Facultad de
Comunicación y Documentación

UNIVERSIDAD DE GRANADA

GRADO EN INFORMACIÓN Y DOCUMENTACIÓN

TRABAJO FIN DE GRADO

Desarrollo de una herramienta para la extracción y análisis de datos procedentes de redes sociales. Caso práctico: la API de Twitter

Presentado por:

Dª. Miriam Carrasco Alanís

Tutor:

Prof. Dr. Antonio Gabriel López Herrera

Curso académico 2018 / 2019

D.: Antonio Gabriel López Herrera, tutor del trabajo titulado **Desarrollo de una herramienta para la extracción y análisis de datos procedentes de redes sociales. Caso práctico: la API de Twitters** realizado por el alumna **Miriam Carrasco Alanís**, INFORMA que dicho trabajo cumple con los requisitos exigidos por el Reglamento sobre Trabajos Fin del Grado en *Información y Documentación* para su defensa.

Granada, 05 de Julio de 2019

LOPEZ HERRERA ANTONIO GABRIEL - 75101574L

Firmado digitalmente por LOPEZ HERRERA ANTONIO GABRIEL - 75101574L
Nombre de reconocimiento (DN): c=ES, serialNumber=IDCES-75101574L, givenName=ANTONIO GABRIEL, sn=LOPEZ HERRERA, cn=LOPEZ HERRERA ANTONIO GABRIEL - 75101574L
Fecha: 2019.07.05 11:49:52 +02'00'

Fdo.: _____

Por la presente dejo constancia de ser la autora del trabajo titulado **Desarrollo de una herramienta para la extracción y análisis de datos procedentes de redes sociales. Caso práctico: la API de Twitters** que presento para la materia Trabajo Fin de Grado del Grado en **Información y Documentación**, tutorizado por el profesor Antonio Gabriel López Herrera durante el curso académico 2018- 2019.

Asumo la originalidad del trabajo y declaro que no he utilizado fuentes (tablas, textos, imágenes, medios audiovisuales, datos y software) sin citar debidamente, quedando la Facultad de Comunicación y Documentación de la Universidad de Granada exenta de toda obligación al respecto.

Autorizo a la Facultad de Comunicación y Documentación a utilizar este material para ser consultado con fines docentes dado que constituyen ejercicios académicos de uso interno.

05 / 07 / 2019

Fecha Firma

AGRADECIMIENTOS

En primer lugar, quiero agradecer a mi tutor, Antonio Gabriel, por sus consejos, sus continuos mensajes de ánimo y su paciencia. Y sobre todo por confiar en mi hasta el último momento.

Gracias a mis familiares y seres queridos, por animarme constantemente y ofrecerme todo su apoyo durante esta etapa. Papá, mamá, Raúl gracias por hacerlo posible.

Por último, gracias a mi pareja por todo su sacrificio y paciencia, y por saber como sacar lo mejor de mí.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RESUMEN

Dada la masiva cantidad de datos que la sociedad moderna en la que vivimos, produce y digiere casi a diario, es prácticamente una necesidad el poder hacer frente y dar forma a estas estructuras de datos, transformarlas en información, y seguidamente en conocimiento. Es por ello, que el presente Trabajo Fin de Grado tiene como objetivo hacer uso de las herramientas de recuperación y tratamiento de la información, en el contexto de las tecnologías de la información y comunicación, y más concretamente de lo que conocemos como redes sociales.

Para ello se plantea un modelo de extracción y visualización de datos obtenidos desde la red social *Twitter,* los cuales serán almacenados en dos bases de datos diferentes y procesados a través de una aplicación que desarrollaremos para este fin.

Abstract

Given the massive amount of data that the modern society in which we live, produces and digests almost daily, it is practically a necessity to be able to face and shape these data structures, transform them into information, and then into knowledge. It is for this reason, that the present end of degree project has as objective to make use of the tools of retrieval and treatment of information, in the context of information and communication technologies, and more specifically of what we know as social networks.

For this purpose, a model of extraction and visualization of data obtained from the social network Twitter is proposed, this data will be stored in two different databases and processed through an application that we will develop.

1.- INTRODUCCIÓN

Es con el rápido crecimiento y desarrollo de la tecnología de Internet que surge otro tipo de servicios más relacionados con la comunicación. Estos servicios se vuelcan de lleno en la aparición de plataformas que hoy en día conocemos como redes sociales.

Las redes sociales fueron creadas con el propósito de hacer posible la comunicación y la socialización. Hoy en día, sin embargo, su cometido ha evolucionado mucho más allá de ser una puerta para la comunicación entre personas de diferentes puntos de nuestra geografía, y es que, actualmente sirven como un trampolín para la creación de nuevas modalidades de negocio, como el *marketing digital*, del cual surgen nuevas profesiones o perfiles tales como el *community manager,* el *analista web* o los relacionados con el posicionamiento en buscadores, entre otros. Lo que todos estos perfiles profesionales tienen en común, es que trabajan con los datos que los propios usuarios producen y ofrecen. De esta forma, es posible acceder a los millones de datos públicos que a diario se almacenan en gigantescas infraestructuras pertenecientes a las grandes empresas del sector. Esto hace fundamental el poder analizar, procesar y generar decisiones basadas en el comportamiento de los usuarios siendo así interpretadas y transformadas en información útil, dejando de ser únicamente un dato almacenado más.

Somos conscientes de la importancia a nivel global que tienen los medios sociales, desde el punto de vista de la información, y de cómo el tratamiento de dicha información es fundamental para conocer al usuario y suplir sus necesidades. Esto se identifica claramente como uno de los objetivos de cualquier futuro graduado en Información y Documentación.

A lo largo de este trabajo mostramos: un método de extracción de datos, haciendo uso de la API de Twitter con la que desarrollamos una herramienta que permita extraer y almacenar datos específicos en una base de datos concreta. Reflexionando finalmente sobre el alcance e importancia que tiene el análisis de redes sociales en una sociedad como la nuestra.

A lo largo de esta introducción hablaremos del concepto de red social, y de todos

aquellos que estén asociados a la red social Twitter y que sean necesarios para la compresión del presente trabajo. Hablaremos también sobre la historia que precede al surgimiento de Internet y las redes sociales, a la vez que se mostrará una evolución y clasificación de las mismas.

1.1- Antecedentes

Con la aparición en los años sesenta de la Agencia de Proyectos para la Investigación Avanzada de Estados Unidos (ARPA) y, hasta que se consolidad en 1970 la red ARPANET, se desata toda una revolución, siendo entonces establecidas las bases para lo que conocemos hoy como correo electrónico, esto provoco que el número de ordenadores conectados fuera aumentando. Es ya en el año 1983, con el uso por primera vez del protocolo TCP/IP, cuando surge la primera denominación de la palabra "Internet", derivada de la red "Arpa Internet". Años más tarde, en 1989, Tim Berners Lee desarrollaría por primera vez lo que se conoce como la *World Wide Web*, provocando que en la década de los 90 el crecimiento de la creación de sitios web se incrementara notablemente (Bahillo, 2019).

La web 1.0 revolucionó entonces el acceso a la información. Sin embargo esta carecía de algo que hoy en día es fundamental, el poder del usuario de editar e interactuar con el contenido, es decir, carecía de dinamismo, siendo todos los sitios webs estáticos en cuanto a contenido y estructura. Un ejemplo de Web 1.0 sería el de un sitio web con contenido estático hecho en código HTML y sin ofrecer la capacidad para interactuar mediante foros o chats. Debido a esto aparece en 2004 el término "Web 2.0", acuñado por Dale Dougherty y ligado a Tim O´Reilly y su editorial O´Reilly Media. "Aunque el término sugiere una nueva versión de la World Wide Web, no se refiere a una actualización de las especificaciones técnicas de la web, sino más bien a cambios acumulativos en la forma en la que desarrolladores de software y usuarios finales utilizan la Web" (Salazar, 2011).

Es gracias a la Web social o Web 2.0 que se hace posible el que surjan posteriormente las primeras redes sociales, siendo las más conocidas y longevas, *LinkedIn* (2002), *Flickr* (2004), *Facebook* (2005), *YouTube* (2005), *Twitter* (2006), entre otras.

Más tarde y a partir del crecimiento de estas redes sociales surgen derivadas de ellas, los primeros sistemas de publicidad, como *Facebook Ads* o *Google Ads*, también gracias a la aparición de los primeros dispositivos móviles con acceso a internet se facilitaba el crecimiento del *e-commerce* o comercio electrónico, con empresas como *Amazon, ASOS, Alibaba* y muchas otras.

Todo esto nos lleva a los grandes avances que le precederían a lo largo de los siguientes años llegando a alcanzar un exponencial crecimiento en la última década.

1.2- Concepto de red social

Es importante no confundir los medios sociales (*Social media* en inglés) con el concepto de red social. Es por ello que debemos aclarar que nos referimos a medios sociales como forma de englobar a una serie de herramientas que generan y participan en su propio contenido, lo cual viene a decir que sin interacción, no es un medio social, esto incluye a los blogs, marcadores sociales, las plataformas de multimedia (*YouTube, Vimeo, Itunes),* la geolocalización y más destacadamente a las redes sociales(*social networks)*, por esto, ambos conceptos están tan íntimamente ligados, siendo las redes sociales un producto de los medios sociales.
Sin embargo, y ya que en español no se diferencia demasiado esta terminología, nos referiremos únicamente a red social a lo largo del trabajo.

El concepto de red social ha ido evolucionando y adquiriendo una gran relevancia durante los últimos años. Dado que las definiciones sobre red social son diversas y varían dependiendo de la fuente he decidido recopilar algunas de ellas de forma que se pueda entender este término de una forma más amplia.

- Las redes sociales son sitios de Internet formados por comunidades de individuos con intereses o actividades en común (como amistad, parentesco, trabajo) y que permiten el contacto entre estos, de manera que se puedan comunicar e intercambiar información (Raffino, 2019).
- En sentido amplio, una red social es una estructura social formada por personas o entidades conectadas y unidas entre sí por algún tipo de relación o interés

común (Ponce, 2012).

- Una Red Social es una estructura social integrada por personas, organizaciones o entidades que se encuentran conectadas entre sí por una o varios tipos de relaciones como ser: relaciones de amistad, parentesco, económicas, relaciones sexuales, intereses comunes, experimentación de las mismas creencias, entre otras posibilidades (Lorenz, 2010).

Debemos, eso sí, diferenciar el concepto de red social con el de Servicio de red social:

- Los servicios de redes sociales son la infraestructura tecnológica sobre la que se crean las relaciones y, por tanto, las redes sociales. Es decir, son aplicaciones que ponen en contacto a las personas a través de Internet. (de Haro, 2010).

1.2.1.- Clasificación de redes sociales

Existen diferentes formas de clasificar a las redes sociales (ver Fig.1), en la siguiente tabla mostraremos las más destacadas y algunos ejemplos (Ponce, 2012), (Civico Cabrera, 2017):

Tipos	Definición	Ejemplos
Horizontal	Dirigidas a cualquier usuario y sin tener una temática definida y se centra en los contactos. Siendo su función principal la relacionar a las personas y ofrecer para ello el uso y creación de perfiles, generar listas de contactos y compartir contenidos entre estos.	*Google, Facebook, Instagram, Line, Twitter, Badoo.*
Vertical	Este tipo de redes están dedicadas a la especialización, y pueden a su vez subdividirse según el asunto que traten, por actividad, el tipo de contenido o la ubicación.	
	1. Por temática: 1.1 Profesionales 1.2 Identidad cultural 1.3 Aficiones / viajes	 1.1 LinkedIn, ResearchGate 1.2 Spaniards 1.3 Trover, Foursquare
	2. Por actividad: 2.1 Microblogging 2.2 Juegos 2.3 Marcadores sociales	 2.1 Tumblr, Twitter 2.2 Haboo, Discord, World of Warcraft.
	3. Por contenido compartido: 3.1 Fotos 3.2 Música	 3.1 Flickr, Pinterest,

	3.3 Vídeos 3.4 Documentos 3.5 Presentaciones 3.6 Noticias 3.7 Lectura	WeHeartIt 3.2 Musically, Myspace 3.3 YouTube 3.4 Scribd 3.5 SlideShare 3.6 Feedly, Reddit 3.7 Wattpad
	4. Según la localización geográfica: 4. 1 Sedentarias: se actualizan acorde a los contenidos compartidos en ellas y las relaciones entre sus usuarios. 4.2 Nómadas: van cambiando conforme los usuarios se van moviendo.	 4.1 Blogger 4.2 Foursquare

Tabla 1: Clasificación redes sociales. *Origen:* elaboración propia

1.3- Origen y evolución de las redes sociales

Partimos desde la teoría de los seis grados de separación, según la cual toda la gente del planeta estaría conectada a través de no más de seis personas. Esta teoría o concepto sería el punto de partida de las redes sociales a mediados de los años 90.

Y es en torno a 1995 cuando surge el sitio web classmates.com que permitía conectar con antiguos compañeros en colegios, institutos y universidades en Estados Unidos.

Más tarde, para 2003, ya eran numerosas las plataformas existentes (MySpace, LinkedIn, Friendster, entre otras) que aunque con diferentes propuestas destacaban con un único fin común: el de conectar a la gente. Sin embargo fueron solo aquellas que consiguieron renovarse a lo largo de los años las que sobrevivieron.

Facebook que surgía también dentro del ámbito académico como forma de conectar a los estudiantes en las universidades acabaría ampliando su espacio al resto de usuarios de Internet convirtiéndose actualmente en la red social más grande e influyente a nivel mundial.

1.4- El concepto de *microblogging* y Twitter

Ya que este trabajo está centrado en la extracción de datos exclusivamente de la red social Twitter. Dedicaremos esta parte a conocer brevemente algunos aspectos básicos sobre la misma.

Conocemos a Twitter como un servicio de *microblogging,* ¿pero a qué nos referimos exactamente con este término? Entendemos por *microblogging* el servicio de enviar y publicar mensajes breves, generalmente entorno a los 140 caracteres, aunque como veremos con Twitter, esto ha ido cambiando. Esto obliga al usuario a sintetizar y adaptar su contenido a un reducido número de palabras. Otro ejemplo de red social de *microblogging* es Tumblr.

Twitter surge por primera vez en 2006 de la mano de Jack Dorsey, Noah Glass, Biz Stone, Evan Williams, siendo estos sus principales fundadores, aunque se atribuye su creación a Jack Dorsey (Egea, 2007).

Su crecimiento progresó rápidamente y es que actualmente cuenta con un total de 300 millones de usuarios activos (We Are Digital, 2019).

El funcionamiento de esta red es sencillo, el usuario registrado publica un mensaje de no más de 280 caracteres[1] y aquellos usuarios que formen parte de su red de seguidores podrán visualizar todo lo que este publique en su página principal o perfil. A su vez este puede seguir a otros usuarios. Al mismo tiempo los mensajes son públicos para todo el mundo, aunque esto puede configurarse para que sean privados y solo los seguidores que sean aprobados puedan verlos.

Llamamos seguir o *follow* al acto de subscribirse a los *tweets (*tuits en español) o mensajes de un usuario, convirtiéndose en seguidores o *followers* al que lo hace.
Es decir un usuario puede tener seguidores (*followers)* y al mismo tiempo seguir a otros usuarios (hacer *follow).*
Es sabido, además, que se la considera también como una red de difusión, ya que actualmente se priorizan muchísimo las tendencias de información. Siendo muchos los

[1] Anteriormente Twitter solo permitía un máximo de 140 caracteres por tweet, y fue en 2017 cuando el número de caracteres se duplicaron (Blog.twitter.com, 2017)

usuarios que utilizan este medio como uno de los primeros lugares a los que acudir cuando quieren informarse sobre una noticia, por delante de otros medios digitales, prensa, etc. (Ver Figura 1).

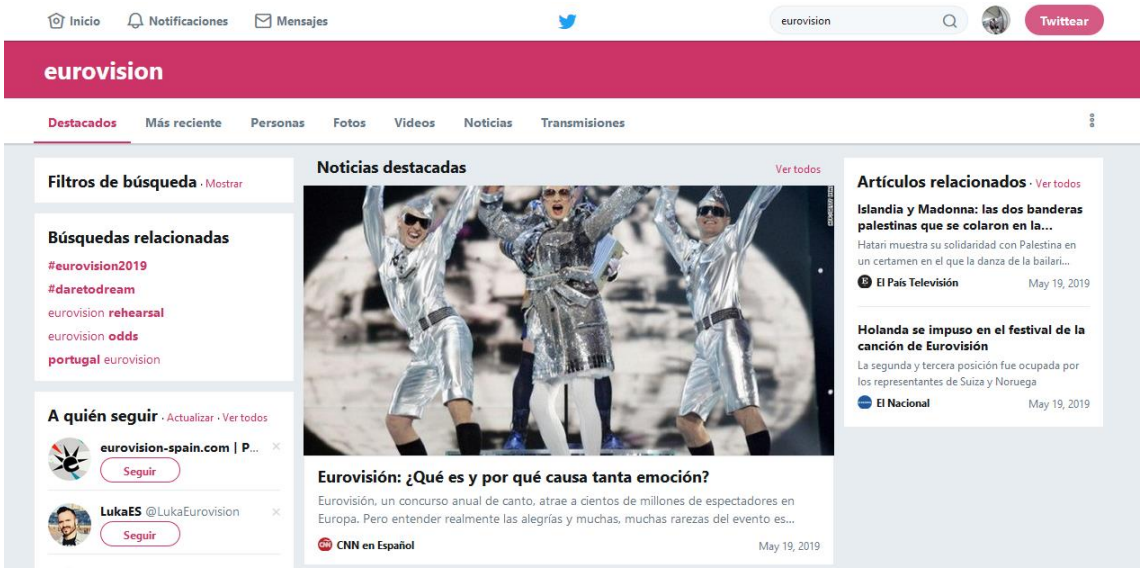

Figura 1: Noticias destacadas como resultado de la búsqueda por un término

El hecho de que Twitter sea un medio para dar cobijo a los sucesos en tiempo real y repercuta en el uso y visualización de la información es una de las razones por las que queremos centrarnos en esta red social y no en otra.

Twitter cuenta además con otras características y términos propios como:

- *Hashtag*: "*es una cadena de caracteres formada por una o varias palabras concatenadas y precedidas por una almohadilla o numeral* (#)" (Jarould, 2019).
- *Timeline*: página principal de Twitter en la cual aparecen los mensajes de aquellos usuarios a los que sigues.
- *Trending Topic* (TT): se refiere a las tendencias sobre un tema en tiempo real que se agrupan en forma de listas enumeradas según el país o globalmente.
- Retuit (Retweet): permite que la información publicada por otro usuario forme parte de nuestro perfil y pueda ser vista por nuestros seguidores.
- Me gusta: antes conocido como "favorito", permite marcar con un símbolo en forma de corazón el tweet que has leído indicando que te interesa.

2.- ESTADO DEL ARTE

Twitter es una red social que ofrece la oportunidad a los desarrolladores de trabajar con sus servicios y crear aplicaciones en beneficio de la comunidad. A lo largo de esta sección hablaremos sobre los distintos métodos y herramientas de extracción a nivel general.

2.1-Métodos de extracción de datos: diferencias entre API, Servicios Web o *Web Services y Web Scraping*

Hoy en día existen multitud de métodos con los que extraer información procedente de redes sociales. En nuestro caso nos centraremos particularmente en tres de ellos, las APIs, los Servicios Web y el *Web Scraping*.

Servicio Web

Un servicio Web está diseñado para tener una interfaz que se representa en un formato procesable por una máquina. Además no están ligados a ningún lenguaje de programación ni a sistemas operativos, dado que el protocolo que utilizan para la comunicación es HTTP. Los Servicios Web también utilizan SOAP, REST y XML-RPC como medio de comunicación.

Dicho en otras palabras, permite intercambiar datos entre aplicaciones utilizando protocolos y formatos procesables por un equipo informático.

API

Las APIs (por su siglas en inglés, A*pplication Programming Interface)* o Interfaz de Programación de Aplicaciones son un paso más en torno a lo que conocemos como Servicios web. Permiten que un programa se comunique con otro a través de una serie de reglas. Normalmente la API (Go4it.solutions, 2019) lleva a cabo sus funciones desde dentro de un programa de software. Cuando la API debe enviar datos a través de una red, entra en escena el Servicio Web.

Las APIs simplifican bastante el trabajo de un programador, ya que no tiene que «escribir» códigos desde cero. Y es que, le permiten usar funciones predefinidas para interactuar con el programa.

En (Go4it.solutions, 2019) apunta que una de las diferencias entre API y Servicio Web es que la API es capaz de definir con total exactitud el modo, el método o métodos que un programa usará para comunicarse con otros. Otra de las diferencias entre API y Servicio Web es que este último no contiene todas las reglas que facilitan la comunicación. Por eso, son capaces de realizar menos funciones que las APIs.

Web Scraping

Otro de los métodos más populares para la extracción de datos es el Web scraping, que traducido literalmente al español vendría a ser "raspado web" o "escarbar en la web". Ya que no existe una traducción exacta al español nos dirigiremos a este método por su término en inglés.

Algunas de las definiciones que mejor describen este término son:

"También conocido como Web harvesting o Web data extraction, es el proceso de rastreo y descarga de sitios web de información y la extracción de datos no estructurados o poco estructurados a un formato estructurado. Para lograrlo, se simula la exploración humana de la World Wide Web, ya sea por implementación de bajo nivel del protocolo de transferencia de hipertexto, o la incorporación de ciertos navegadores web." (Tadeo Hernández *et al.,* 2015)

"Web "scraping" (also called "web harvesting," "web data extraction," or even "web data mining"), can be defined as "the construction of an agent to download, parse, and organize data from the web in an automated manner." Or, in other words: instead of a human end user clicking away in a web browser and copy-pasting interesting parts into, say, a spreadsheet, web scraping offloads this task to a computer program that can

execute it much faster, and more correctly, than a human can." (Vanden Broucke y Baesens, 2018).

Ambas coinciden en que este método simula la acción de un ser humano al buscar en la web, facilitando en última instancia esta tarea y reduciendo de forma importante el tiempo de búsqueda que si fuera de forma manual.

Algunas características propias del *Web scraping* son el uso de expresiones regulares, además algunas de las razones que pueden llevarte a la elección del *scraping* como método de extracción, es que el sitio del que queremos extraer información no disponga de una API, que esta esté limitada (en cuanto al número de veces que podemos extraer en un periodo determinado de tiempo como es el caso de la API de Twitter), que la API no sea de acceso gratuito y el sitio web sí, y por último, que no obtenga todos los datos necesitados y que el sitio web si los ofrezca.

2.2- Herramientas para la extracción de datos

Con el propósito de poder entender las estructuras que forman las redes, los investigadores deben utilizar métodos como la minería de datos o las técnicas que hemos mencionado anteriormente. Para ello, se ayudan de una gran variedad de herramientas construidas para un amplio conjunto de propósitos.

- Twitter API: cuando nos referimos a la API de Twitter deberíamos estar refiriéndonos a las APIs en plural, ya que son varios los servicios de API que Twitter ofrece con diferentes finalidades cada una de ellas. Además en los últimos años ha sufrido una evolución de lo que conocíamos como REST API, Search API y Streaming API a las nuevas propuestas de Twitter, las cuales se diferencian en herramientas gratuitas y de pago. Primero haremos un repaso de las tradicionales para poder conocer su funcionamiento.

    - La REST API (Congosto, 2019) ofrece a los desarrolladores el acceso al núcleo de los datos de Twitter. Todas las operaciones que se pueden hacer vía web son posibles realizarlas desde la API, como ver los perfiles

de los usuarios, quienes son sus seguidores y seguidos, los tuits que publican, cuales son los trending topics, etc. En la REST API existe una limitación variable en función del método solicitado. Esta restricción se mide en solicitudes que se pueden realizar durante una ventana de 15 minutos. Los valores oscilan entre 15 y 900. Los métodos más restrictivos son los que proporcionan la listas de seguidores o seguidos de los usuarios que están limitados a 15 solicitudes. Los menos restringidos son las solicitudes de tuits de usuarios que permiten 900 consultas.

- La Search API provee los tuits de los últimos 7 días. Permite filtrar por lenguaje y localización. La limitación es de 180 peticiones cada 15 minutos (Congosto, 2019).
- El Streaming API permite extraer tuits casi en tiempo real al establecer una conexión permanente con los servidores de Twitter. Se puede filtrar con varios tipos de parámetros diferentes, siendo los más habituales palabras claves, usuarios y localizaciones. También es posible descargar una muestra aleatoria de tuits (statuses/sample). En la Streaming API, al ser un flujo continuo, la restricción se aplica al caudal recibido que nunca será mayor a 50 tuits por segundo (Congosto, 2019).

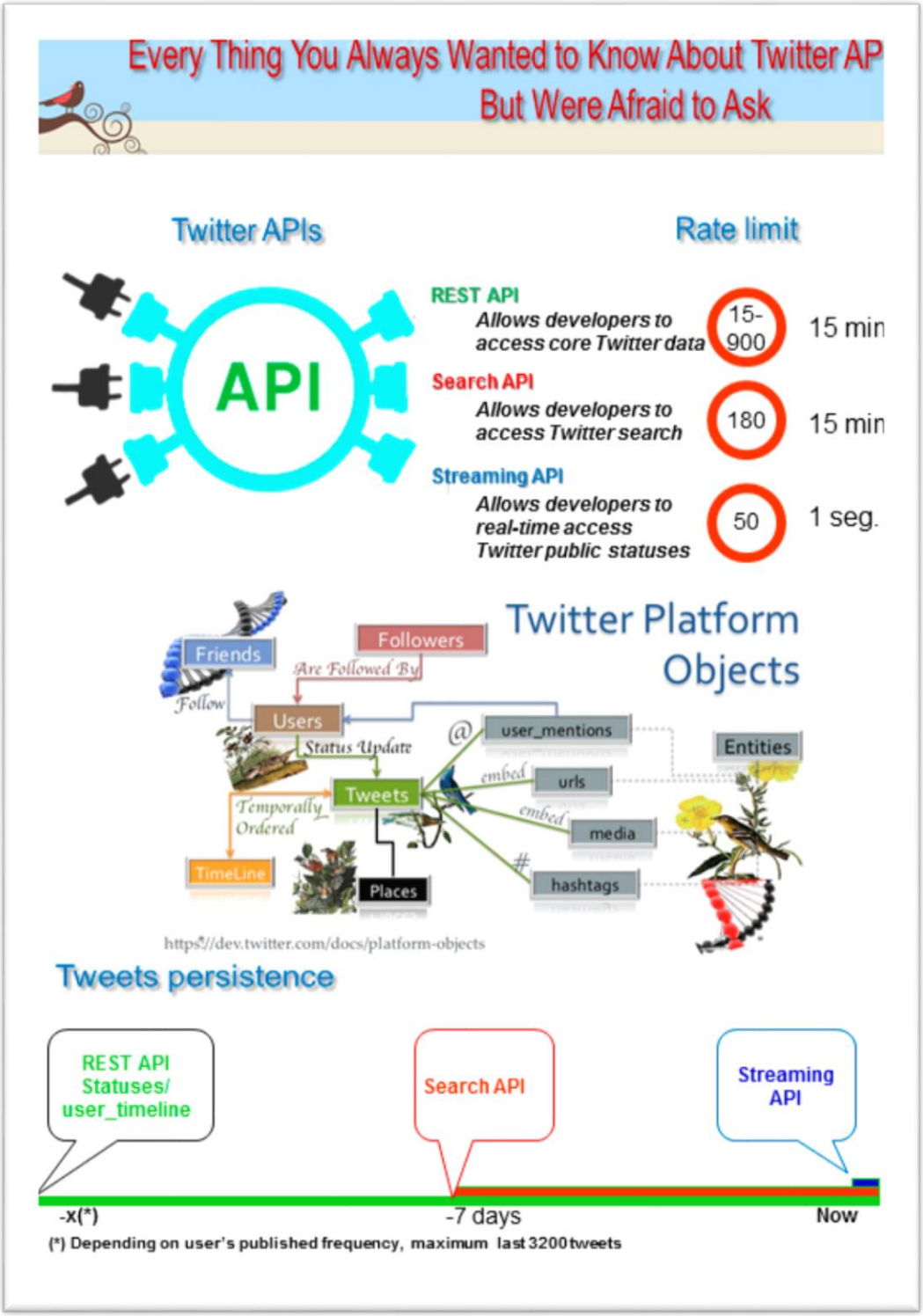

Figura 2: infografía sobre la obtención de tuits y limitaciones de cada una de las APIs. (Congosto, 2017)

Los nuevos servicios en los que ha evolucionado actualmente Twitter se desglosan en (Developer.twitter.com, 2019):

- Search tweets: nueva denominación de la Search API.
- Filter realtime Tweets: utiliza herramientas de filtrado de la Streaming API.
- Direct Message API: esta es más reciente y no implica a las APIs tradicionales.
- Ads API: su fin es gestionar publicidad.
- Account Activity API: permite seguir la actividad de más de 15 cuentas.
- Twitter for websites: para incluir los tuits y timelines en una página web.
- API reference index: contiene lo que antes se denominaba REST API.

- OpenSocial: consiste en un conjunto de APIs que prestan servicio a Google, Yahoo!, MySpace y otros muchos asociados, permite construir aplicaciones o redes sociales. Posibilita que programar aplicaciones para diferentes sitios de redes sociales sea común (Farisori, 2019).
- Scrapy: Es una herramienta pensada para programadores con conocimientos más avanzados de Python y para proyectos que no estén centrados en la visualización de datos (Martí, 2019).
- ScraperWiki: es una herramienta para la extracción de datos dispuestos en tablas en un PDF. Es tan sencillo como cargar el archivo y exportar. Ofrece una vista previa con todas las páginas y las distintas tablas y la posibilidad de descargar los datos de forma ordenada y separada (BBVAOpen4U, 2019).
- Jsoup de Java: es una biblioteca Java de código abierto diseñada para analizar, extraer y manipular datos almacenados en documentos HTML (Hedleyv, 2019).
- BeautifulSoup (Python): es una biblioteca de Python para realizar web scraping sobre documentos HTML. Esta biblioteca crea un árbol con todos los elementos del documento y puede ser utilizado para extraer información de sitios web (Kizar, 2019).

3.- OBJETIVOS

Una vez constituidos aquellos conceptos que consideramos básicos, se pretende fundamentar en esta sección del trabajo cual es la principal finalidad y cuáles son sus objetivos específicos.

Los objetivos que se establezcan deberán ser coherentes con las limitaciones temporales; al ser desarrollado en tan solo un semestre, y las tecnológicas relativas a la utilización de software y hardware existente.

El objetivo principal consiste en mostrar una forma de extraer datos concreta desde una red social para posteriormente analizar los datos obtenidos y de esta manera ayudar a otros futuros trabajos a profundizar en su conocimiento sobre el uso y tratamiento de redes sociales.

Los objetivos específicos que se cubre son:

a) Reunir la información existente relativa a la extracción de datos en redes sociales, recogiendo los métodos y las herramientas que se conocen actualmente. Poder diferenciar entre ellas y elegir cual es la más adecuada para nuestro caso de estudio.
b) Trabajar con estas herramientas y conocer su funcionamiento, de modo que puedan ser aplicadas y puestas en práctica con el objetivo final de demostrar su uso y facilitar el entendimiento de las mismas.
c) Desarrollo de un programa capaz de trabajar con estas herramientas en lenguaje de programación Python, que conecte con Twitter y pueda extraer la información desde los campos que determinemos para finalmente exportarse en un formato concreto.
d) Se realiza una exploración de los diferentes modelos de datos, como son el de tipo relacional con SQLite y el de tipo documental a partir de MongoDB.
e) Transformación de los datos en información interpretable, la cual será analizada de forma estadística y se podrá visualizar de manera clara y concisa.

4.- METODOLOGÍA

En este apartado establecemos la metodología del presente trabajo, el cual se basa en un estudio descriptivo y seccional de la red social Twitter, incluyendo una revisión de las diferentes técnicas y métodos de extracción que existen a nivel general.

El primer paso a tener en cuenta será el de buscar y recopilar la información necesaria para afianzar y profundizar en los conocimientos que se tendrán en cuenta a lo largo del desarrollo de este proyecto. Para facilitar y organizar el almacenamiento de las referencias bibliográficas se utiliza Mendeley (https://www.mendeley.com/) como gestor bibliográfico.

Tras esto, se hace una valoración de las herramientas que se utilizarán en el desarrollo de la parte práctica. En nuestro caso particular, se elige el lenguaje de programación Python, debido a su sencillez y flexibilidad para trabajar en diferentes entornos. También se opta a la hora de almacenar los datos resultantes por una base de datos de tipo no relacional o NoSQL, como es MongoDB y por otra de tipo relacional (SQLite).

En tercer lugar, se ha procederá a organizar la división del problema en 5 partes: 1. Construcción de un programa que descargue datos desde Twitter, 2. Volcado en una base de datos, 3. Procesamiento de los tuits, 4. Obtención de un fichero en formato XML 5. Visualización de los datos.

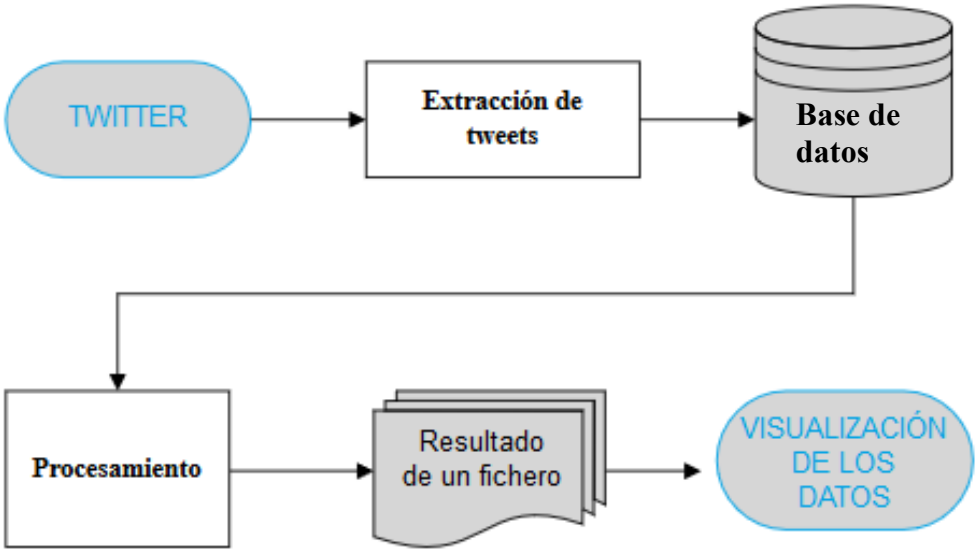

Figura 3: Diagrama de la arquitectura del sistema.

Finalmente, se ha desarrollado una herramienta siguiendo los pasos y los requerimientos mencionados en los párrafos anteriores, la cual al ser puramente práctica será anexionada al final del trabajo. También se hace una reflexión de los resultados obtenidos y las aplicaciones que se pueden poner en práctica.

5.- DESARROLLO Y RESULTADOS

Tal y como se ha mencionado anteriormente, introducimos una herramienta de elaboración propia, que permite la conexión directa con la red social Twitter la cual será explicada y detallada en la totalidad de su proceso. Hablaremos también de las diferencias entre los datos obtenidos con cada una de las diferentes APIs que está red social ofrece. Por último, repasaremos cuales son las aplicaciones y usos de la extracción de datos en redes sociales.

5.1. Creación de una herramienta para la extracción de datos

Una de las principales motivaciones del presente trabajo era la creación de una herramienta propia que nos permitiera explorar las opciones y posibilidades que ofrece la API de Twitter. Como veremos a lo largo de este punto es necesario llevar a cabo una serie de pasos previos antes de comenzar directamente con la extracción de datos. Lo

cual puede suponer distintas limitaciones, ya sean, las limitaciones propias que Twitter impone para la descarga de datos, así como las limitaciones temporales y de extensión del trabajo. Teniendo en cuenta dichas limitaciones vamos a explicar en que ha consistido la elaboración de esta herramienta.

Se tiene en cuenta que dependiendo del lenguaje de programación que se elija debemos trabajar con unas librerías u otras. En nuestro caso, *Tweepy* (https://www.tweepy.org/) es la librería por excelencia para la conexión con Twitter utilizando el lenguaje de programación *Python*, que como ya hemos mencionado antes, es el lenguaje que estaremos implementando en nuestra herramienta.

El objetivo específico de la herramienta es conseguir conectarnos a Twitter mediante unas credenciales creadas con este propósito. Dar la posibilidad al usuario de que interactúe con las búsquedas y haga él mismo sus propias consultas, y a partir de ellas, extraer un número significativo de tuits. Tras esto se llevará a cabo un procesamiento de los mismos, destacando aquellos campos que sean relevantes, para posteriormente volcarlos en dos bases de datos de diferente tipología. De esta forma, no limitaremos a una base de datos el almacenamiento y daremos más opciones de integración. Finalmente se facilita la visualización de los datos y se hace un pequeño análisis de sentimientos que mostraremos más adelante.
Lo primero que se ha tenido en cuenta es la instalación de un entorno de trabajo para *Python,* concretamente se ha utilizado el editor de código fuente *Thonny (https://thonny.org/)*.

Antes de empezar formalmente con la programación del código, era necesario obtener una cuenta de desarrollador en Twitter, la cual permite acceder a las APIs de Twitter.
Para ello es necesario contar previamente con una cuenta en Twitter, con la que necesitarás acceder para, después, contestar y justificar una serie de cuestiones relativas al uso que se pretende hacer tanto de las APIs como de los datos procedentes de Twitter.

Una vez que se ha procedido a realizar la solicitud, el tiempo de espera hasta saber si ha sido aprobada es de entorno a un día (si es rechazada puede ser más tiempo). También es importante mencionar que la solicitud es en inglés, por lo que puede dificultar el

proceso a algunos usuarios que no conocen el idioma. Se necesita también verificar el número de teléfono para poder seguir adelante. Un ejemplo de los preguntas que se deben contestar durante la solicitud sería el siguiente:

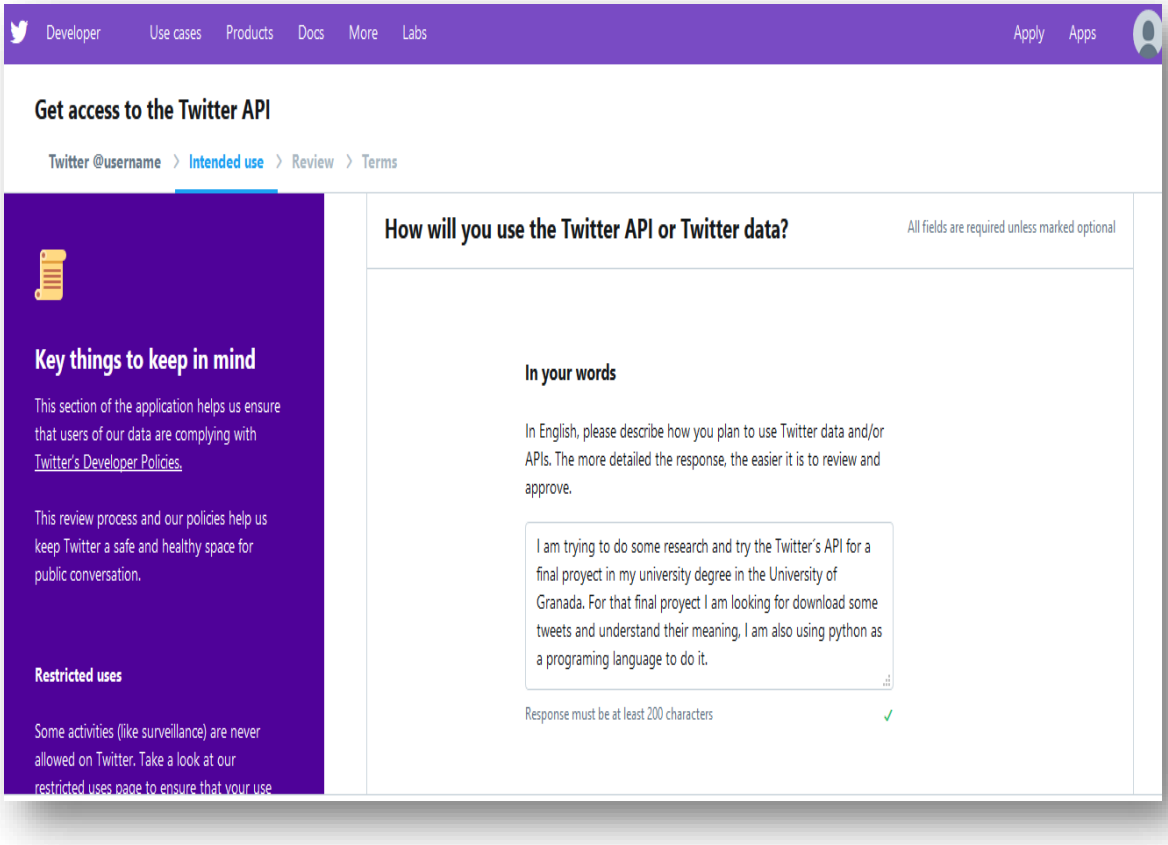

Figura 4: Ejemplificación del proceso de solicitud de la cuenta de desarrollador en Twitter.

Tras la aprobación de la solicitud se da paso a la creación de una aplicación, dentro de la cual se generarán unas credenciales que utilizaremos más adelante en el momento de la conexión con Twitter.

Las credenciales están divididas en:

- *Consumer API key*
- *Consumer API secret key*
- *Access token*
- *Access token secret*

Una vez generadas, únicamente habrá que pegarlas en el código y utilizar el método "OAuth" específico para realizar esta conexión.

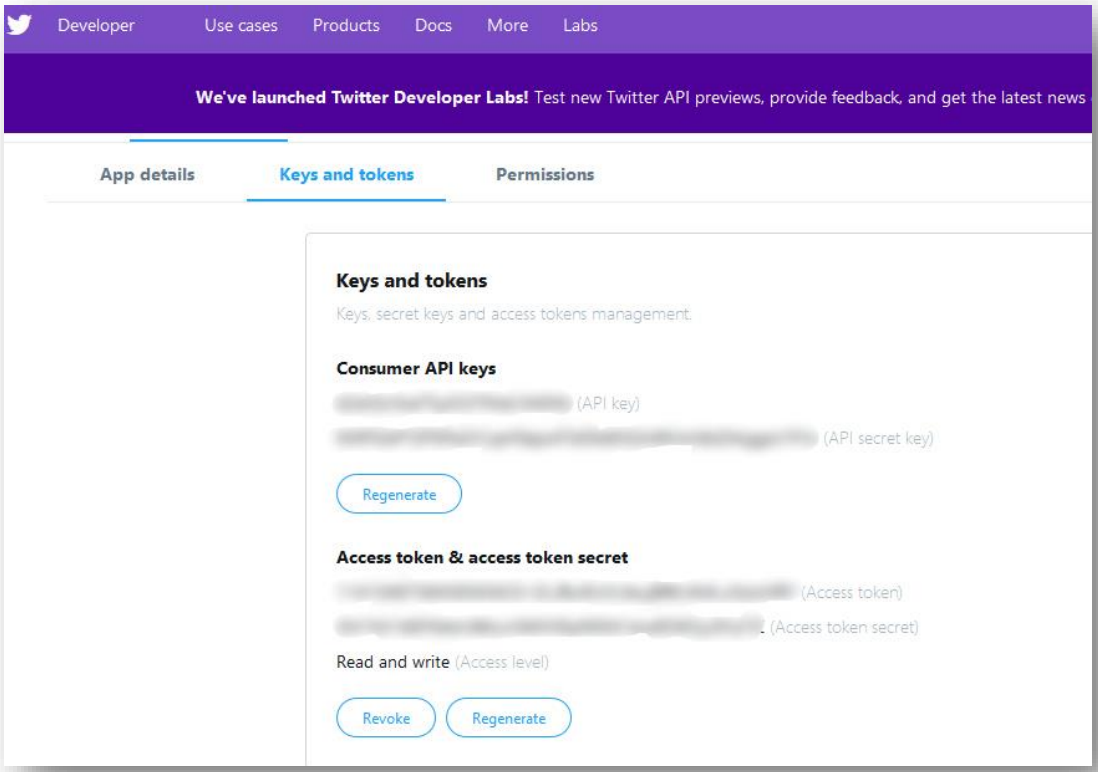

Figura 5: Credenciales para el uso de las APIs de Twitter

5.1.1.- Procesamiento de datos

El procesamiento de datos permite producir información relevante, lo cual implica la disminución de grandes cantidades de datos de carácter irrelevante para quedarse solo con aquellos de interés.

Partiendo de esta base se han estructurado los datos que ofrece Twitter, los cuales se encuentran en formato JSON (ver Figura 6), y que al ser solicitados a través de la API son devueltos de forma masificada y desestructurada por lo que existen una gran abundancia de símbolos y palabras que no admiten ningún tipo de interpretación (ver Figura 7), al menos a primera vista. Por eso era tan importante poder identificar y quedarse con aquellos campos que realmente aporten algo significativo al resto del proyecto. En la siguiente tabla se hace un pequeño repaso de los metadatos que están contenidos en los tuits extraídos de Twitter:

Etiquetas	Descripción
Id	Expresa la representación entera del identificador único para este Tweet. También se puede obtener en forma de cadena con la etiqueta id_str.
Created_at	Muestra la fecha y hora de publicación del tuit.
Text	El texto que contiene el tuit en formato de codificación UTF-8.
Source	Contiene la fuente desde la que ha sido enviado el tuit, ya sea desde un dispositivo móvil, el sistema operativo del mismo, o desde el sitio web.
Retweet_count	Número de veces que un tuit ha sido retuiteado. El número puede ir desde 0 hasta un potencial número infinito.
Favorite_count	Indica cuantas veces ha gustado a los usuarios de Twitter este tuit.
Entities	Son entidades localizadas fuera del texto del tuit. Aquí incluimos, los hashtags, menciones, símbolos y URLs, entre otros.

Tabla 2: Metadatos asociados a los tuits.

(Developer.twitter.com, 2019)

Esta ilustración muestra los datos de un tuit a partir de la búsqueda con el hashtag "Felizmartes" desde la Streaming AP:


```
{
  "created_at": "Tue Jul 02 01:40:21 +0000 2019",
  "id": 1145869788918493200,
  "id_str": "1145869788918493185",
  "text": "RT @TorresAren: YO ME PREGUNTO.\nPara que pagar impuestos\nEn una DICTADURA como\nLa CHAVISTA en VENEZUELA\n
    devuelva…",
  "source": "<a href=\"http://twitter.com/download/android\" rel=\"nofollow\">Twitter for Android</a>",
  "truncated": false,
  "user": {
    "id": 263734415,
    "id_str": "263734415",
    "name": "El Reporte Seguro",
    "screen_name": "reporteseguro1",
    "location": "Maracay, Venezuela",
    "url": null,
    "description": "Somos un medio integral, brindamos un sistema \nde comunicación directa hacia nuestros \nconsumidor
      y \ntelevisión.",
    "translator_type": "none",
    "protected": false,
    "verified": false,
    "followers_count": 1272,
    "friends_count": 3463,
    "listed_count": 18,
    "favourites_count": 445,
    "statuses_count": 15329,
    "created_at": "Thu Mar 10 16:51:15 +0000 2011",
    "geo_enabled": true,
    "lang": null,
```


Figura 6: Ejemplo de los datos de un tuit estructurados en formato JSON.

En esta, sin embargo, se observa la misma serie de tuits a partir del hashtag “Felizmartes” pero de forma desorganizada e ininteligible. Esto sería lo que nos devuelve el programa en un primer momento y con lo que hemos trabajado para estructurar y destacar la información más relevante.


```
{"created_at":"Wed Jul 03 15:47:57 +0000 2019","id":1146445481393811456,"id_str":"1146445481393811456","text":"RT @RosaMa_Hdez: \ud83d\udde3#For
oSNTE\ud83d\udcdc\n\n\ud83d\udd34 Atienden problem\u00e1tica educativa\u2705 estudiantil\u2705 magisterial\u2705 propuestas #SNTE sobre Leyes Se
cundarias\ud83d\udcd4 a @sen\u2026","source":"\u003ca href=\"http:\/\/twitter.com\/download\/android\" rel=\"nofollow\"\u003eTwitter for Android
\u003c\/a\u003e","truncated":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id
_str":null,"in_reply_to_screen_name":null,"user":{"id":1397598223,"id_str":"1397598223","name":"Arnoldo Limas R","screen_name":"ArnoldoLimas","l
ocation":null,"url":null,"description":"Secretario de Organizaci\u00f3n IV  Secci\u00f3n 8 SNTE","translator_type":"none","protected":false,"ver
ified":false,"followers_count":137,"friends_count":143,"listed_count":3,"favourites_count":1938,"statuses_count":2331,"created_at":"Thu May 02 1
6:26:52 +0000 2013","utc_offset":null,"time_zone":null,"geo_enabled":true,"lang":null,"contributors_enabled":false,"is_translator":false,"profil
e_background_color":"C0DEED","profile_background_image_url":"http:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_image_u
rl_https":"https:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_tile":false,"profile_link_color":"1DA1F2","profile_sideb
ar_border_color":"C0DEED","profile_sidebar_fill_color":"DDEEF6","profile_text_color":"333333","profile_use_background_image":true,"profile_image
_url":"http:\/\/pbs.twimg.com\/profile_images\/798029170487934976\/yRFn_vvC_normal.jpg","profile_image_url_https":"https:\/\/pbs.twimg.com\/prof
ile_images\/798029170487934976\/yRFn_vvC_normal.jpg","default_profile":true,"default_profile_image":false,"following":null,"follow_request_sent"
:null,"notifications":null},"geo":null,"coordinates":null,"place":null,"contributors":null,"retweeted_status":{"created_at":"Tue Jul 02 19:03:19
+0000 2019","id":1146132259935854593,"id_str":"1146132259935854593","text":"\ud83d\udde3#ForoSNTE\ud83d\udcdc\n\n\ud83d\udd34 Atienden problem\u
00e1tica educativa\u2705 estudiantil\u2705 magisterial\u2705 propuestas #SNTE sobre Leyes Secundarias\ud83d\udcd4\u2026 https:\/\/t.co\/L6tYs4YO
3J","display_text_range":[0,140],"source":"\u003ca href=\"http:\/\/twitter.com\/download\/android\" rel=\"nofollow\"\u003eTwitter for Android\u0
03c\/a\u003e","truncated":true,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str
":null,"in_reply_to_screen_name":null,"user":{"id":764634925609869312,"id_str":"764634925609869312","name":"Rosa Mar\u00eda Hern\u00e1ndez","scr
een_name":"RosaMa_Hdez","location":"Chihuahua, M\u00e9xico","url":"https:\/\/www.facebook.com\/Rosa-Mar%C3%ADa-Hern%C3%A1ndez-Madero-93912625288
0001\/","description":"Secretaria General de la Secci\u00f3n 8 del SNTE","translator_type":"none","protected":false,"verified":false,"followers_
count":704,"friends_count":87,"listed_count":5,"favourites_count":789,"statuses_count":1830,"created_at":"Sun Aug 14 01:29:07 +0000 2016","utc_o
ffset":null,"time_zone":null,"geo_enabled":true,"lang":null,"contributors_enabled":false,"is_translator":false,"profile_background_color":"00000
0","profile_background_image_url":"http:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_image_url_https":"https:\/\/abs.t
wimg.com\/images\/themes\/theme1\/bg.png","profile_background_tile":false,"profile_link_color":"FF691F","profile_sidebar_border_color":"000000",
"profile_sidebar_fill_color":"000000","profile_text_color":"000000","profile_use_background_image":false,"profile_image_url":"http:\/\/pbs.twimg
.com\/profile_images\/1091027443652182016\/3mskJFWW_normal.jpg","profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/109102744365
2182016\/3mskJFWW_normal.jpg","profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/764634925609869312\/1536783848","default_profile":
false,"default_profile_image":false,"following":null,"follow_request_sent":null,"notifications":null},"geo":null,"coordinates":null,"place":null
,"contributors":null,"is_quote_status":false,"extended_tweet":{"full_text":"\ud83d\udde3#ForoSNTE\ud83d\udcdc\n\n\ud83d\udd34 Atienden problem\u
00e1tica educativa\u2705 estudiantil\u2705 magisterial\u2705 propuestas #SNTE sobre Leyes Secundarias\ud83d\udcd4 a @senadomexicano\ud83c\udfeb\
n\n#SNTEconM\u00e9xico #FelizMartes M\u00e9xico #AMLOFest AMLO Oja\u00e1 #scelc19 Z\u00f3calo #sof\u00e1 El Senado #Opini\u00f3n legislaci\u00f
3n secundaria #Ecatepec Pemex https:\/\/t.co\/9wyEjFsacV","display_text_range":[0,271],"entities":{"hashtags":[{"text":"ForoSNTE","indices":[1,1
0]},{"text":"SNTE","indices":[85,90]},{"text":"SNTEconM\u00e9xico","indices":[136,150]},{"text":"FelizMartes","indices":[151,163]},{"text":"AMLO
Fest","indices":[171,180]},{"text":"scelc19","indices":[192,200]},{"text":"sof\u00e1","indices":[208,213]},{"text":"Opini\u00f3n","indices":[224
```


Figura 7: Conjunto de datos en formato JSON para un tuit.

Para este objetivo se importan dos librerías *Pandas(https://pandas.pydata.org/)* y *Numpy (*https://www.numpy.org/*)* que ayudan a la estructuración en columnas y filas de los tuits, además, utilizando algunas funciones podemos extraer los campos con los que vayamos a trabajar. Ya sean, el texto, la fecha de publicación, el número de me gustas, retuits, etc. Esto ayuda a visualizar mejor la información que se nos presenta y a poder trabajar con ella ya sea en el volcado de datos en una base de dates u otro tipo de tareas que veremos más adelante. Como podemos ver a continuación los datos se verían estructurados en filas y columnas. Las filas corresponden a cada tuit individualmente, y las columnas a los diferentes campos. (Ver Figuras 7 y 8).

	A	B
1	tweets	id
2	Iâ€™m admittedly biased, but this article gibes wit	1146460662832540000
3	Congrats to the USWNT! A great performance from	1146180070467390000
4	No one changes the world alone. Thatâ€™s why the	1146140164122690000
5	The most important job in our democracy is citizen	1145724532700980000
6	50 years ago, history was written at the Stonewall I	1144650730256680000
7	What a gift to come across this interview with Julia	1144336622814670000
8	Now weâ€™re talking! Congrats to Team USA for m	1142133432698450000
9	On Juneteenth, we celebrate our capacity to make	1141390520956190000
10	Outside the Oval Office, I kept a painting of a small	1141390367352400000
11	This is worth a read: a thought-provoking reminder	1141047405560940000

Figura 8: Vista de los datos estructurados procedentes de un tuit. Parte 1

C	D	E	F	G
len	date	source	likes	retweets
140	03/07/2019 16:48	Twitter for iPhone	47430	6336
105	02/07/2019 22:13	Twitter for iPhone	175839	16042
140	02/07/2019 19:34	Twitter for iPhone	28973	4565
140	01/07/2019 16:03	Twitter for iPhone	51085	12426
140	28/06/2019 16:56	Twitter for iPhone	181512	27046
139	27/06/2019 20:08	Twitter for iPhone	65494	6894
131	21/06/2019 18:13	Twitter Web Client	89664	8056
140	19/06/2019 17:01	Twitter for iPhone	159208	23585
140	19/06/2019 17:00	Twitter for iPhone	291112	48391
140	18/06/2019 18:17	Twitter for iPhone	46773	9254

Figura 9: Vista de los datos estructurados procedentes de un tuit. Parte 2.

5.1.2.- Almacenamiento de bases de datos

Una vez concluido el procesamiento de los tuits se procede al volcado de aquella información que hemos seleccionado como relevante, en una base de datos.

Aunque en nuestro caso se ampliamos esta tarea a dos bases de datos de diferente tipología. Una de tipo relacional como es SQLite y otra de tipo NoSQL orientada a documentos y de código abierto como es MongoDB, el cual es el mejor recurso para trabajar con JSON, que como ya hemos visto en el procesamiento es la forma que tiene Twitter de estructurar los datos.

De esta manera, nuestra herramienta permite el trabajar en diferentes entornos, sin

limitarse tan solo a uno de ellos. Esto también permite demostrar que existen diferentes maneras de tratar con la información extraída de las redes sociales y que las posibilidades son ilimitadas.

Para MongoDB se ha creado la base de datos “tweets_tfg”, la cual a su vez contiene una colección llamada “tweets”, en la que se guardan los tuits en forma de “documentos”. Cómo se puede ver en la siguiente Figura 9.

En el caso de SQLite se crea una base de datos llamada también “tweets_tfg”, con su respectiva tabla “tweets”, que contiene los diferentes campos que hemos extraído.

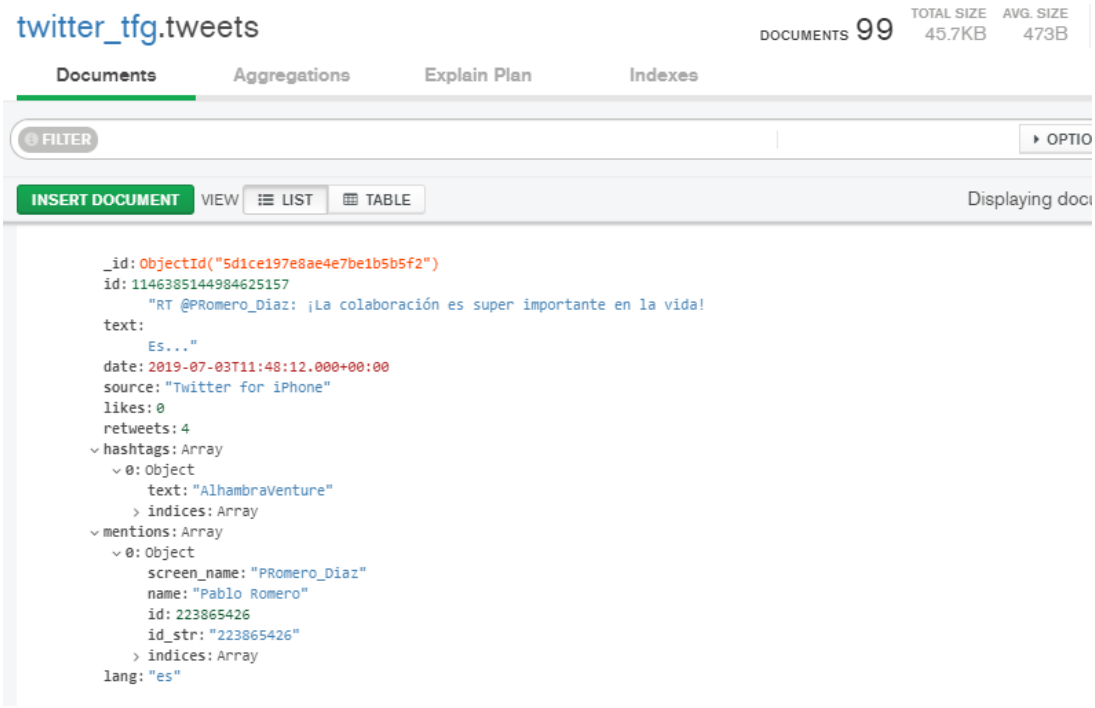

Figura 10: Demostración de almacenamiento en MongoDB

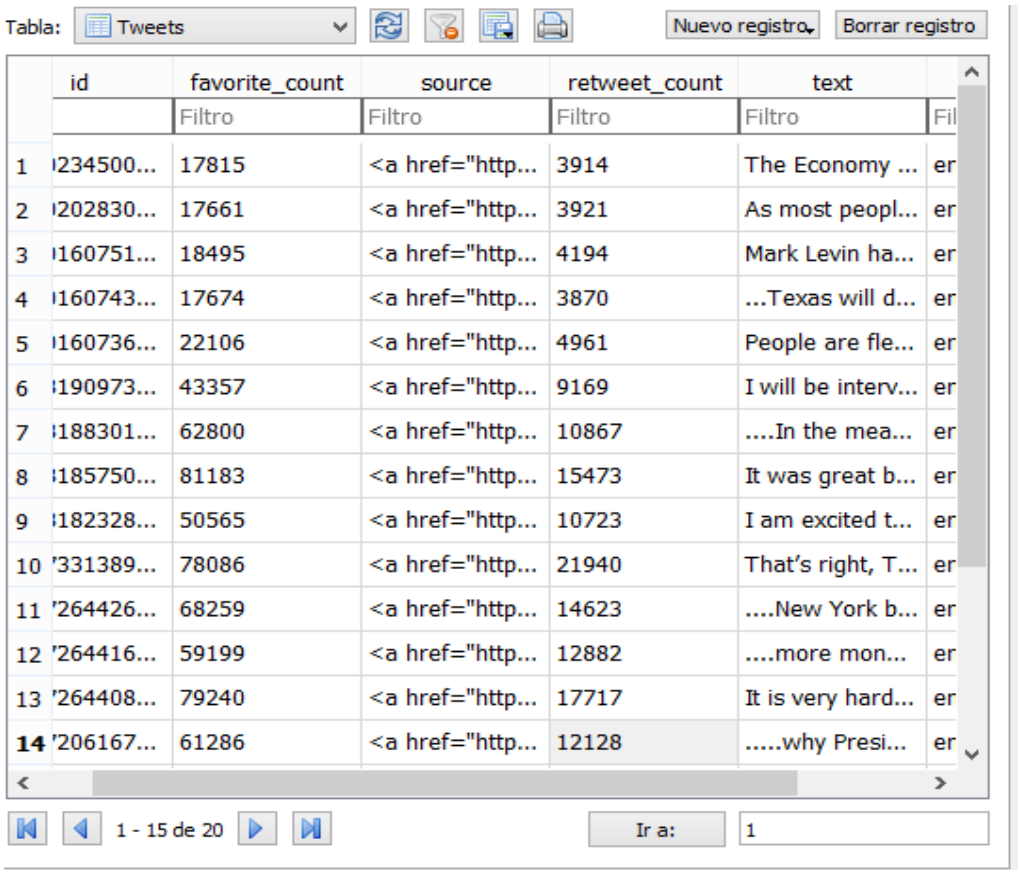

Figura 11: Demostración de almacenamiento en SQLite

5.2. Aplicaciones y usos de la extracción de datos en redes sociales

Desde el principio ha sido necesario preguntarnos hacia donde se puede dirigir la extracción de datos en redes sociales respecto a sus aplicaciones.
Existen diferentes contextos en los que la extcomo pueden ser la investigación social, la toma de decisiones, la formación de opinión política, la calidad de la información o los problemas de seguridad (Pfeffer, Mayer et al. 2018).

Aunque nosotros hablaremos más específicamente del caso que nos ocupa, la extracción de datos en Twitter. Es por ello, que en esta parte del trabajo reflexionamos sobre cuales son las aplicaciones y usos de la API de Twitter en el mundo profesional.

Debido a la repercusión que tiene Twitter como fuente de información, muchos periodistas y profesionales de la información acuden a Twitter para identificar o noticias o incluso como fuente de investigación. También son numerosas las compañías que

utilizan esta red para monitorizar y o promocionar sus marcas. Siendo esto así, la API de Twitter es la que permite al investigador no limitar sus búsquedas al filtrado por palabras clave o cuentas. Otro de sus grandes reclamos es el poder acceder a grandes cantidades de tweets históricos o para la detección de tendencias en tiempo real (Pfeffer, Mayer et al. 2018).

Actualmente y como ya hemos mencionado son numerosos los investigadores, periodistas y compañías que hacen uso de las tecnologías que la plataforma ofrece, en el caso de estas últimas utilizan el análisis de sentimientos para para detectar posibles campañas publicitarias. También son muchas las consultas políticas que se realizan para conocer el transcurso de unas elecciones, y todo tipo de situaciones derivadas de esto.

Existen numerosas aplicaciones que tienen su origen en el uso de la API de Twitter y que son en esencia bots, y es que un estudio, Varol *et al.* (2017) revela que entorno al 15% de los usuarios en Twitter son bots. Esto puede tener su motivación en que las corporaciones hagan más sencillas sus interacciones con sus clientes/usuarios mediante la utilización de estos recursos que permiten detectar determinadas situaciones en las que el usuario requiere de información sobre un tema en concreto y el bot es capaz de detectar que necesita y generar una respuesta acertada para cada situación.

Finalmente comentaremos algunos casos prácticos que pueden darse haciendo uso de la API de Twitter. Por ejemplo las empresas que prestan servicios conocen perfectamente que las redes sociales son uno de los primeros sitios donde el cliente acude a quejarse o dar su opinión respecto a un servicio. Descargar datos sobre los tuits que hagan mención de alguna marca o compañía puede ayudar a conocer cual es la opinión general de los usuarios sobre sus servicios y en última instancia poder mejorarlos u ofrecer soluciones a algunas peticiones. También ayuda a las empresas y emprendedores a detectar posibles nichos de mercado en los que centrar su atención. Incluso puede servir para saber que ha podido ir mal en una campaña publicitaria o cuales han sido las razones para que un usuario se haya dado de baja en sus servicios.

En esta sección presentamos los principales resultados que el programa desarrollado es capaz de generar. Como veremos a continuación hemos realizado una exportación de los

datos a XML, sugerido parámetros de análisis desde el punto de vista estadístico, demostrado las opciones de visualización que existen, y por último nos introducimos en el mundo del análisis de sentimientos describiendo un ejemplo de análisis entorno a una serie de tuits.

Una vez fueron procesados se ha programado la exportación de los mismos en formato XML. Cumpliendo así con una de las partes del planteamiento inicial del problema.

```
<?xml version=1.0 encoding=UTF-8?>
<tuits>
        <tuit>
                <id> 1146460662832541697 </id>
                <texto> I'm admittedly biased, but this article gibes with my experience about what all of us
                <fecha> 2019-07-03 16:48:16 </fecha>
                <fuente> Twitter for iPhone </fuente>
                <likes> 54783 </likes>
                <id> 7295 </id>
                <hashtags> [] </hashtags>
                <menciones> [] </menciones>
                <lenguage> en </lenguage>
        </tuit>
        <tuit>
                <id> 1146180070467391489 </id>
                <texto> Congrats to the USWNT! A great performance from a great team—looking forward to Sunday
                <fecha> 2019-07-02 22:13:18 </fecha>
                <fuente> Twitter for iPhone </fuente>
                <likes> 177282 </likes>
                <id> 16129 </id>
                <hashtags> [{'text': 'OneNationOneTeam', 'indices': [88, 105]}] </hashtags>
                <menciones> [] </menciones>
                <lenguage> en </lenguage>
        </tuit>
</tuits>
```

Figura 12: Exportación de datos procedentes de Twitter a formato XML

5.3. Análisis estadístico y visualización de los datos

Comentaremos en este apartado los resultados que se han obtenido relacionados con los datos estadísticos, algunos de ellos se verán representados para su visualización y mejor entendimiento.

Hemos podido extraer por ejemplo el número medio de caracteres que utiliza este usuario cada vez que envía un tuit. O por ejemplo también el número medio de me gustas que recibe por publicación.

Otros datos que se pueden observar es el tuit con mayor número de retuits, que en este caso correspondería con el primero de ellos. También sabemos que el tuit con mayor número de me gustas supera los cuatrocientos mil.

```
                                                tweets  ...  retweets
0  Happy Fourth of July, everybody! This is alway...  ...     49701
1  I'm admittedly biased, but this article gibes ...  ...      9951
2  Congrats to the USWNT! A great performance fro...  ...     16568
3  No one changes the world alone. That's why the...  ...      4848
4  The most important job in our democracy is cit...  ...     12760
5  50 years ago, history was written at the Stone...  ...     27124
6  What a gift to come across this interview with...  ...      6950
7  Now we're talking! Congrats to Team USA for mo...  ...      8070
8  On Juneteenth, we celebrate our capacity to ma...  ...     23622
9  Outside the Oval Office, I kept a painting of ...  ...     48433

[10 rows x 7 columns]
nº medio de caracteres:  135.4
nº medio de likes:  156394.8
El tweet con mayor número de likes tiene:  433936
El tweet con mayor número de retuits tiene:  49701
Tuits insertados en SQLite...
```

Figura 13: datos estadísticos para un conjunto de tuits

Si por ejemplo descargamos un número superior de tuits, como por ejemplo 200, como en este caso las cifras pueden también variar. Por lo que dependerá de cuantos datos se hayan extraído en el momento del estudio.

```
nº medio de caracteres:  134.445
nº medio de likes:  310606.495
El tweet con mayor número de likes tiene:  1943231
El tweet con mayor número de retuits tiene:  474700
```

Figura 14: datos estadísticos para un conjunto mayor de tuits

Por último representamos en forma de gráfico de líneas, utilizando la librería Matplotlib (https://matplotlib.org/), el número de me gustas y retuits en un período de tiempo. Podemos observar que cuando el número de me gustas aumenta, también lo hace aunque de forma mucho más leve el de retuits.

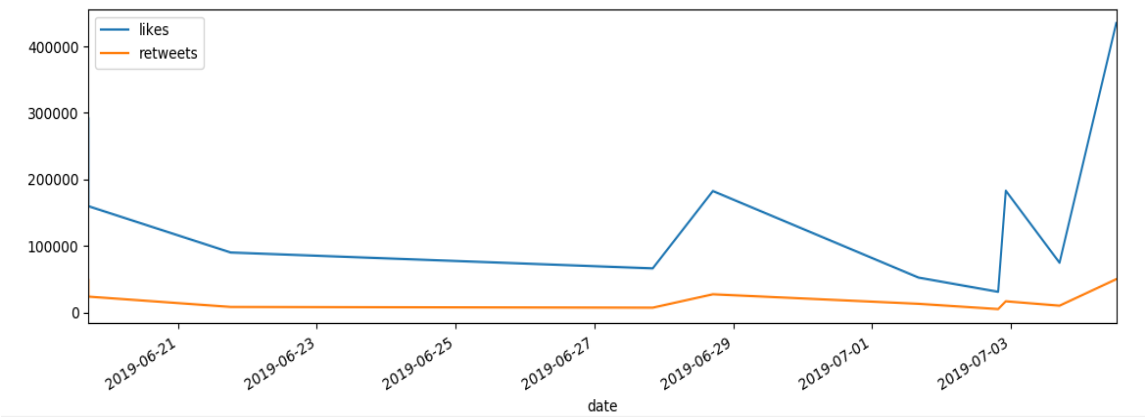

Figura 15: Número de me gustas y retuits en un período de tiempo

5.4. Análisis de sentimientos

Llegado este punto explicamos en que ha consistido el análisis llevado a cabo sobre un grupo de tuits provenientes de una cuenta concreta. Para empezar se ha necesitado hacer una limpieza de todos aquellas palabras, artículos o pronombres que pudieran intervenir en la tarea de procesamiento del texto. Esta limpieza se ha hecho utilizando una serie de expresiones regulares dentro del propio programa. También se ha utilizado un paquete TextBlob (https://textblob.readthedocs.io/en/dev/) que utiliza el lenguaje natural para el procesamiento del texto.

A continuación observamos los 10 últimos tuits obtenidos a través de la REST API de una cuenta de Twitter en inglés.

tweets	sentiment
Happy Fourth of July, everybody! This is always a great day in the Obama family: a chance to celebrate Americaâ€"andâ€	positivo
Iâ€™m admittedly biased, but this article gibes with my experience about what all of us might take away from the â€œa	neutro
Congrats to the USWNT! A great performance from a great teamâ€"looking forward to Sunday. #OneNationOneTeam	positivo
No one changes the world alone. Thatâ€™s why the @ObamaFoundation is connecting emerging leaders from South Afr	neutro
The most important job in our democracy is citizen. If youâ€™re tired of politicians manipulating maps and ignoring tâ€¦	positivo
50 years ago, history was written at the Stonewall Inn when New York's LGBT community stood up, spoke out, and starâ€	positivo
What a gift to come across this interview with Julia â€œHurricaneâ€ Hawkins. Iâ€™m as grateful for her life advice as Iâ€	neutro
Now weâ€™re talking! Congrats to Team USA for moving on, and thanks for continuing to make us all proud. #USA https:	positivo
On Juneteenth, we celebrate our capacity to make real the promise of our founding, that thing inside each of us thaâ€¦	positivo
Outside the Oval Office, I kept a painting of a small crowd huddled around a pocketwatch, waiting for the moment thâ€¦	negativo

Figura 16: análisis de sentimientos tuits

En ella podemos ver en la columna de la izquierda el texto de los tuits que se han recogidos, y a la derecha la valoración entorno a cada uno de los tuits. Diferenciamos en la misma entre positivos, negativos y neutros. Para realizar esta clasificación, el programa tiene en cuenta el tipo de palabras que existen en cada tuit, las detecta, las clasifica y enumera dependiendo de la cantidad de palabras de cada tipo que haya en el tuit. Finalmente las valora y decide si se trata de:

- Tuit positivo: con palabras como por ejemplo "feliz", "enhorabuena" o "celebración".
- Tuit neutral: o bien no detecta ni palabras positivas ni negativas o bien existen el mismo número de palabras de cada tipo.
- Tuit negativo: palabras de carácter negativo.

tweets	sentiment
Happy Fourth of July, everybody! This is always a great day in the Obama family: a chance to celebrate Americaâ€"andâ€	positivo
Iâ€™m admittedly biased, but this article gibes with my experience about what all of us might take away from the â€œea	neutro
Congrats to the USWNT! A great performance from a great teamâ€"looking forward to Sunday. #OneNationOneTeam	positivo
No one changes the world alone. Thatâ€™s why the @ObamaFoundation is connecting emerging leaders from South Afr	neutro
The most important job in our democracy is citizen. If youâ€™re tired of politicians manipulating maps and ignoring tâ€¦	positivo
50 years ago, history was written at the Stonewall Inn when New York's LGBT community stood up, spoke out, and starâ€	positivo
What a gift to come across this interview with Julia â€œHurricaneâ€ Hawkins. Iâ€™m as grateful for her life advice as Iâ€	neutro
Now weâ€™re talking! Congrats to Team USA for moving on, and thanks for continuing to make us all proud. #USA https:	positivo
On Juneteenth, we celebrate our capacity to make real the promise of our founding, that thing inside each of us thaâ€¦ I	positivo
Outside the Oval Office, I kept a painting of a small crowd huddled around a pocketwatch, waiting for the moment thâ€¦	negativo

Figura 17: Palabras positivas encontradas en el análisis

Con estos resultados se pretende demostrar que no se necesitan demasiados recursos para poner analizar las opiniones de los usuarios en Internet, pudiendo aplicarse como hemos explicado anteriormente en el entorno de cualquier empresa.

6.- CONCLUSIONES

Las principales conclusiones que podemos extraer de la realización de esta memoria de TFG es que tanto las redes sociales como todo lo que las rodea tienen un fuerte impacto

en la sociedad en la que vivimos. No es difícil darse cuenta de que la forma de comunicarse ha evolucionado hasta el punto de que una gran mayoría de personas vive conectada a la red a diario o tiene alguna cuenta en una red social. Esto se traduce en vastas cantidades de información procedentes de personas de todo el mundo. Lo que hace necesario el hecho de ser capaz de aprender a gestionarla y analizarla con diferentes fines.

Otra conclusión paralela al trabajo es que existen numerosas herramientas para la extracción automática de datos, y que muchas de ellas están siendo explotadas diariamente, por lo que poder conocerlas y trabajar con ellas es fundamental desde el punto de vista profesional y puede ser aplicado de muchas maneras y con diferentes propósitos. También pueden ser de vital importancia para el/la profesional de la información. Ya que hace que su perfil no se reduzca únicamente al del sector público, abarcando de esta manera nuevas áreas de trabajo que cada vez más se abren a recibirlos.

Destacar el hecho de haber formado parte de la primera promoción en la mención de Gestión de la Información en la Web con la que he podido adquirir muchos conocimientos que han sido de gran importancia para la realización de este trabajo.

A modo de resumen, esta tabla incluye los conocimientos y competencias generales que se han puesto en práctica en la realización de este TFG:

Asignaturas	Competencias generales
Archivos electrónicos	**Estructuración y tratamiento de datos XML.**
Bases de datos	Estructuración y almacenamiento de datos.
Estadística	Conocimientos básicos en estadística descriptiva
Fundamentos de informática	Conceptos de fichero y sistemas operativos.

Fundamentos de programación	Puesta en práctica del lenguaje de programación Python y de la mayoría de los conceptos aprendidos en la asignatura. Y ampliación del conocimiento sobre la misma.
Inglés	Consulta de documentos en inglés y redacción del resumen.
Metodología de la investigación en información y documentación	Aplicación de los conocimientos aprendidos en la asignatura. Y organización del trabajo entorno a la misma.
Recursos de Información	Conceptos de dato e información
Técnicas avanzadas de recuperación y representación de la información	Procesamiento automático de la información, y recuperación. Visualización de la información.
Técnicas de recuperación de información	Conceptos básicos sobre recuperación de información.
Tratamiento masivo de datos	Conexión a bases de datos desde programación. Uso de sistemas distribuidos basados en tecnologías web, en nuestro caso la API de Twitter.

Tabla 3: Competencias y asignaturas en relación al TFG

7.1.- Trabajos futuros

Para trabajos futuros lo principal sería poder continuar con el desarrollo de la herramienta y/o mejorar y añadir las propuestas que se desarrollan a continuación:

- Modelar mejor los datos, sobre todo en el caso de MongoDB.
- Hacer posible un modelo entidad / relación más elaborado añadiendo nuevas tablas y creando relaciones entre ellas.
- La exportación y/o interoperabilidad a otros formatos que permitan enviar los datos a otro software, con el que establecer nuevas funciones y casos de estudio.

- También sería interesante la exploración de otros modelos más avanzados de datos como las bases de datos en red, sería el caso por ejemplo de Neo4j[2].
- Análisis más elaborados de los datos extraídos y mayor profundización en la visualización de la información.

[2] https://neo4j.com/

BIBLIOGRAFÍA

Bahillo, L. (2019). *Historia de Internet: ¿cómo nació y cuál fue su evolución?*. [online] Marketing 4 Ecommerce - Tu revista de marketing online para e-commerce. Available at: https://marketing4ecommerce.net/historia-de-internet/ [Accessed 5 May 2019].

BBVAOpen4U. (2019). *Herramientas de extracción de datos: para principiantes y profesionales*. [online] Available at: https://bbvaopen4u.com/es/actualidad/herramientas-de-extraccion-de-datos-para-principiantes-y-profesionales [Accessed 19 Jun. 2019].

Blog.twitter.com. (2017). *Tuitear más fácil*. [online] Available at: https://blog.twitter.com/official/es_es/topics/product/2017/280caracteres.html [Accessed 20 May 2019].

Calderón Maldonado, A. and Ibarra Orozco, R. (2015). *Metodologías paraanálisis político utilizando Web Scraping*. [ebook] Chiapas: Research in Computing Science 95. Available at: http://www.rcs.cic.ipn.mx/rcs/2015_95/Metodologias%20para%20analisis%20politico%20utilizando%20Web%20Scraping.pdf [Accessed 16 Jun. 2019].

Civico Cabrera, J. (2017). *Tipos de Redes Sociales*. [online] Web App Design. Available at: https://webappdesign.es/tipos-de-redes-sociales/ [Accessed 12 May 2019].

Congosto, M. (2019). *Lo que siempre quiso saber del API de Twitter y nunca se atrevió a preguntar (actualizado en 2017) - Barriblog*. [online] Barriblog. Available at: https://www.barriblog.com/2017/10/lo-siempre-quiso-saber-del-api-twitter-nunca-se-atrevio-preguntar-actualizado-2017/ [Accessed 20 Jun. 2019].

de Haro, J. J. (2010). Redes Sociales en Educación. [En línea] Available at: http://www.cepazahar.org/recursos/pluginfile.php/6425/mod_resource/content/0/redes_sociales_educacion.pdf [Accessed 7 May 2019].

Developer.twitter.com. (2019). *Docs*. [online] Available at:

https://developer.twitter.com/en/docs [Accessed 25 Jun. 2019].

Egea, I. (2007). *Twitter*. [online] Wikipedia Available at: https://es.wikipedia.org/wiki/Twitter [Accessed 19 May 2019].

Farisori (2019). *OpenSocial*. [online] Wikipedia. Available at: https://es.wikipedia.org/wiki/OpenSocial [Accessed 20 Jun. 2019].

Go4it.solutions. (2019). *Diferencias entre API y servicio Web | Go4IT Solutions*. [online] Available at: https://go4it.solutions/es/blog/diferencias-entre-api-y-servicio-web [Accessed 15 May 2019].

Hedleyv, J. (2019). *Jsoup*. [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/Jsoup [Accessed 23 Jun. 2019].

Jarould. (2019). *Etiqueta (internet)*. [online] Wikipedia Available at: https://es.wikipedia.org/wiki/Etiqueta_(internet) [Accessed 22 May 2019].

Kizar (2019). *Beautiful Soup*. [online] Wikipedia. Available at: https://es.wikipedia.org/wiki/Beautiful_Soup [Accessed 27 Jun. 2019].

Lorenz, C. (2010). *Definición de Red social*. [online] Definición ABC. Available at: https://www.definicionabc.com/social/red-social.php [Accessed 7 May 2019].

Martí, M. (2019). *Qué es el Web scraping? Introducción y herramientas*. [online] Sitelabs. Available at: https://sitelabs.es/web-scraping-introduccion-y-herramientas/ [Accessed 25 Jun. 2019].

PÉREZ SALAZAR, G., 2011. La Web 2.0 y la sociedad de la información. Revista mexicana de ciencias políticas y sociales, 56(212), pp. 57-68

PFEFFER, J., MAYER, K. and MORSTATTER, F., 2018. Tampering with Twitter's Sample API. *EPJ Data Science*, **7**(1), pp. 50.

Ponce, I. (2012). *MONOGRÁFICO: Redes Sociales - Definición de redes sociales | Observatorio Tecnológico*. [online] Recursostic.educacion.es. Available at: http://recursostic.educacion.es/observatorio/web/eu/internet/web-20/1043-redes-sociales?start=1 [Accessed 6 May 2019].

Raffino, M. (2019). *Red Social: Concepto, Tipos, Evolución y Aspectos negativos*. [online] Concepto.de. Available at: https://concepto.de/redes-sociales/ [Accessed 6 May 2019].

Tadeo Hernández, A., Gómez Vázquez, E., Berdejo Rincón, C., Montero García, J., Calderón Maldonado, A. and Ibarra Orozco, R. (2015). Metodologías paraanálisis políticoutilizando Web Scraping. [ebook] Chiapas: Research in Computing Science 95. Available at: http://www.rcs.cic.ipn.mx/rcs/2015_95/Metodologias%20para%20analisis%20politico%20utilizando%20Web%20Scraping.pdf [Accessed 16 Jun. 2019].

Tadeo Hernández, A., Gómez Vázquez, E., Berdejo Rincón, C., Montero García, J., Calderón Maldonado, A. and Ibarra Orozco, R. (2015). Metodologías paraanálisis políticoutilizando Web Scraping. [ebook] Chiapas: Research in Computing Science 95. Available at: http://www.rcs.cic.ipn.mx/rcs/2015_95/Metodologias%20para%20analisis%20politico%20utilizando%20Web%20Scraping.pdf [Accessed 16 Jun. 2019].

Vanden Broucke, S. & Baesens, B. (2018) Practical Web Scraping for Data Science Best Practices and Examples with Python . [Online]. Berkeley, CA: Apress.

VAROL, O., FERRARA, E., DAVIS, C.A., MENCZER, F. and FLAMMINI, A.,(2017). Online human-bot interactions: Detection, estimation, and characterization, *Eleventh international AAAI conference on web and social media* 2017.

We Are Digital. (2019). "Digital 2019 - GLOBAL REPORT," 221. https://datareportal.com/.

ANEXOS

A. CÓDIGO EN PYTHON

```python
from tweepy import API
from tweepy import Cursor
from tweepy import OAuthHandler
from pymongo import MongoClient

from textblob import TextBlob

import simplejson
import json
import sqlite3
import twitter_credenciales1
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import re


### PROGRAMA PARA LA DESCARGA DE DATOS EN TWITTER ###
# Autora: Miriam Carrasco Alanis
# Tutor: Antonio Gabriel López Herrera
# Trabajo Fin de Grado en Información y Documentación, julio 2019

MONGO_HOST= "mongodb://localhost:27017/"
```

```python

MONGO_HOST= "mongodb://localhost:27017/"


# # # # TWITTER CLIENTE # # # #
class TwitterClient():
    def __init__(self, twitter_user=None):
        self.auth = TwitterAuthenticator().authenticate_twitter_app()
        self.twitter_client = API(self.auth)

        self.twitter_user = twitter_user

    def get_twitter_client_api(self):
        return self.twitter_client

    def get_user_timeline_tweets(self, num_tweets):
        tweets = []
        for tweet in Cursor(self.twitter_client.user_timeline, id=self.twitter_user).items(num_tweets):
            tweets.append(tweet)
        return tweets

#    def get_friend_list(self, num_friends):
#        friend_list = []
#        for friend in Cursor(self.twitter_client.friends, id=self.twitter_user).items(num_friends):
#            friend_list.append(friend)
#        return friend_list

```

```
#        return friend_list

#    def get_home_timeline_tweets(self, num_tweets):
#        home_timeline_tweets = []
#        for tweet in Cursor(self.twitter_client.home_timeline, id=self.twitter_user).items(num_tweets):
#            home_timeline_tweets.append(tweet)
#        return home_timeline_tweets


# # # # TWITTER AUTHENTICATER # # # #
class TwitterAuthenticator():

    def authenticate_twitter_app(self):
        auth = OAuthHandler(twitter_credenciales1.CONSUMER_KEY, twitter_credenciales1.CONSUMER_SECRET)
        auth.set_access_token(twitter_credenciales1.ACCESS_TOKEN, twitter_credenciales1.ACCESS_TOKEN_SECRET)
        return auth



class TweetAnalyzer():

    def clean_tweet(self, tweet):
        return ' '.join(re.sub("(@[A-Za-z0-9]+)|([^0-9A-Za-z \t])|(\w+:\/\/\S+)", " ", tweet).split())

    def analyze_sentiment(self, tweet):
        analysis = TextBlob(self.clean_tweet(tweet))
```

```
    def analyze_sentiment(self, tweet):
        analysis = TextBlob(self.clean_tweet(tweet))

        if analysis.sentiment.polarity > 0:
            return "positivo"
        elif analysis.sentiment.polarity == 0:
            return "neutro"
        else:
            return "negativo"

    def tweets_to_data_frame(self, tweets):
        df = pd.DataFrame(data=[tweet.text for tweet in tweets], columns=['tweets'])

        df['id'] = np.array([tweet.id for tweet in tweets])
        df['len'] = np.array([len(tweet.text) for tweet in tweets])
        df['date'] = np.array([tweet.created_at for tweet in tweets])
        df['source'] = np.array([tweet.source for tweet in tweets])
        df['likes'] = np.array([tweet.favorite_count for tweet in tweets])
        df['retweets'] = np.array([tweet.retweet_count for tweet in tweets])
        #df['entities'] = np.array([tweet.entities['hashtags'] for tweet in tweets])

        # # GUARDAR EN MONGODB # #
        print("Conexión con mongoDB ...")
        client = MongoClient(MONGO_HOST)
        print("Conectada!")
        print("Abriendo base de datos ...")

        print("Abriendo base de datos ...")

        db = client.twitter_tfg
        collection = db.tweets
        tuits = client.twitter_tfg.tweets
        for tweet in tweets:
            tuit = {}
            tuit["id"] = tweet.id
            tuit["text"] = tweet.text
            tuit["date"] = tweet.created_at
            tuit["source"] = tweet.source
            tuit["likes"] = tweet.favorite_count
            tuit["retweets"] = tweet.retweet_count
            tuit["hashtags"] = tweet.entities['hashtags']
            tuit["mentions"] = tweet.entities['user_mentions']
            tuit["lang"] = tweet.lang
            try:
                collection.insert_one(tuit)
                print ("se ha insertado un nuevo tweet")

            except Exception as e:
                print (e)

        ### FORMATO XML ###
        print("  ================================================")
        print ("<?xml version=1.0 encoding=UTF-8?>")
```

```
        ### FORMATO XML ###
        print("  ================================================")
        print ("<?xml version=1.0 encoding=UTF-8?>")
        print ("<tuits>")
        for tweet in tweets:

            print ("\t<tuit>")
            print ("\t\t<id>",tweet.id,"</id>")
            print ("\t\t<texto>",tweet.text, "</texto>")
            print ("\t\t<fecha>",tweet.created_at, "</fecha>")
            print ("\t\t<fuente>",tweet.source,"</fuente>")
            print ("\t\t<likes>",tweet.favorite_count,"</likes>")
            print ("\t\t<id>",tweet.retweet_count,"</id>")
            print ("\t\t<hashtags>",tweet.entities['hashtags'],"</hashtags>")
            print ("\t\t<menciones>",tweet.entities['user_mentions'],"</menciones>")
            print ("\t\t<lenguage>",tweet.lang,"</lenguage>")
            print ("\t</tuit>")

        print ("</tuits>")
        print("  ================================================")
        print("\n")

        return df

if __name__ == '__main__':
```

```
if __name__ == '__main__':



    twitter_client = TwitterClient()

    tweet_analyzer = TweetAnalyzer()

    api = twitter_client.get_twitter_client_api()

    screen_name = input ("Introduce el nombre de usuario que desees consultar: ")

    tweets = api.user_timeline(screen_name, count=10)

    df = tweet_analyzer.tweets_to_data_frame(tweets)
    df['sentiment'] = np.array([tweet_analyzer.analyze_sentiment(tweet) for tweet in df['tweets']])

    print(df.head(10))


    df.to_csv('prueba.txt', header=True, index=False, sep='\t', mode='a')


    # Obtención del número medio de caracteres por tweet
    print("nº medio de caracteres: ", np.mean(df['len']))
    # Obtención del número medio de likes
```

```
    # Obtención del número medio de caracteres por tweet
    print("nº medio de caracteres: ", np.mean(df['len']))
    # Obtención del número medio de likes
    print("nº medio de likes: ", np.mean(df['likes']))
    # Obtención del tweet con mayor número de likes
    print("El tweet con mayor número de likes tiene: ", np.max(df['likes']),)
    # Obtención del tweet con mayor número de retweets
    print("El tweet con mayor número de retuits tiene: ", np.max(df['retweets']))



    # Representación del nº de likes según la fecha
    #time_favs = pd.Series(data=df['likes'].values, index=df['date'])
    #time_favs.plot(figsize=(16, 4), color='r')
    #plt.show()

    # Representación del nº de retweets según la fecha
    #time_retweets = pd.Series(data=df['retweets'].values, index=df['date'])
    #time_retweets.plot(figsize=(16, 4), color='r')
    #plt.show()

    # Comparación del nº de likes con el nº de retweets según la fecha
    time_likes = pd.Series(data=df['likes'].values, index=df['date'])
    time_likes.plot(figsize=(16, 4), label="likes", legend=True)

    time_retweets = pd.Series(data=df['retweets'].values, index=df['date'])
```

```
    time_retweets = pd.Series(data=df['retweets'].values, index=df['date'])
    time_retweets.plot(figsize=(16, 4), label="retweets", legend=True)
    plt.show()


# # GUARDAR EN BD SQLITE # #

db = sqlite3.connect('tweets_tfg.sqlite')
cursor = db.cursor()

cursor.execute('DROP TABLE IF EXISTS Tweets')


cursor.execute('''CREATE TABLE Tweets (id, favorite_count, source, text, retweet_count, lang)''')

for status in tweets:

    cursor.execute('''SELECT *
                FROM Tweets
                WHERE Tweets.id = {0}'''.format(status._json["id"]))
    row = cursor.fetchone()
    if row is None:

        cursor.execute('''INSERT INTO Tweets(id, favorite_count, source, text, retweet_count, lang)
            VALUES ({0}, {1}, '{2}', "{3}", {4}, "{5}")'''.format(status._json['id'], status._json['favorite_count'], status._json['source'],
                                                    status._json['text'], status._json['retweet_count'], status._json['lang']))
```

```
    else:
        print("""El tweet con identificador = {0} ya
                está en la base de datos!!!""".format(status._json["id"]))

print ("Tuits insertados en SQLite...")

db.commit()

# Cerrar la base de datos ...
cursor.close()


```