y "doctor".

Al realizar el experimento con este tema y realizar una consulta en la base de datos se obtienen los resultados mostrados en la Tabla 5.1 (se encuentran los 8 términos con mayor peso). Los términos obtenidos se muestran en orden de importancia según el algoritmo TF-IDF.

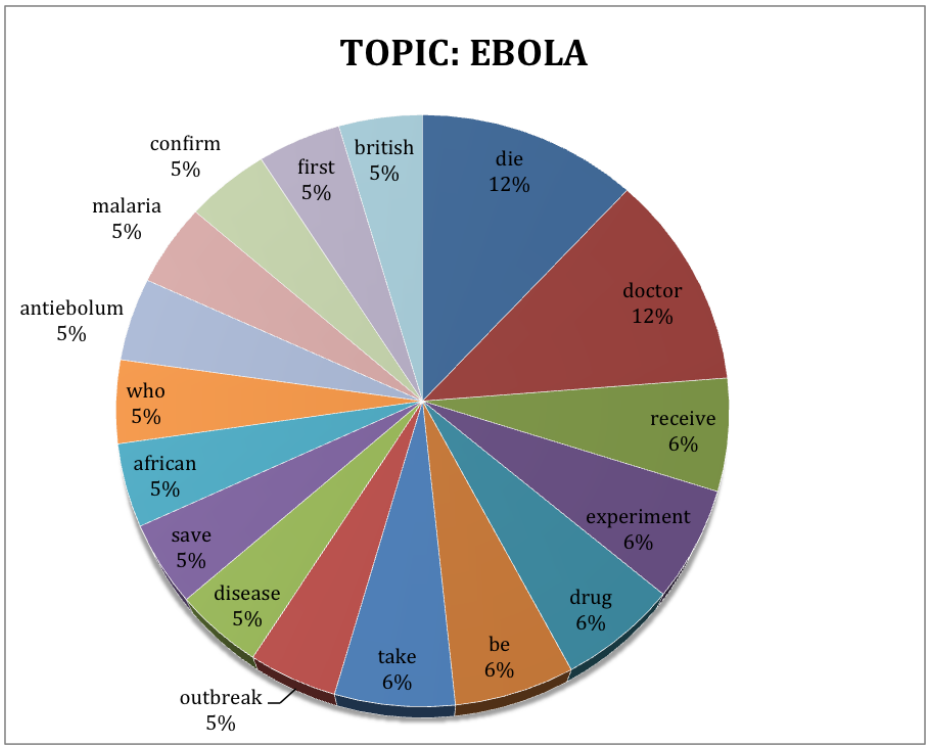

Figura 5.11: Extracción de datos obtenidos para el tema 'ebola'.

Término	**Valor TF-IDF**
die	0.761916231
doctor	0.750077117
receive	0.391415687
drug	0.391415687
take	0.391415687
experiment	0.391415687
do	0.391415687
outbrake	0.289711998

Tabla 5.1: Lista de términos obtenidos para el tema 'Ebola'.

Para lograr ver los resultados obtenidos en la etapa de preprocesamiento se realizó el mismo experimento con la herramienta Twitter Arquivist. Esta herramienta permite ver sus resultados en una imagen que muestra las palabras en distintos tamaños de acuerdo a su importancia. En la Figura 5.12 se puede observar el resultado obtenido.

Haciendo una comparación con base en los resultados de ambas herramientas, se pueden destacar los siguientes puntos:

- La presencia de datos en más de un idioma en los resultados de Twitter Arquivist. Esta característica que es considerada en *Twitter AT* permite definir mejor el conjunto

Figura 5.12: Resultados obtenidos para el tema 'Ebola' con Twitter Arquivist.

de datos que se quiere considerar.

- En los resultados también se encuentran palabras como 'así' y 'con' que son conectores que no representan información relevante. Este tipo de elementos se eliminan en nuestra etapa de preprocesamiento.
- El considerar URL's por si solas tampoco representa información que sea de utilidad, motivo por el que no son consideradas en nuestro análisis.

Tópico: Immigration Reform

Para este segundo experimento el procedimiento es el mismo que con el primer tópico. La cantidad de tweets que fueron coleccionados para el tema fueron un total de 8000 elementos. Los resultados obtenidos al aplicar el algoritmo TF-IDF con este tema se muestran en la Figura 5.13.

Los términos más sobresalientes obtenidos al realizar el análisis de Immigration Reform se mencionan en la Tabla 5.2 en orden de importancia.

Término	**Valor TF-IDF**
need	0.262574756
vote	0.262574756
immigrant	0.246087196
say	0.218949643
support	0.214392853
illegal	0.212903407
powerful	0.196056535
foe	0.19214526

Tabla 5.2: Lista de términos obtenidos para el tema 'Immigration Reform'.

Al igual que con el primer tópico, también se realizó el mismo experimento con la herramienta Twitter Arquivist. En la Figura 5.14 se puede observar el resultado obtenido.

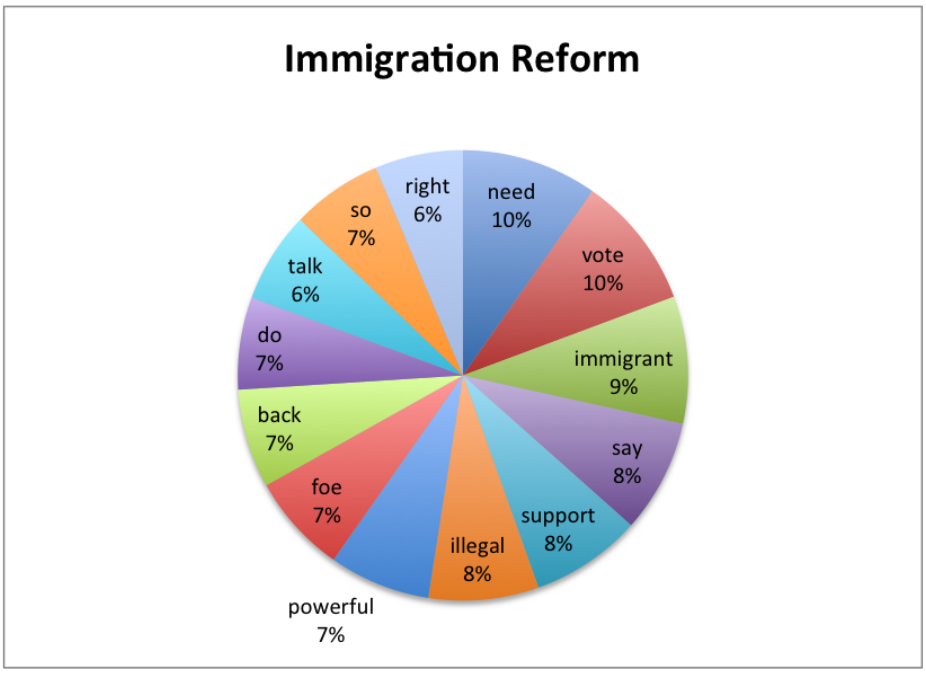

Figura 5.13: Extracción de datos obtenidos para el tema 'Immigration Reform'.

Figura 5.14: Resultados obtenidos para el tema 'Immigration Reform' con Twitter Arquivist.

Haciendo una comparación nuevamente es posible notar la presencia de URL's. Analizando los resultados de Twitter Arquivist también puede notarse la importancia del proceso de lematización que *Twitter AT* incluye. Por ejemplo, considerar palabras en plural y singular puede causar duplicidad de información si ambas existen en el conjunto de datos que se analiza.

5.5. Discusión

Twitter AT tiene como propósito implementar la etapa de preprocesamiento para datos de Twitter con base en temas. La finalidad es contar con datos que mejoren los resultados obtenidos en la etapa de análisis. En este caso, poder analizar tópicos que son de interés en determinado tiempo para un gran número de usuarios de Twitter.

Aunque este tema ha sido abarcado en múltiples investigaciones, una de las problemáticas existentes es la falta de conjuntos de datos, particularmente si se quiere analizar un área en particular. Parte de de nuestro trabajo se enfocó en la solución de este problema. Se planteó el desarrollo de una herramienta que permita obtener tweets relacionados con temas que son de interés social en el momento. Además se consideró la idea de tener una representación más completa de los datos, por lo que no sólo se guardan los textos, sino también información como la ubicación y el idioma, lo cual es necesario para poder analizar grandes cantidades de información y tener representaciones con mayor detalle.

A la fecha esta presente el problema de la cantidad de datos que se requiere procesar para llevar a cabo análisis de redes sociales en línea. Por tal motivo, el módulo principal de *Twitter AT* se encarga de la reducción de datos mediante el preprocesamiento de datos con base en temas. El enfoque considerado permite eliminar información innecesaria y facilitar la etapa de análisis.

Como se puede apreciar en los experimentos presentados, *Twitter AT* permite obtener una mejor representación del contenido de los tweets, que en este caso son las palabras clave. La mejora de los resultados es debido a la etapa de preprocesamiento de Twitter AT. Cabe destacar que el tiempo de análisis también se ve beneficiado gracias a los datos obtenidos en el módulo de preprocesamiento. La estructura de los datos es lo que facilita el análisis de los tweets tomando como referencia las temáticas. Además es posible identificar los tweets como parte de un corpus específico.

Finalmente en la Tabla 5.15 se hace una comparación de las características que tiene Twitter AT (Herramienta desarrollada) y las de herramientas ya existentes. Para este análisis se consideraron dos aplicaciones, una gratuita (SONDY) y una comercial (Twitter Arquivist). Los características que se consideran son: antigüedad de tweets a obtener, pará-

Herramienta	Twitter Arquivist	SONDY	Twitter AT
Tipo	Comercial	Libre	Libre
Antigüedad de tweets a obtener	7 días	Base de datos proporcionada por el usuario	7 días
Parámetro de búsqueda	Por usuario o hashtag	No se tiene	Temas específicos
Preprocesamiento	No se menciona en su página	Eliminación de palabras vacías y lematización	Eliminación de símbolos especiales, palabras vacías, entidades de Twitter (URLS, nombres de usuarios); etiquetación y lematización.
Resultados	Palabras más relevantes (reportes estadísticos)	Detección y visualización de eventos	Términos relacionados con el tópico dado
Idioma	No se contempla un filtro	Inglés	Inglés
Almacenamiento de datos	Sí	No	Sí
Cantidad de tweets considerada	Total de tweets almacenados	Los que proporcione el usuario	Total de tweets almacenados

Figura 5.15: Comparación entre la herramienta desarrollada y dos existentes.

metro de búsqueda, preprocesamiento, resultados, idioma, cantidad de tweets considerada y almacenamiento de datos.

Respecto a la antigüedad de tweets que es posible obtener, tanto Twitter Arquivist como *Twitter AT* tienen como límite 7 días. Mientras que SONDY no establece una periodo, ya que depende de la base de datos que el usuario proporcioné. El parámetro de búsqueda que se considera difiere para las 3 herramientas, SONDY no cuenta con un módulo de colección de datos que defina dicha entrada, Twitter Arquivist considera dos parámetros: usuario y hashtag, y *Twitter AT* realiza la búsqueda de acuerdo a temas específicos; lo que hace que nuestra herramienta sea la que considera el espacio de búsqueda más amplio. En cuanto al preprocesamiento, las fuentes de Twitter Arquvist no mencionan las características del mismo, SONDY sólo contempla eliminación de palabras vacías y la lematización; *Twitter AT* incluye eliminación de símbolos especiales, palabras vacías, entidades de Twitter, proceso de etiquetación y lematización, lo que permite obtener mejores resultados en la etapa de análisis. Los resultados finales que da cada herramienta varían ya que son muchos los tipos de análisis que pueden efectuarse con datos de redes sociales en línea. En idioma, *Twitter AT* y SONDY consideran el inglés; sin embargo Twitter Arquivist no contempla ningún filtro, ya que los tweets que obtiene se encuentran en múltiples idiomas, lo que genera ruido en al llevar a cabo la etapa de análisis. Respecto a la cantidad de tweets a considerar no hay un límite definido en ninguna de las tres herramientas, aunque el tiempo de respuesta sin duda alguna varía

de acuerdo al cantidad de tweets que se tenga. Por último, en relación al almacenamiento de datos, SONDY no cuenta con esta funcionalidad, Twitter Arquivist almacena únicamente el texto de los tweets, mientras que *Twitter AT* cuenta con una base de datos que permite acceder a los tweets de una forma más estructurada.

Capítulo 6

Conclusiones y trabajo futuro

En este capítulo se presentan las conclusiones a las que se llegaron con el desarrollo de esta tesis. De igual forma se mencionan posibles mejoras y las perspectivas de *Twitter AT* (herramienta que se implementó).

6.1. Conclusiones

El trabajo presentado en esta tesis es una contribución importante en el campo de estudio de las redes sociales, particularmente el problema que existe con la cantidad de datos que es necesaria analizar. Para obtener una mejora en la etapa de análisis de datos, *Twitter AT* cuenta con un módulo que se encarga del preprocesamiento de los tweets. Tarea que permite eliminar elementos no necesarios en los textos recopilados. Este módulo logra reducir el conjunto de datos que se tiene, lo que inevitablemente mejora y facilita la etapa de análisis.

También se abarca parte del tema de colección de datos, el cual es importante considerar para poder realizar análisis de opinión de temas que son mencionados por los usuarios sólo en un periodo de tiempo específico. *Twitter AT* puede ser de utilidad para investigaciones que estén relacionadas con las diferentes vertientes de minería de textos, ya que puede ayudar a obtener conjuntos de datos relacionados con un tema en particular.

Para complementar la funcionalidad de *Twitter AT* también se consideró la implementación del algoritmo TF-IDF. Este algoritmo permite identificar términos relacionados con un tópico específico, lo que permite de cierta forma conocer el impacto de dicho tema en esta red social. Es importante mencionar que el haber utilizado la API REST para el desarrollo de *Twitter AT* permite al usuario poder analizar temas que son noticia en una región específica y en determinado tiempo.

Actualmente *Twitter AT* es una primera versión que además de obtener datos de manera

inmediata de Twitter, cuenta con una base de datos en la que se almacena información que en un futuro se puede utilizar para realizar diversos estudios y así, mediante técnicas de minería de datos, identificar distintos grupos de usuarios y analizar las características de cada uno.

6.2. Trabajo futuro

La herramienta implementada en esta investigación (*Twitter AT*) se enfoca en la recopilación y preprocesamiento de datos en Twitter, lo que da oportunidad a múltiples opciones que pueden ser contempladas como trabajo futuro.

En primera instancia pueden incrementarse la cantidad elementos a configurar antes de realizar la búsqueda de datos. Por ejemplo, buscar tweets de usuarios específicos. También sería interesante tomar en cuenta los vínculos existentes entre usuarios para así abarcar el tema de similitud de intereses.

Respecto al idioma, cabe mencionar que *Twitter AT* también podría considerar el idioma español. Sin embargo, debe tenerse en cuenta que se requieren de distintos algoritmos para la etapa de preprocesamiento de los tweets, ya que este idioma es más complejo.
Otra posibilidad es tener más redes sociales como fuente de información, una propuesta sería integrar Facebook, ya que la mayoría de los usuarios cuentan con perfiles en estas dos redes sociales.

Referente a la interfaz de usuario, se puede desarrollar un módulo en el que un mapa muestre la diferencia de opinión respecto a un tema en particular de acuerdo a la ubicación de los usuarios. Por ejemplo, se podría saber lo que la gente piensa a cerca de un tema específico, una marca o una persona pública dependiendo de la ubicación de los usuarios que comparten los tweets.

Apéndice A

Tecnologías empleadas

La herramienta de extracción y análisis de datos de Twitter implementada requirió de diversas tecnologías para su desarrollo. En este anexo se incluye la descripción de la funcionalidad de cada una.

A.1. Twitter API

Actualmente existen 3 diferentes APIS proporcionadas por twitter para poder interactuar con los datos de esta red social, en la Figura A.1 se pueden observar las diferencias entre sí.

SEARCH API:
La API Search de Twitter es parte de la API REST v1.1. Permite consultas en los índices de tweets recientes o populares, pero la función de búsqueda disponible en móviles o web clientes de Twitter no es exactamente como la búsqueda en la página web.

Es importante saber que esta API se centra en la relevancia de los tweets. Esto significa que algunos Tweets y los usuarios no se encuentren, pero esa es tarea de la API Streaming.

- Permite a un usuario consultar el contenido de twitter.
- Búsqueda de un conjunto de tweets con palabras clave específicas.
- Búsqueda de los tweets que hacen referencia a un usuario específico.

REST API:
La API REST proporciona acceso programático para leer y escribir datos de Twitter. Conocer al autor de un nuevo tweet, leer perfiles de usuarios y datos de seguidores, entre otras características. La API REST proporciona las respuestas en formato JSON.

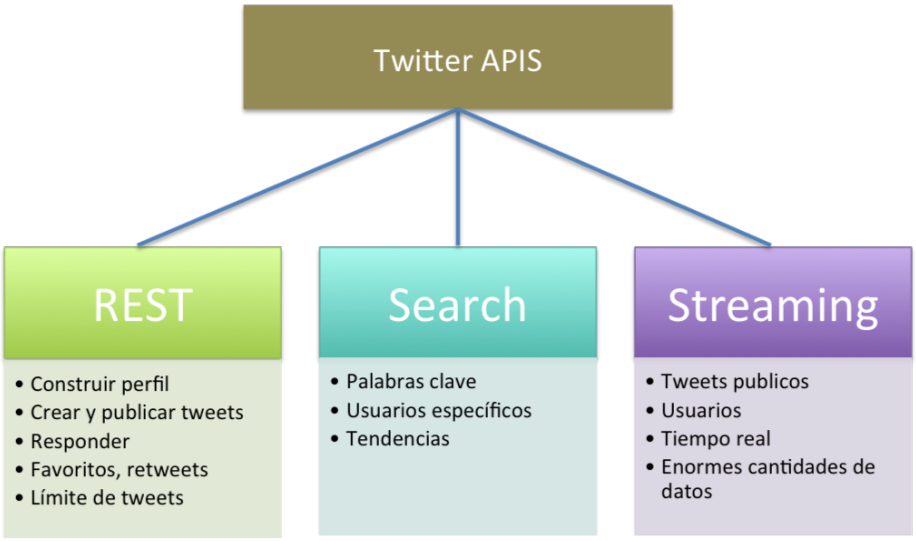

Figura A.1: APIS de Twitter.

- Permite acceder a algunas de las primitivas básicas de twitter, tales como actualizaciones de estado, y la información del usuario.
- Ofrece acceso mediante programación a la línea de tiempo, el estado, y los objetos de usuario.
- Brinda oportunidades de integración para interactuar con twitter. Se pueden crear y publicar tweets, así como dar respuesta a los tweets.

STREAMING API:

- Es la muestra en tiempo real de twitter.
- Utilizada para tareas de minería de datos e investigación analítica.
- Permite el uso de palabras, recuperar los tweets de una determinada región, o tener el conjunto de los estados públicos de un determinado usuario.

Es importante saber cómo realizar la paginación de las peticiones de una manera óptima, para poder obtener la mayor cantidad de datos posibles. En la tabla A.1 se encuentran las restricciones que cada una de las APIS tienen.

A.2. MySQL

MySQL es un sistema gestor de base de datos de código abierto, desarrollada por la compañía sueca MySQL AB. MySQL es soportado por distintas plataformas, incluyendo Microsoft

API	Límite de tiempo	Límite de tamaño
Streaming	Sólo tiempo real	-
Search	7 días	1500 últimos tweets
REST	-	3200 últimos tweets

Tabla A.1: Limitación temporal de las APIS de Twitter.

Windows, las principales distribuciones de Linux, UNIX y Mac OS X. MySQL tiene versiones gratuitas y de pago, en función de su uso (no comercial ó comercial) y características.

El servidor de base de datos SQL, con el que cuneta este sistema gestor, tiene un excelente rendimiento, es robusto, permite acceso multiusuario y cuenta con un entorno multithread (múltiples hilos a la vez).

A.3. Standford NLP

Para el desarrollo de nuestra herramienta se utilizó software que es importante describir para comprender la utilidad del mismo.

El grupo de desarrolladores de Stanford NLP implementa parte del software de procesamiento del lenguaje natural, el cual está disponible de forma gratuita. Se cuenta con una serie de herramientas para abordar algunos de los principales problemas de lingüística computacional. Dichos paquetes pueden ser incorporados en aplicaciones que necesiten trabajar con el lenguaje humano.

Todo el software que Stanford NLP distribuye está escrito en Java. Como resultado, gran parte de este software también puede ser fácilmente utilizado desde Python, Ruby, Perl, Javascript, y otros lenguajes NET.

A.3.1. CoreNLP

Stanford CoreNLP proporciona un conjunto de herramientas de análisis de lenguaje natural que al dar como entrada texto en inglés, se obtiene como salida las formas base de palabras. De igual forma permite identificar partes de una oración, como nombres de empresas, personas, etc.; normalizar las fechas, horas, cantidades; y marcan la estructura de las oraciones en cuanto a frases y dependencias de palabras.

Stanford CoreNLP es un marco integrado, que hacen que sea muy fácil de aplicar un cúmulo de herramientas de análisis de lenguaje a una parte del texto. Sus análisis proporcionan los bloques de construcción fundamentales para la de más alto nivel y aplicaciones

de comprensión de texto específicos de dominio.

Stanford CoreNLP integra muchas de las herramientas de NLP, incluyendo el etiquetador (POS), el llamado reconocedor de entidad (NER), el analizador, el sistema de resolución de la correferencia y el análisis de los sentimientos. La distribución básica proporciona los archivos de modelo para el análisis de textos en inglés, pero el motor es compatible con los modelos para otros idiomas.

El objetivo de este proyecto es permitir a los usuarios obtener rápidamente y sin dificultades anotaciones lingüísticas completas de textos en lenguaje natural. Está diseñado para ser altamente flexible y extensible. Con una opción que permite cambiar qué herramientas deben estar habilitadas y cuáles deben ser desactivadas.

A.3.2. POS Tagger

A Part-Of-Speech Tagger (POS Tagger) es un software que lee el texto en un idioma y asigna partes de la oración a cada palabra (token), como sustantivo, verbo, adjetivo, etc; aunque generalmente aplicaciones computacionales utilizan las etiquetas más especificas como 'sustantivo-plural'.

Varias descargas están disponibles. La descarga básica contiene dos modelos tagger capacitados para el idioma inglés, sin embargo, es importante mencionar que el etiquetador se puede formar en cualquier idioma, dado texto de entrenamiento para el idioma que se requiera.

Las principales razones por la cuales se eligió este etiquetador son su velocidad, rendimiento, facilidad de uso y soporte para otros idiomas.

Apéndice B

Administración de proyectos de Twitter

Twitter cuenta con un módulo para desarrolladores, el cual permite tener acceso a determinado conjunto de datos para posteriormente poder realizar experimentos con ellos. Para poder realizar peticiones al servidor de Twitter, es necesario contar con cuatro claves de acceso. En este apéndice se muestra el procedimiento a seguir para poder crear una aplicación, la cual nos proporciona las claves necesarias para poder obtener datos de esta red social.

1. Acceder a la página https://apps.twitter.com/.

2. Crear una nueva aplicación.

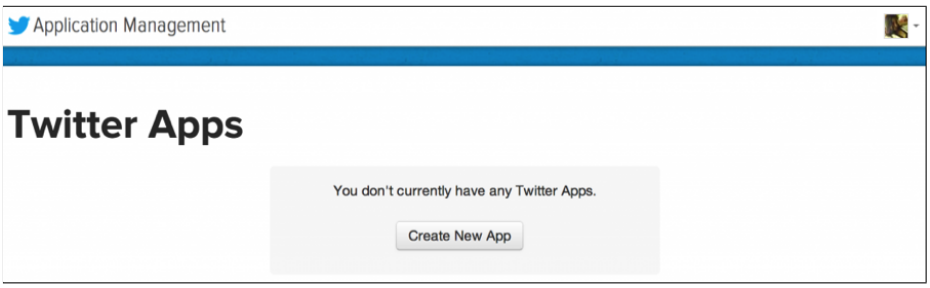

Figura B.1: Creación de una nueva aplicación.

3. Llenar los campos con información de la aplicación: nombre, descripción, sitio web y URL.

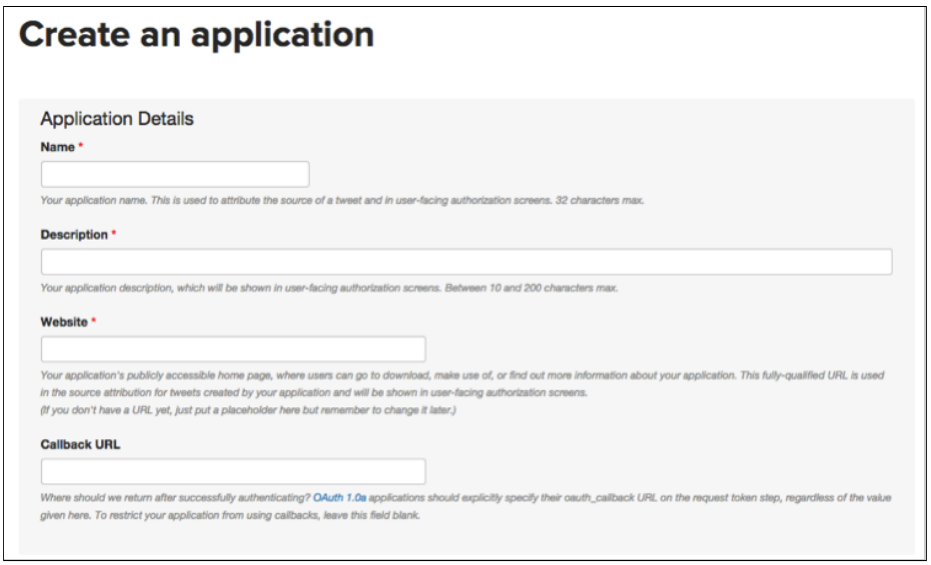

Figura B.2: Información de la aplicación a crear.

4. Guardar la nueva aplicación creada.

5. Configurar los permisos de lectura, escritura y acceso a mensajes directos.

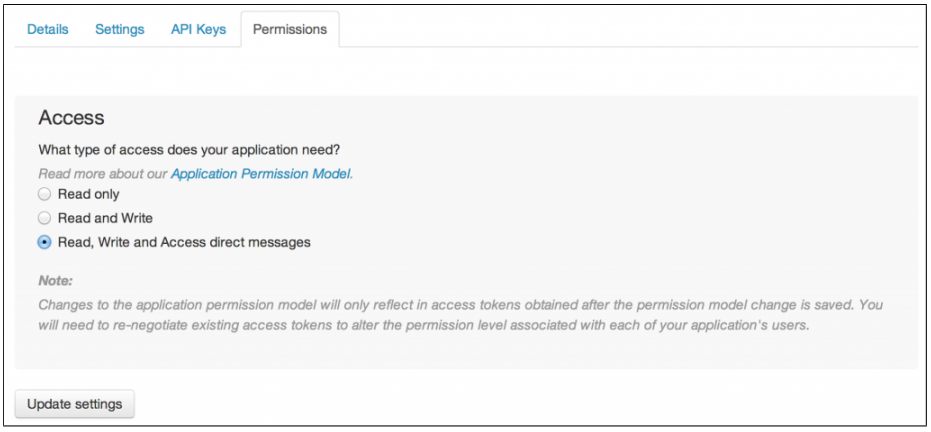

Figura B.3: Configuración de permisos.

6. En la pestaña 'API Keys' se encontrara las claves generadas para la aplicación en cuestión: 'API key' y 'API secret', dichas claves van a permitir la autenticación de la aplicación.

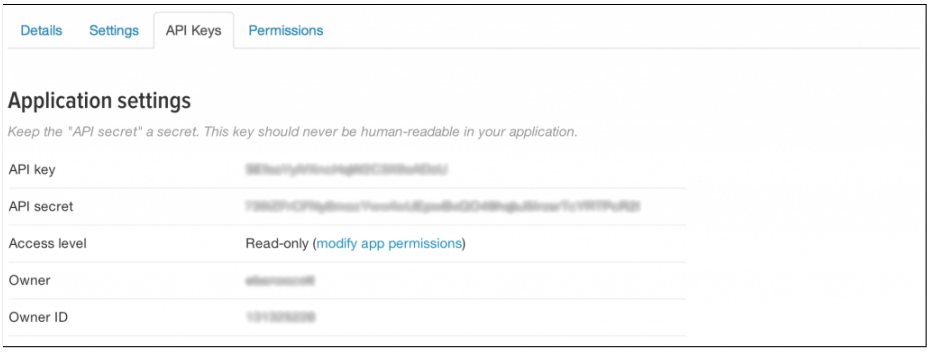

Figura B.4: Creación de las claves de la API.

7. Posteriormente hacer clic en el botón 'Create my access token' para la generación de dos claves más.

8. Finalmente se obtendrán dos cifras más, 'Access token' y 'Access token secret'. Estas claves son necesarias para realizar peticiones al servidor de Twitter, ya que permiten la autenticación del desarrollador de la aplicación.

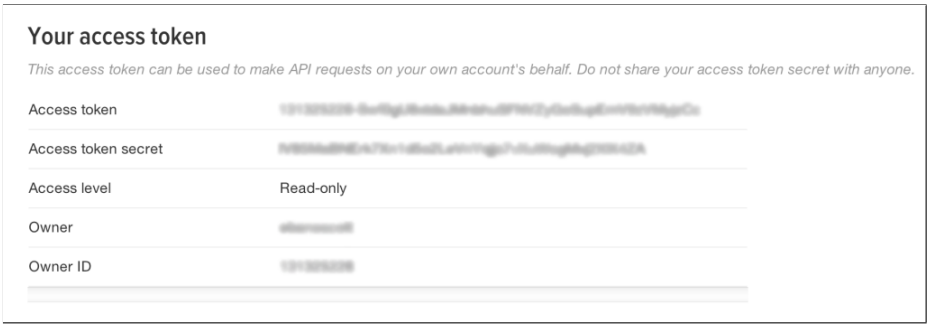

Figura B.5: Generación de tokens.

Referencias

[1] Tim O'Reilly. What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. MPRA Paper 4578, University Library of Munich, Germany, March 2007.

[2] Darcy DiNucci. Fragmented future. *Print*, 53(4):32+.

[3] Bing Liu. *Web Data Mining. Exploring Hyperlinks, Contents, and Usage Data*. Second edition edition, 2011.

[4] http://www.emarketer.com/article/social-networking-reaches-nearly-one-four-around-world/1009976, Junio 2013.

[5] P. Maes. Agents that reduce work and information overload. *Communications of the ACM*, 37(7):30–40, 1994.

[6] R. Kosala, H. Blockeel. Social network analysis: Methods and applications. *SIGKDD Explorations*, 2(1):1–15, 2000.

[7] O. Etzioni. The world wide web: Quagmire or gold mine. *Communications of the ACM*, 39(11):65–68, 1996.

[8] J. Cowie, W. Lehnert. Information extraction. *Communications of the ACM*, 39(1):80–91, 1996.

[9] N. Fiedman, L. Getoor, D. Koller, A. Pfeffer. Learning probabilistic relational models. In *IJCAI99*, pages 1300–1309, 1999.

[10] K. Kersting, Luc De Raedt. Bayesian logic programs. In *Proceedings of the Work-in-Progress Track at the 10th International Conference on Inductive Logic Programming*, pages 138–155, 2000.

[11] P. Flach, N. Lachiche. A first-order bayesian classifier. In *Proceedings of the Ninth International Workshop on Inductive Logic Programming (ILP99)*, pages 92–103, 1999.

[12] J. Neville, D. Jensen. Supporting relational knowledge discovery: Lessons in architecture and algorithm design. In *Proceedings of the Data Mining Lessons Learned Workshop, 19 th International Conference on Mechine Learning*, 2002.

[13] S. Muggleton. *Inductive Logic Programming.* Academic Press, 1992.

[14] S. Dzeroski, N. Lavrac. *Relational Data Mining.* Springer, 2001.

[15] S. Wassermann, K. Faust. Social network analysis: Methods and applications. In *Cambridge: Cambridge University Press*, 1994.

[16] H. Kosala, R. y Blockeel. Web mining research: A survey. *ACM SIGKDD Explorations Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining*, 2(1):1–15, Junio 2000.

[17] M. Kamber J. Han and J. Pei. *Data Mining: Concepts and Techniques.* Morgan Kaufmann Publishers, second edition edition, 2006.

[18] Jesús García Herrero. José Manuel Molina López. Técnicas de análisis de datos, 2006.

[19] G. W. Furnas. *The vocabulary problem in human system communication*, volume 30. Communications of the ACM, Noviembre 1987.

[20] Max Bramer. *Principles of Data Mining.* Springer, 2007.

[21] Stephanie Rich Weiguo Fan, Linda Wallace and Zhongju Zhang. Tapping into the power of text mining. *ACM*, 2005.

[22] Peter Jackson and Isabelle Moulinier. *Natural Language Processing for Online Applications: Text retrieval, extraction and categorization.* John Benjamins, second edition edition, 2002.

[23] Pennacchiotti M. y Zanzotto F.M. Pazienza, M.T. *Terminology extraction: an analysis of linguistic and statistical approaches*, pages 255–280. Knowledge Mining. Studies in Fuzziness and Soft Computing, 2005.

[24] https://dev.twitter.com/, 2014.

[25] Itai Himelboim Ben Shneiderman Marc A. Smith, Lee Rainie. Mapping twitter topic networks: From polarized crowds to community clusters. Technical report, Pew research Center, Febrero 2014.

[26] NIMA ASADI and JIMMY LIN. Fast candidate generation for real-time tweet search with bloom filter chains. *ACM Transactions on Information Systems*, 13(3), Julio 2013.

[27] Jimmy Lin Craig Macdonald Iadh Ounis Dean McCullough Richard McCreadie, Ian Soboroff. On building a reusable twitter corpus. *SIGIR'12*, Agosto 2012.

[28] *EvenTweet: Online Localized Event Detection from Twitter*, 2013.

[29] Jianshu Weng, Yuxia Yao, Erwin Leonardi, and Francis Lee. Event detection in twitter. Technical report, Noviembre 2011.

[30] K. P. Subbalakshmi Rohan D.W Perera, S. Anand and R. Chandramouli. Twitter analytics: Architecture, tools and analysis. In *The Military Communications Conference. Cyber Security and Network Management.*, 2010.

[31] https://www.tweetarchivist.com/, Enero 2015.

[32] http://topsy.com/, Enero 2015.

[33] Cheng Li Feida Zhu Juan Du, Wei Xie and Ee-Peng Lim. Twicube: A real-time twitter off-line community analysis tool. In Springer-Verlag., editor, *DASFAA*, volume Part II, LNCS 7826, pages 458–462, Berlin Heidelberg., 2013.

[34] B.E. Teitler M.D. Lieberman J. Sankaranarayanan, H. Samet and J. Sperling. Twitterstand: News in tweets. In ACM, editor, *The 17th ACM SIGSPATIAL*, pages 42–51, 2009.

[35] http://nlp.stanford.edu/software/, 2014.

[36] M. Mitra A. Singhal, G. Salton and C. Buckley. Document length normalization. *Information Processing and Management*, 5(32):619–633, 1996.