INSTITUTO POLITÉCNICO NACIONAL

Unidad Profesional
Interdisciplinaria en Ingeniería y
Tecnologías Avanzadas

Academia Telemática

PROYECTO TERMINAL I

"Sistema de análisis para la ponderación de usuarios de Twitter en temas de tendencia"

Presentan:

C. Ortiz Romero Héctor Arturo

C. Sánchez López Alan Rodrigo

Asesoras:

Dra. Obdulia Pichardo Lagunas

Dra. Bella Citlali Martínez Seis

Noviembre 2018

C O N T E N I D O

Índice

Índice de ilustraciones

Índice de tablas

Resumen

Este proyecto tiene como finalidad realizar la ponderación y evaluación de la relación que existe entre temas de tendencia, fuente y líderes de opinión en la red social Twitter. A través de un análisis temporal de los temas de tendencia y la obtención de los líderes de opinión mediante el uso de procesamiento de lenguaje natural, minería de datos y análisis de grafos.

Palabras clave: Twitter, grafos, minería de datos, aprendizaje automático.

Abstract

The purpose of this project is to assess and evaluate the relationship between trends, sources and opinion leaders in the social network Twitter. Through a temporal analysis of the themes of tendency and the obtaining of opinion leaders using natural language processing, data mining and graph analysis.

Keywords: Twitter, graphs, data mining, machine learning.

Capítulo 1. Introducción

En los últimos años las redes sociales se han convertido en un importante medio de comunicación.

Hoy en día casi todo el mundo que disponga de conexión a Internet tiene un perfil en alguna de ellas. La mayoría las usa activamente, compartiendo un flujo de información importante que crece cada día de manera exponencial. Además de darle un uso cotidiano a este servicio, como puede ser contactar con amigos o compartir información con el resto del mundo, se puede ir más lejos y sacarles partido a todos esos datos. Haciendo uso de técnicas de minería de datos se puede extraer información útil del contenido que se genera en la red social Twitter.

El símbolo #, conocido como *hashtag*, es utilizado en redes sociales para marcar las palabras clave de las publicaciones. Enriqueciéndose en la web, el *hashtag* es una forma de etiquetar o clasificar los

mensajes de Twitter, de tal forma que se puedan agrupar alrededor de un tema en común. Si es una frase ésta tiene que escribirse junta y sin espacios. Un seguidor en Twitter es un usuario que recibe las actualizaciones de los estados que publicas, entonces cuantos más seguidores tenga un usuario mayor alcance tendrán las publicaciones que emita.

El estudio Factores que inciden en la variación de seguidores [1] nos indica como es la distribución de géneros en Twitter esta información se presenta en forma gráfica en la ilustración 1.

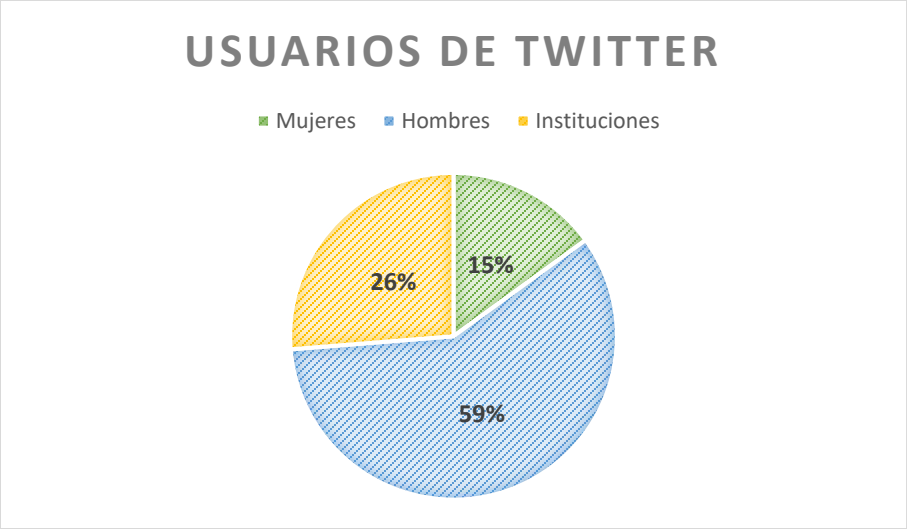


Ilustración 1 Grafica de la distribución de usuarios en Twitter

También nos indica que porcentaje de los usuarios prefieren mantener sus perfiles privados como se muestra en la ilustración 2.

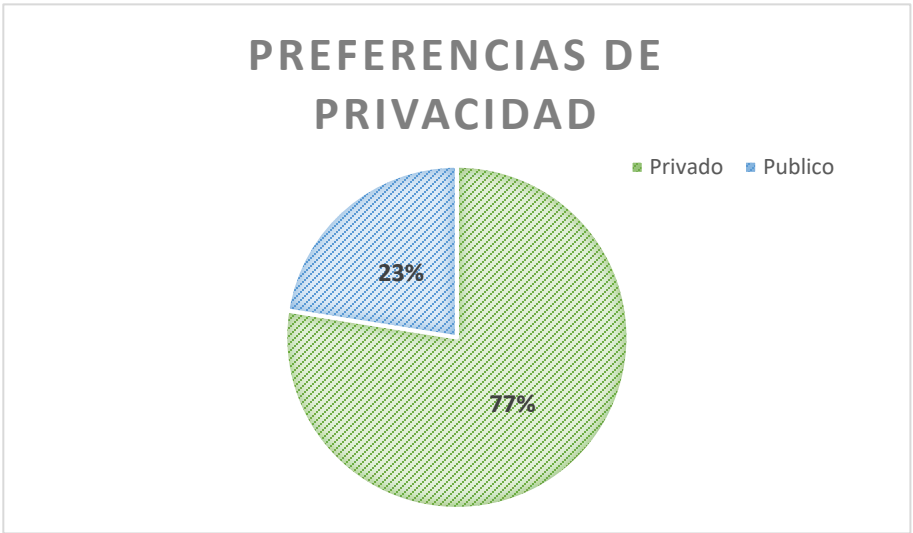


Ilustración 2. Preferencias de privacidad de los usuarios en Twitter

En la ilustración 3 se muestra a los usuarios de Twitter por su ocupación.

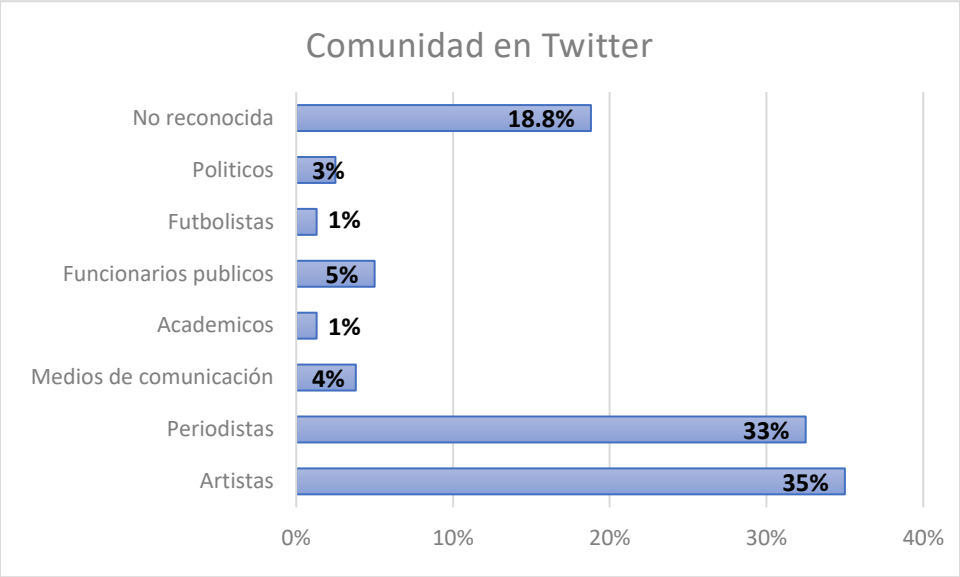


Ilustración 3. Tipos de usuarios en Twitter

En México hay 35.3 millones de usuarios mensuales activos de Twitter, de los cuales aproximadamente el 60% de las cuentas activas se encuentran en la Ciudad de México [2]. Debido a esta gran cantidad de usuarios el flujo de información en la plataforma es muy grande, lo cual ofrece la posibilidad de obtener datos que no se podrían conocer a simple vista, para hacerlo es necesario aplicar técnicas de minería de datos y aprendizaje automático.

Por lo cual este proyecto tiene como finalidad estudiar la difusión de la información en la red social Twitter, analizando los temas de tendencia y los usuarios que participan en ellos, además de ver las relaciones entre estos usuarios, es decir, de quién son seguidores y quiénes son sus seguidores en la red social. Utilizando técnicas de aprendizaje automático para la selección, filtrado y transformación y minería de datos para el análisis de dicha información.

Resulta importante saber quiénes son los líderes de opinión en los temas de tendencia debido a la influencia que tienen sobre los demás usuarios en la red social, ya que su posición como líder de opinión puede impulsar de manera positiva el lanzamiento de un nuevo producto

o servicio, promocionar un determinada acción o marca, así como dar a conocer un evento.

Planteamiento del problema

Existen temas que llegan a ser tendencia y que son de utilidad para diversos fines. En marketing, rastrear el origen de la información es de utilidad para encontrar aquellos miembros que influyen en la difusión de la información por lo que establecer contacto con ellos permitiría llegar a los clientes potenciales dependiendo de una clasificación previamente obtenida.

Los líderes de opinión no sólo son aquellos que tienen más número de seguidores, sino además de los que surgen las ideas. Por ejemplo, el miembro más importante de un grupo criminal no suele ser el que tiene la mayor cantidad de contactos ya que vulneraría su seguridad, pero seguramente tiene conexión con el elemento que tiene más enlaces.

El rastreo de información y de los líderes de opinión es un campo de estudio abierto en busca de una estrategia táctica. Ya que, en la actualidad, existen diversos trabajos que se encargan de analizar la información generada en las redes sociales, sin embargo, la mayoría se enfocan en analizar el texto para conocer la opinión de los usuarios acerca de productos, política o eventos sociales [4] [5], por lo que se detecta que existe la necesidad de un estudio que relacione la información y los elementos que la generan.

Es importante saber quién da origen a un tema de tendencia, para conocer si es un líder de opinión, si es *spam*, si surge de manera natural o, si fue pagado o impuesto, esto indica la intención con la que se publicó tal información. Conocer estos datos resulta útil en marketing porque permite medir la difusión de la información, que líderes de opinión impulsan la difusión de ciertos temas y con qué propósito.

Por ello, es de interés estudiar la difusión de la información en la red social Twitter, analizando los temas de tendencia y los usuarios que participan en ellos, además de ver las relaciones entre estos usuarios es decir de quién son seguidores y quiénes son sus seguidores en la red social.

Por lo señalado anteriormente surge la siguiente pregunta de investigación:

¿Cómo desarrollar un sistema que permita identificar a los líderes de opinión en los temas de tendencia en la Ciudad de México y que además posibilite el rastreo del origen de la información vertida en esos temas?

Propuesta de solución

Teniendo en cuenta que ya existen varios trabajos relacionados con la clasificación de tuits a partir del análisis de texto, esta solución se enfoca en analizar la relación que existe entre los usuarios y los tuits haciendo un análisis estructural de la red.

Se logrará a través de un estudio de los tuits relacionados con cada tema de tendencia. En el protocolo de registro se propuso que se descargasen los tuits en tiempo real por medio de robots programados para obtener dos bases de datos, una con todos los tuits publicados en la zona de estudio y la otra filtrando los tuits por tema de tendencia. Sin embargo, al comenzar con las descargas se hizo evidente que los tuits en tiempo real no aportan los datos necesarios para el estudio, puesto que al ser descargados un instante después de su publicación, no tienen conteo de favoritos ni *retweets*. Por esta razón los tuits se descargan al final del día, usando como filtro un *hashtag* o una frase que distinga a un tema de tendencia y la zona geográfica donde se estudian los temas de tendencia. Posteriormente se normaliza el texto para su fácil manipulación y para poder identificar que usuario lo emitió, en qué momento, cuantos seguidores tiene y que efecto causo su participación en el tema. Así se podrá identificar que usuarios son los más influyentes en cada tema de tendencia y cuál es el origen de la información.

Este proceso se divide en 5 etapas:

Extracción

El proceso de extracción se llevará a cabo en la red social Twitter desde la cual se obtendrán dos tipos de datos: texto e información de usuario. La primera se refiere a los tuits de los usuarios en la

Ciudad de México y la segunda a los seguidores y *retweets* que obtuvo cada tuit. La extracción será hará por medio de un script que busque los temas de tendencia por día para proceder a descargar los tuits por tema de tendencia. Esta información se irá almacenando en una base de datos de tipo NoSQL.

Filtrado

De la base de datos ya obtenida, extraeremos los atributos claves como “id de usuario, id de tuit, fecha de publicación”, entre otros. Teniendo que normalizar el texto para poder hacer una representación matricial del texto, sin perder el contexto de la información.

Transformación de datos

Después de tener el texto filtrado, se obtendrán los nombres de los usuarios que originaron cada tuit, el número de *retweets* y los seguidores de cada usuario. Creando un grafo de todos los usuarios que participan en los temas de tendencia. De forma paralela se generará un grafo por tema con todos los tuits que participaron. Estos subgrafos serán almacenados en una base de datos de tipo NoSQL.

Análisis

En la etapa de análisis se distinguen 3 módulos:

- Obtención del origen de la información, se llevará a cabo mediante algoritmos de difusión en grafos que ayuden a identificar el origen de la información dentro del subgrafo
- Obtención de líderes de opinión, este proceso se realizará mediante el análisis de la centralidad de cada nodo, se considera que un nodo es prominente si sus enlaces hacen que este nodo sea particularmente visible para los demás nodos del subgrafo.
- Clasificación, consiste en seleccionar y agrupar a los líderes de opinión y la fuente de información por cada subgrafo que se haya analizado.

Presentación

Los resultados obtenidos se presentarán a través de una interfaz web mostrando los líderes de opinión y fuente de cada tema de tendencia analizado.

Todo el procedimiento se representa con sus 5 etapas a través del diagrama de bloques presentado en la ilustración 4.

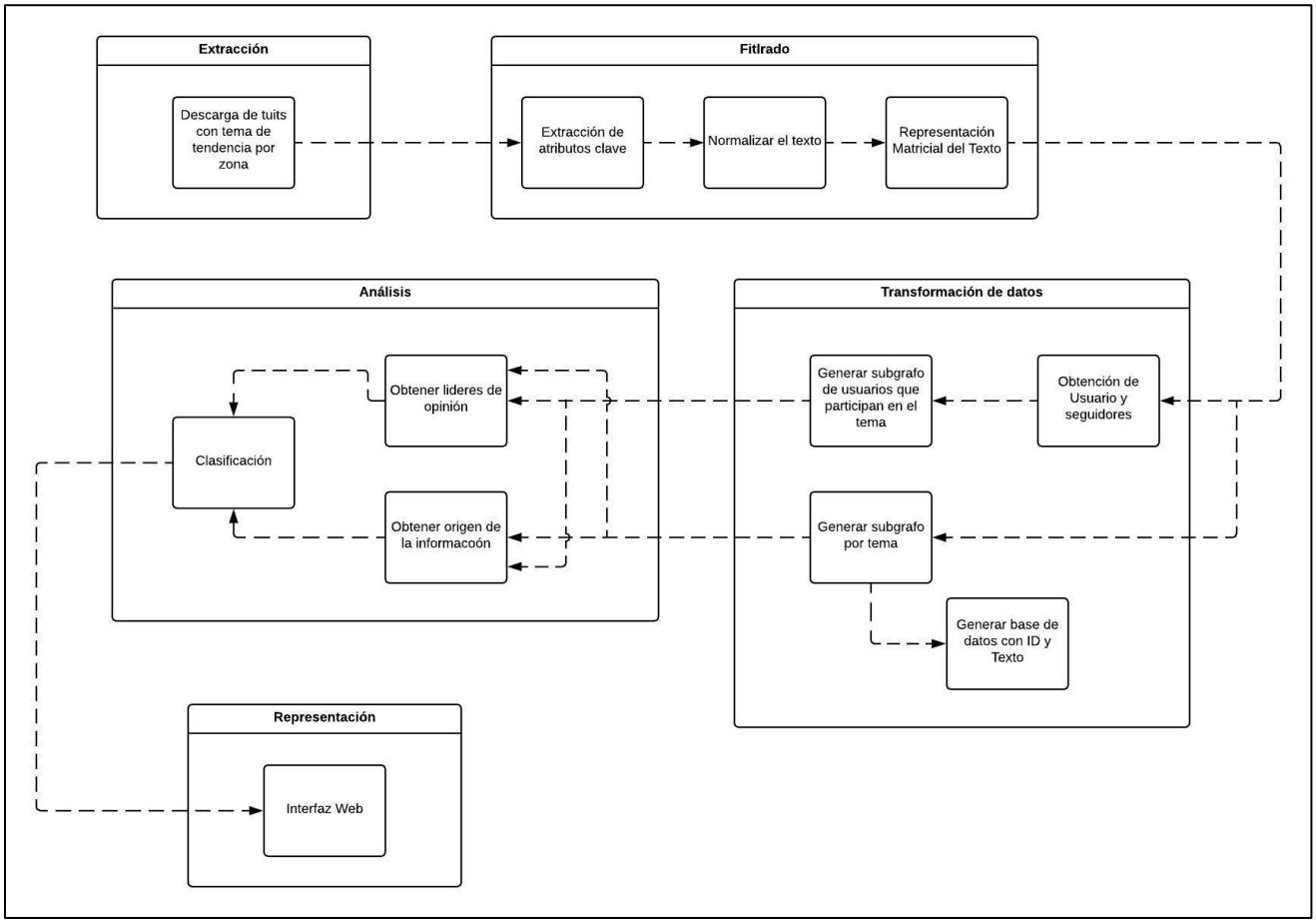


Ilustración 4 Diagrama de bloques del sistema

Alcances

El resultado final del análisis de los líderes de opinión se verá representada con gráficas por tema de tendencia, como primer limitante se encuentran las contracciones de palabras que utilizan los usuarios, al no poder delimitar los parámetros se debe considerar qué palabras son importantes dentro de cada tuit para obtener la información esencial del que habla sin perder el sentido de la oración, como área de trabajo se decidió que la zona de estudio será la ciudad de México al tener la mayor cantidad de usuarios activos en la red social, delimitado en el centro histórico con un radio de 35 km a la redonda.

Debido a la gran cantidad de información que se publica diariamente, se recolectarán los tuits en un documento semana con semana hasta tener almacenados seis meses, ya que Twitter no permite descargar tuits por más de una semana, siendo que para el estudio que se aplicará se necesitan grandes cantidades de tuits para tener un mejor resultado.

El estudio solo estará enfocado a la relación entre los temas de tendencia y los líderes de opinión, identificarlos y saber por medio de *retweets* qué personas se identifican como líderes.

La base de datos a utilizar se compondrá de tuits descargados en un periodo aproximado de 6 meses. Para construir esta base se realizarán descargas por medio de robots que almacenarán los datos en una computadora. Una captura continua de tuits durante meses significa un volumen de información muy considerable, tanto para almacenar como para procesar. Aproximadamente 50 tuits por segundo, guardando la información en modo texto plano (200 bytes) son 10K por segundo, 600K por minuto, 36 Megas por hora, 864 megas al día, dicha tasa puede variar dependiendo del estado de los servidores de descarga y la API de Twitter.

El tiempo que se necesitará para hacer las pruebas y desarrollo del análisis sobre los tuits ya almacenados conlleva a trabajar en los primeros cuatro puntos del cronograma desde el proyecto terminal 1, teniendo que trabajar primero en los robots que permitan crear la base de datos.

Como primeros resultados presentables en proyecto terminal 1 serán todos los datos almacenados descargados por los robots de al menos 3 meses, debido a que el sistema para el análisis se estará implementando y probando durante proyecto terminal 2 para presentar sus resultados completos.

Justificación

De acuerdo con el artículo ¿Por qué los líderes de opinión influyen ahora más que nunca? [29] las discusiones que antes se limitaban a un puñado de personas alrededor de una mesa, ahora es a escala global a través de Facebook, Twitter, wiki y miles de redes. Cualquier persona puede opinar, pero los más influyentes son aquellos a los que la gente escucha, aprende de ellos y adoptan medidas en función de sus opiniones.

En México el 66% de los usuarios de internet usan Twitter [31], generando tráfico de información y opiniones en tan poco tiempo debido a que sus publicaciones son cortas. En el mundo se tuitean más de 55 mil tuits por minuto [30]. En Twitter los temas de tendencia para el usuario final solo le sirven para saber de forma puntual de que se habla en ese momento. Identificar el origen de una tendencia no suele ser tan fácil. Existe tanta información en la red que seguirla es casi imposible pero cuando muchas personas tienen la misma información, se hace notar esa misma información. Los temas de tendencia se escogen por un algoritmo que mide no solo la popularidad de los temas (en base al número total de tuits), sino la velocidad a la que surgen ciertos debates entre usuarios [32].

Este estudio se enfoca en la CDMX por ser la zona de México que contiene al 60% de los usuarios activos de Twitter [2], además este estudio es de fácil adaptación para estudiar otras zonas. Con el fin de poder identificar a los usuarios más influyentes y su relación con las tendencias. Decir si son participantes en el debate de una tendencia y su impacto que tuvo su opinión con el resto de los usuarios.

Se debe poder identificar la fuente de la información, los líderes de opinión y también saber si pueden llegar a ser los mismos, los trabajos existentes se encargan de analizar la información generada

en la Red, sin embargo, la mayoría se enfoca en analizar el texto y no la relación que existe entre cada participante de la información.

Conocer y analizar la información es útil para medir la difusión en la red y si existen individuos que impulsaron ese tema, ¿de qué manera?, ¿Con algún propósito en especial?, ¿De forma inconsciente?, ¿Esperaban que se propagara? En la red social Twitter se genera información de manera exponencial debido a la gran cantidad de usuarios por lo que existe relación entre usuarios a través de la información que se comparte de forma pública.

Objetivo general

Realizar un sistema que sea capaz de identificar a los líderes de opinión en temas de tendencia en Twitter y de rastrear el origen de la información con técnicas de minería de datos.

Objetivos específicos

- Implementar un sistema que realice la descarga diaria de los tuits relacionados con los temas de tendencia usando los hashtags.

- Implementar un sistema que descargue todos los tuits publicados en la región de la Ciudad de México.

- Preprocesar el contenido de los tuits para limpieza del texto en español.

- Filtrar los tuits similares a los temas de tendencia a través de procesamiento de lenguaje natural.

- Obtener información de los usuarios que participaron en cada tema de tendencia.

- Generar un subgrafo para cada tema de tendencia, donde cada usuario que haya participado en el tema de tendencia será un nodo y los arcos serán la relación entre los usuarios.

- Se medirá la influencia de un usuario por el número de seguidores, la cantidad de favoritos que tengan sus tuits y las conexiones que tenga con otros tuits. Las conexiones podrán ser *retweets* o respuestas.

- Aplicar un algoritmo para el análisis de redes complejas, para conocer la relación que existe entre los temas de tendencia, los usuarios y los líderes de opinión.

- Implementar una interfaz web para presentar la información obtenida del análisis de los subgrafos.

Capítulo 2 Estado del arte

En el trabajo Estudio de tendencias diarias en Twitter [3] como objetivo se crea una aplicación que es capaz de contextualizar los tuits publicados en la red de microblogging, mostrando información al usuario para que pueda conocer el contenido de las tendencias que han surgido en Twitter mientras el usuario no estaba presente en dicha plataforma. También existe un trabajo [4] donde se propone utilizar un pequeño conjunto de características específicas de dominio extraídas del texto y del perfil del autor, para clasificar la información difundida en Twitter, debido a que la plataforma limita el número de caracteres por mensaje los algoritmos de clasificación tradicionales no obtienen suficiente información.

Hashtracking [8] es una herramienta que ofrece información sobre los tuits y *retweets*, búsquedas o impresiones de un *hashtag* en específico. AL colocar el *hashtag* en la búsqueda se obtiene información que ayuda a las empresas a darle marketing a sus marcas para poder decidir qué hashtags son ideales para promocionar.

Los costos de esta herramienta Hashtracking se muestran en la tabla 1:

Tabla 1 Costos de Hashtracking

Plan	Personal	Bronce	Plata	Oro	Platino
Costo al mes	$50	$100	$250	$550	$1500
Seguidores	3	6	15	40	100
Publicaciones	50 mil tuits	150 mil tuits	400 mil tuits	1 millón de tuits	3 millones de tuits
Límite de publicaciones	10 mil publicaciones de Instagram	25 mil publicaciones de Instagram	75 mil publicaciones de Instagram	200 mil publicaciones en Instagram	600 mil publicaciones en Instagram
Pared de transmisión	Básica	Básica	Básica	Básica	Moderada
Análisis en tiempo	Real	Real	Real	Real	Real
Cantidad de archivos	Ilimitados	Ilimitados	Ilimitado	Ilimitado	Ilimitado

En la UPIITA se tiene un trabajo terminal con el nombre “Extracción y clasificación de la información del tráfico vehicular a través de Twitter y su visualización en Google Maps” [5]. En este trabajo hacen uso la API de Twitter para la extracción de los tuits, después por medio del escaneo y análisis de los mensajes con un esquema ontológico de palabras clave se realiza una comparación con cadenas de texto para identificar si el mensaje está relacionado con el tema de tráfico vehicular. Una vez que se identificaron esos tuits, por medio del algoritmo de Bayes y otros métodos de clasificación definidos en la biblioteca WEKA se extrae la ubicación geográfica y se procesa para identificarlo en un mapa de una aplicación web.

En otro trabajo terminal con el título “Integración de la red social de Twitter a un ambiente de TV interactiva”, se utiliza la red social

aplicada con una programación de TV, esto para mostrar tuits con información del programa, del elenco, menciones o por usos de hashtags [6].

A nivel institucional los trabajos relacionados con el análisis de redes sociales son solo para obtención de datos y con fines de marketing.

El Centro de Investigación en computación (CIC) tiene la tesis "Análisis automático de opiniones de productos en redes sociales" [7] donde con el uso de procesamiento de lenguaje natural en análisis de textos cortos de opinión pueden clasificar los tuits como muy positivos, positivos, neutros, negativos, muy negativos, sin opinión o sentimiento donde con un diccionario con 556,210 formas y por pruebas se llegó a una exactitud máxima de 56.75% generando un diccionario de características multipolares de 575 lemas.

En el estudio de tendencias diarias en Twitter [3] se desarrolla una aplicación que permita a los usuarios explorar el conjunto de tendencias que han ido apareciendo a lo largo del tiempo, permitiendo ver cómo se desarrolla cada tema de tendencia que se desarrolla en Twitter, dejando los datos solicitados en la aplicación para un posterior análisis. Dicha aplicación ofrece clasificación y agrupamiento de temas de tendencia, así como la visualización gráfica de la evolución en el tiempo real de las tendencias más importantes, también permite al usuario buscar los tuits que más han destacado en la comunidad con respecto a cada tema de tendencia. Esta aplicación es capaz de informar al usuario de las tendencias que han surgido en Twitter a lo largo del tiempo, también mostrando la información necesaria para que pueda percatarse sobre qué temas se ha hablado, cuáles son los mensajes más populares y cómo se relacionan los temas de tendencia entre sí.

La aplicación online Followerwonk [23] realiza búsquedas en Twitter por temas, muestra la actividad del usuario en la red como, por ejemplo: número de retuits y número de menciones directas, para saber la relevancia que tiene con el resto de los usuarios. Ayuda a encontrar información en la biografía de un usuario y también compara la relación de los tuits para identificar si existen relaciones entre usuarios y cuáles son los que más actividad tienen en la red social,

analiza a los seguidores por ubicación, biografía, a quien sigue y quienes los siguen.

La herramienta SocialBro [24] por otro lado sirve para medir nuestra propia influencia con la red, el resultado de este análisis es indicar el momento adecuado para hacer un tuit dependiendo del tema que queremos publicar identificando audiencias. Esta herramienta tiene costos, se divide en dos planes, el *Audiense Insights* que se enfoca en generar nueva audiencia mostrando informes de cómo se divulga la información sin importar que tan especifica sea la audiencia y el *Audiense Connect* está orientado a hacer contacto directo con la audiencia por medio de mensajes privados enviados con *chatbots* personalizados, incluso permite crear árboles de conversación para personalizar la interacción, enviando ocasionalmente informes del análisis, gestión y monitoreo de la comunidad. Los cobros por plan y características de la herramienta se muestran en la Tabla 2.

Tabla 2 Costos del uso de la herramienta Audiense de SocialBro

Cobro	Gratis	Motor de arranque	Estándar
Mensual	0	10 Dólares	99 Dólares
Anual	0	96 Dólares	948 Dólares
Cuentas Twitter	1	2	2
Chatbots	No	No	Si
Modo *Follow/Unfollow*	Si	Si	Si
Perfiles de búsqueda	200 perfiles por búsqueda	Ilimitado	Ilimitado
Alertas definidas por el cliente	No	No	Si
Audiencias adaptadas	No	Si	Si
Análisis de tuit	No	No	Si
El mejor momento para tuitear	Si	Si	Si

Para poder identificar a los líderes de opinión por medio de temas específicos o por hashtag está la aplicación Tweetlevel [25] que al

final separa a los usuarios por la forma en la que interactuación con la información buscada siendo espectador, comentador, comisario, emprendedor y difusor. El difusor es el de interés ya que son los que gozan de mayor popularidad e influencia. Este análisis por esta aplicación es obteniendo información de la audiencia, se califica la audiencia y la relación con los segmentos del cliente de la aplicación, tratando de entender las actitudes, comportamientos y hábitos, contenido de la información (usando técnicas de investigación elaboran comunicaciones altamente dirigidas, controlan a dónde va la información y los mensajes, como se comparte y como entra a grupos de usuarios más cerrados), contexto (comprenden el entorno de mercado del cliente para poder identificar como implementar la información a divulgar, logrando esto, evitan audiencia que genere ruido o estorbe la propagación de la información) e impacto (por medio de modelos estadísticos demuestran el impacto comercial y la recompensa financiera del programa frente al costo, cambios, progreso y el nuevo impacto en la comunidad), los costos de esta aplicación no están disponibles.

Motores de búsqueda en la web son muchos y muy variados, dependiendo del resultado que esperemos es lo específico y lo intenso de la búsqueda, Topsy [26] es un motor impulsado por diferentes redes sociales. En especial sirve para buscar tuits, que ordena según la influencia de quienes envían los mensajes. Solo busca por tendencias y mensajes publicados con anterioridad, estos resultados los usan con fines de marketing para identificar a los líderes de opinión e identificar el momento más oportuno para publicar un tuit y hacerlo lo más viral posible, es fácil de usar porque no utiliza una configuración previa y su interfaz de usuario es sencilla de usar. Esta herramienta tiene análisis de sentimientos, lo que da un clasifica con puntajes de sentimiento instantáneo para cualquier termino, hashtag o identificador de Twitter, permite mandar alertas inmediatas cuando las menciones en la red aumentan con los temas de interés, algunas marcas que usan este motor es Netflix, Fox Sports, The Washington post, entre otros.

Muchas herramientas y aplicaciones existen en la web, muy pocas sobresalen para la sociedad, dependiendo del uso, calidad de información y fama que logren en la era digital, Lithium [27] es una empresa que a absorbió varias herramientas para crear estudios más completos para sus clientes, la herramienta klout de Lithium es la más conocida debido a que tiene dos segmentos. *Social Media Marketing*

la cual planifica, ejecuta y analiza campas en una sola plataforma de red social y el segmento *Social Costumer Service* la cual hace que los usuarios se involucren en la campaña en el momento que se les imponga. Esta herramienta usa una fuente de inteligencia artificial y aprendizaje automático, lo que facilita la gestión, control de la comunidad y mensajería de las redes sociales. La herramienta proporciona un puntaje el cual es proporcional a su nivel de influencia, este puntaje se basa en diversas variables como son número de seguidores, calidad y cantidad de interacción con los demás usuarios, las menciones que se hacen, los precios de uso no son públicos, se solicitan los datos del usuario/empresa para hacer una cotización personalizada, se divide en dos paquetes 1.- Comunidades y 2.- Gestión de redes sociales, esta última se divide en publicación social o respuesta social, la publicación social es cuando se publica el contenido para impulsar un tema y la respuesta social administra y analiza cómo responde la comunidad a las publicaciones o a los temas indicados.

Tabla 3 Tabla comparativa de los trabajos encontrados y parecidos.

Nombre	**Obtener Perfil de usuarios**	**Clasificación por tema**	**Análisis de sentimientos**	**Análisis estructural de la red a través de subgrafos**	**Análisis de la relación usuario - tuit**
Análisis automático de opiniones de productos en redes sociales	✓	✗	✓	✗	✗
Extracción y clasificación de la información del tráfico vehicular a través de twitter y su visualización en Google Maps	✗	✓	✗	✗	✓
Integración de la red social de Twitter a un ambiente de TV interactiva	✗	✓	✗	✗	✓
Hashtracking	✓	✓	✗	✗	✓
Estudio de tendencias diarias en Twitter	✗	✓	✗	✗	✓
Sistema de análisis para la ponderación de usuarios de twitter en temas de tendencia	✓	✓	✗	✓	✓
Followerwonk	✓	✓	✗	✗	✓
SocialBro	✗	✓	✗	✓	✓
Tweetlevel	✗	✓	✗	✗	✓
Topsy	✗	✓	✓	✗	✓
Lithium/Klout	✓	✓	✗	✗	✓

Capítulo 3 Marco teórico

En esta sección damos una breve explicación de los recursos, conceptos, herramientas y tecnologías a usar a lo largo del proyecto.

Twitter

Twitter es una red social en línea que permite a los usuarios enviar y leer mensajes cortos, de 140 caracteres llamados “tuits”. Los usuarios registrados pueden leer y publicar tuits, aunque los que no están registrados sólo pueden leerlos. Los usuarios pueden acceder a Twitter a través de la interfaz web, SMS o aplicación para dispositivo móvil, el objetivo de esta red social es compartir información relevante en tiempo real.

En Twitter se usan los *hashtags* de manera común, un *hashtag* es una palabra que va al lado del símbolo (#). Dependiendo del país, este símbolo puede ser nombrado como numeral, almohadilla e incluso gato. Los hashtags permiten al usuario crear tendencia, diferenciar, destacar y agrupar una palabra específica en esta red social.

El uso del *hashtag* consiste en crear etiquetas para luego poder agruparlas. Por ejemplo, si un usuario está buscando en Twitter las últimas novedades acerca de videojuegos, deberá hacer una búsqueda sencilla escribiendo en el buscador de Twitter “videojuegos”. Con ello obtendrá un gran número de resultados, pero también cualquier tuit que incluya esa palabra, como el de alguien que comente “Videojuegos con mis amigos”.

Tema de tendencia (*Trending topic*)

Un *Trending Topic* es un algoritmo que se encarga de destacar y clasificar aquellos términos que los usuarios utilizan en Twitter. Este algoritmo va clasificando en tiempo real, y son diversos los factores que van ligados a la popularidad de un tema en la plataforma: El número de usuarios distintos que lo están usando, incremento de usuarios que utilizan el término o los *retweets* que incluyan ese término en concreto [1].

Líder de opinión

Un líder de opinión en redes sociales es un usuario, que puede ser una persona o una empresa, qué por su función, posición social, experiencia o por su carisma, tiene la capacidad de influir en las opiniones, actitudes y comportamiento de otros usuarios dentro de la red social [1].

Muchas campañas de comunicación diseñadas para impactar en una comunidad u organización utilizan mensajes dirijidos a las masas o a una comunidad local. Estos mensaje a menudo se diseminan por la comunidad, pero la estructura interna de la comunidad nos les presta mucha atencion.El articulo “Acelerating the difusion of innovations using opinion leaders”[36] detalla los principios teóricos y metodológicos que subyacen a un enfoque de red para promover el cambio social dentro de las organizaciones y comunidades. En dicho articulo tambien se argumenta la importancia de encontrar los personajes mas influyentes de una red para a traves de ellos lograr difundir un mensaje de forma eficiente.

Minería de datos

La minería de datos es un conjunto de técnicas y herramientas que permiten la explotación de los datos para extraer información que no es detectada a simple vista, esto se logra combinando técnicas semiautomáticas de inteligencia artificial, análisis estadístico, bases de datos y visualización gráfica, para la obtención de información que no esté representada explícitamente en los datos.

Las técnicas de minería de datos se enfocan en el descubrimiento automático del conocimiento contenido en la información almacenada en grandes bases de datos, por lo general se buscan patrones, perfiles tendencias con el propósito de auxiliar o soportar procesos de toma de decisión. Sin embargo, la minería de datos es solo una parte del proceso de extracción de conocimiento a partir de datos, este proceso consta de varias fases como lo son la preparación de datos (selección, limpieza y transformación), su exploración y auditoría, desarrollo de modelos y análisis de datos (minería de datos), evaluación, difusión y la presentación y/o utilización de los modelos.

Además, el proceso de extracción de conocimiento incorpora muy diferentes técnicas como, árboles de decisión, regresión lineal redes neuronales, técnicas bayesianas, etc., de diversos campos como el aprendizaje automático y la inteligencia artificial [11].

Análisis de texto

En la etapa de filtrado se llevará a cabo el procesado de texto, para evaluar similitudes entre los tuits y de esta forma agruparlos de acuerdo con cada tema, este proceso consiste en la búsqueda de palabras clave para posteriormente detectar un patrón. Dos algoritmos usados en este tipo de tareas son:

Algoritmo Knuth-Morris-Pratt (KMP)se basa en usar técnicas de precondicionamiento con autómata para poder encontrar de manera eficiente la ocurrencia de un patrón p en una cadena dada con coste en tiempo de pre procesado y de ejecución lineal [12].

Algoritmo de Boyer-Moore El algoritmo se desplaza dentro de la cadena de búsqueda de izquierda a derecha, y dentro del patrón de derecha a izquierda. La mayor eficiencia se consigue minimizando el número de comparaciones entre caracteres, desplazando lo máximo posible la venta de comparación, a costa de una computación previa [12].

Aprendizaje automático

El aprendizaje automático es una rama de la Inteligencia Artificial que tiene por objetivo desarrollar técnicas mediante las cuales las computadoras puedan aprender a desarrollar tareas que los seres humanos hacemos de forma natural y rápida, como, por ejemplo, reconocer imágenes, entender el lenguaje natural, tomar decisiones, etc. El aprendizaje automático se divide en dos áreas principales que son el aprendizaje supervisado y aprendizaje no supervisado.

El aprendizaje supervisado consiste en entrenar un sistema partir de un conjunto de datos etiquetados o patrones de entrenamiento, compuesto por patrones de entrada y la salida deseada. El objetivo del algoritmo es desarrollar una función capaz de deducir el valor

correspondiente a cualquier entrada válida, de manera tal que la salida generada sea lo más cercanamente posible a la verdadera salida dada una cierta entrada. El patrón de salida hace el papel de supervisor [13].

Mientras que, en el aprendizaje no supervisado, muchas veces llamado de autoorganización, el propio sistema trata de identificar algún tipo de regularidad en un conjunto de datos de entrada sin tener conocimiento a priori, solamente requiere de vectores de entrada para adiestrar el sistema. Esto se logra mediante el algoritmo de entrenamiento, que extrae regularidades estadísticas desde el conjunto de entrenamiento [13].

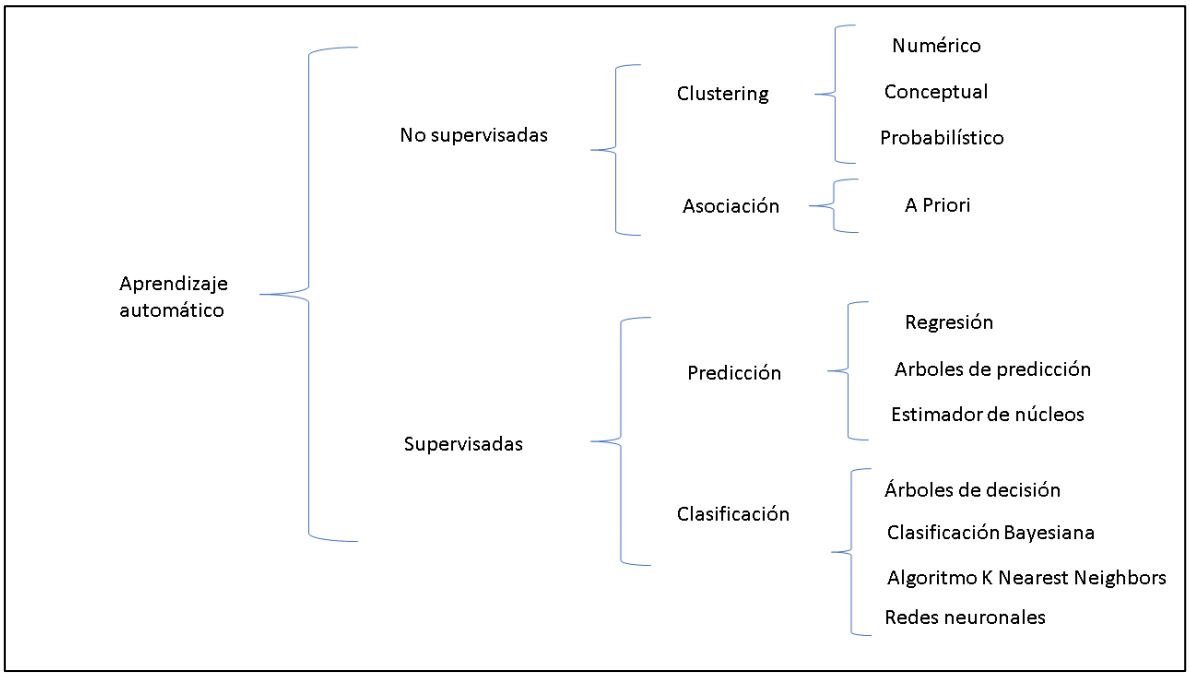


Ilustración 5 Técnicas usadas en Minería de Datos

- *Clustering*

También llamada agrupamiento, permite la identificación de tipologías o grupos donde los elementos guardan gran similitud entre sí y muchas diferencias con los de otros grupos. Así se puede segmentar el colectivo de clientes, el conjunto de valores e índices financieros, el espectro de observaciones astronómicas, el conjunto de zonas forestales, el conjunto de empleados y de

sucursales u oficinas, etc. La segmentación está teniendo mucho interés desde hace ya tiempo dadas las importantes ventajas que aporta al permitir el tratamiento de grandes colectivos de forma pseudoparticularizada, en el más idóneo punto de equilibrio entre el tratamiento individualizado y aquel totalmente masificado. Las herramientas de segmentación se basan en técnicas de carácter estadístico, de empleo de algoritmos matemáticos, de generación de reglas y de redes neuronales para el tratamiento de registros. Para otro tipo de elementos a agrupar o segmentar, como texto y documentos, se usan técnicas de reconocimiento de conceptos. Esta técnica suele servir de punto de partida para después hacer un análisis de clasificación sobre los *clusters*. [13]

- Reglas de asociación

"Este tipo de técnicas se emplea para establecer las posibles relaciones o correlaciones entre distintas acciones o sucesos aparentemente independientes; pudiendo reconocer como la ocurrencia de un suceso o acción puede inducir o generar la aparición de otros. Son utilizadas cuando el objetivo es realizar análisis exploratorios, buscando relaciones dentro del conjunto de datos. Las asociaciones identificadas pueden usarse para predecir comportamientos, y permiten descubrir correlaciones y coocurrencias de eventos" [13]

- Predicción

Es el proceso que intenta determinar los valores de una o varias variables, a partir de un conjunto de datos. La predicción de valores continuos puede planificarse por las técnicas estadísticas de regresión.

Clasificación

La clasificación es el proceso de dividir un conjunto de datos en grupos mutuamente excluyentes, de tal forma que cada miembro de un grupo esté lo más cerca posible de otros y grupos diferentes estén lo más lejos posible de otros, donde la distancia se mide con respecto a las variables especificadas, que se quieren predecir. Algunos métodos que se usan para la clasificación son:

- Algoritmo K Nearest Neighbors (K-NN): Es un algoritmo popular y utilizado en modelos de clasificación. El algoritmo clasifica cada dato nuevo en el grupo que corresponda, según tenga *k* vecinos más cerca de un grupo o de otro. Es decir, calcula la distancia del elemento nuevo a cada uno de los existentes, y ordena dichas distancias de menor a mayor para ir seleccionando el grupo al que debe pertenecer. Es de aprendizaje supervisado, es decir, a partir de un conjunto de datos inicial su objetivo será el de clasificar correctamente todas las instancias nuevas. El conjunto de datos típico de este tipo de algoritmos está formado por varios atributos descriptivos y un solo atributo objetivo (clase). [13]

- Árboles de decisión: El aprendizaje de árboles de decisión está englobado como una metodología del aprendizaje supervisado. Un árbol de decisión se puede interpretar como una serie de reglas compactadas para representarlas en forma de árbol. Teniendo un conjunto de ejemplos, estructurados como vectores de pares ordenados atributo-valor. Cada eje está etiquetado con un par atributo-valor y las hojas con una clase, de forma que la trayectoria que determinan desde la raíz los pares de un ejemplo de entrenamiento alcanza una hoja etiquetada con la clase del ejemplo. [13]

- Clasificación Bayesiana: Los clasificadores Bayesianos son clasificadores estadísticos, que pueden predecir tanto las probabilidades del número de miembros de clase, como la probabilidad de que una muestra dada pertenezca a una clase particular. La clasificación Bayesiana se basa en el teorema de Bayes, y los clasificadores Bayesianos han demostrado una alta exactitud y velocidad cuando se han aplicado a grandes bases de datos. Diferentes estudios comparando los algoritmos de clasificación han determinado que un clasificador Bayesiano sencillo conocido como el clasificador “naive Bayesiano” es comparable en rendimiento a un árbol de decisión y a clasificadores de redes de neuronas. [13]

- Redes neuronales

  Las redes neuronales constituyen una nueva forma de analizar la información, su principal diferencia con las técnicas

tradicionales es, que son capaces de aprender y detectar patrones dentro de los datos. Las redes neuronales se comportan de forma similar al nuestro cerebro aprendiendo de la experiencia y aplicando ese conocimiento a la resolución de problemas. Este aprendizaje se obtiene como resultado del entrenamiento y éste permite la sencillez y la potencia de adaptación y evolución ante una realidad cambiante y muy dinámica. [13]

Bases de datos NoSQL

Las bases de datos NoSQL son sistemas de almacenamiento de información que no cumplen con el esquema entidad-relación. Mientras que las tradicionales bases de datos relacionales basan su funcionamiento en tablas, *joins* y transacciones. Las bases de datos NoSQL no imponen una estructura de datos en forma de tablas y relaciones entre ellas, sino que proveen un esquema mucho más flexible. Las bases NoSQL son adecuadas para una escalabilidad realmente enorme, y tienden a utilizar modelos de consistencia relajados, no garantizando la consistencia de los datos, con el fin de lograr una mayor performance y disponibilidad.

En general se pueden mencionar Sistemas NoSQL clasificados en cuatro categorías:

- *Framework Map-Reduce* (usado por aplicaciones que hacen procesamiento analítico online - OLAP), es un *framework* que proporciona un sistema de procesamiento de datos paralelo y distribuido, está orientado a resolver problemas con conjuntos de datos de gran tamaño [33]. Por ejemplo, Hadoop.

- Almacenamiento Clave-Valor (sistemas que tienden al procesamiento de transacciones online - OLTP), es un tipo de base de datos no relacional que utiliza un método simple de clave-valor para almacenar datos, almacena datos como un conjunto de pares clave-valor en los que una clave sirve como un identificador único [34]. Por ejemplo: Google BigTable, Amazon Dynamo, Cassandra, Voldemort, HBase.

• Almacenamiento de documentos, una base de datos de documentos es un tipo de base de datos no relacional que está diseñada para almacenar datos semiestructurados como documentos En una base de datos de documentos, cada documento puede tener la misma estructura de datos o no, y cada documento es autodescriptivo, incluyendo su posible esquema único, y no depende necesariamente de ningún otro documento. Los documentos se agrupan en "colecciones", que tienen un propósito similar al de una tabla en una base de datos relacional. Ejemplos de bases de datos en documentos son: CouchDB, MongoDB, SimpleDB. [14]

• Sistemas de base de datos Gráficas. Estas bases están diseñadas expresamente para almacenar relaciones y navegar por ellas. Estas bases, usan nodos para almacenar entidades de datos y bordes para almacenar relaciones entre entidades [35]. Por ejemplo: Neo4j, FlockDB, Pregel.

Teoría de grafos

La teoría de grafos, también llamada teoría de gráficas es una rama de las matemáticas y las ciencias de la computación que estudia las propiedades de los grafos, y que no deben ser confundidos con las gráficas que estudia las propiedades de los grafos, y que no deben ser confundidos con las gráficas que tienen una acepción de los grafos, y que no deben ser confundidos con las gráficas que tienen una acepción muy amplia.

Un grafo es un conjunto, no vacío, de objetos llamados vértices (o nodos) y una selección de pares de vértices, llamados aristas que pueden ser orientados o no. Típicamente, un grafo se representa mediante una serie de puntos (los vértices) conectados por líneas (las aristas). [14]

- **Subgrafo**

En matemáticas y ciencias de la computación, un subgrafo es una generalización de un grafo, donde las aristas pueden relacionarse con cualquier cantidad de nodos, en lugar de solo dos como en un

grafo convencional. Un subgrafo representa las interrelaciones que existen entre unidades que interactúan con otras [14].

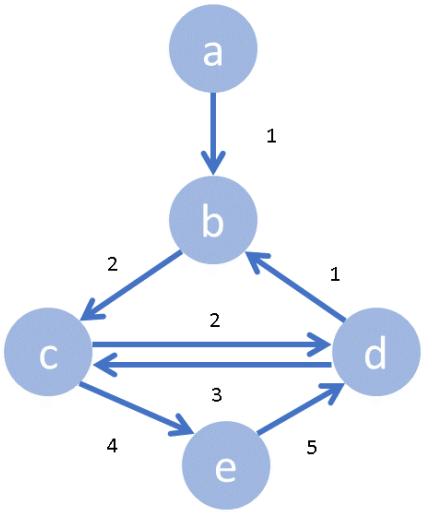
Centro de un Grafo

El centro de un grafo es el vértice con la mínima excentricidad. Encontrar el centro de un grafo se puede realizar aplicando los pasos siguientes:

Aplicar el algoritmo de Floyd para encontrar la longitud de los caminos más cortos entre cualquier par de vértices (El resultado se representa en una matriz $A(i)$).

Hallar el costo máximo en cada columna i (Esto proporciona la excentricidad del vértice i)

Hallar el vértice con la mínima excentricidad (Esto proporciona el centro de G). [18]



A	a	b	c	d	e
a	0	1	3	5	7
b	∞	0	2	4	6
c	∞	3	0	2	4
d	∞	1	3	0	7
e	∞	6	8	5	0

El centro del grafo es d

Ilustración 6 Ejemplo de encontrar el Centro de un Grafo por algoritmo de Floyd

Capítulo 4 Análisis

Es este capítulo se muestran los requerimientos identificados, tanto funcionales como no funcionales. Posteriormente se hace el análisis de las herramientas a emplear, así como de los algoritmos necesarios para la elaboración de este proyecto.

Requerimientos

Requerimientos Funcionales

Los requerimientos funcionales describen las interacciones entre el sistema y su ambiente como los usuarios.

Funcionales:

RF1	El sistema debe tener una propiedad o clase para descargar los tuits y almacenarlos.
RF2	Para la normalización del texto, se deben quitar los emoticonos, enlaces, imágenes, videos y guardar la mayor cantidad de texto sin perder el contexto.
RF3	Para cada tema de tendencia se debe generar un subgrafo.
RF4	El sistema debe poder medir la cantidad de *retweets*, respuestas y seguidores de un participante por tema.
RF5	El sistema debe ser capaz de poder relacionar usuarios entre sí.
RF6	El usuario podrá escoger por tema de tendencia.
RF7	Se mostrarán con graficas los datos sobresalientes y conclusiones del sistema, mostrara por grafos los temas de interés para el usuario.

Requerimientos no funcionales

Los requerimientos no funcionales incluyen características del sistema como restricciones de plataforma o de tiempos.

RNF1	El sistema solo debe descargar de la región de la Ciudad de México
RNF2	Tener separados los tuits de tema de tendencia con los similares
RNF3	El sistema debe de ser capaz de terminar con su análisis y dejar la información lista para ser consultada en el momento que sea por un tercero.
RNF4	Los análisis se deben actualizar semana con semana.
RNF5	Se necesita una clave ID de usuario para poder relacionar con el nodo que corresponda a su tuit en el grafo de tema de tendencia correspondiente.

Diagrama de casos de uso

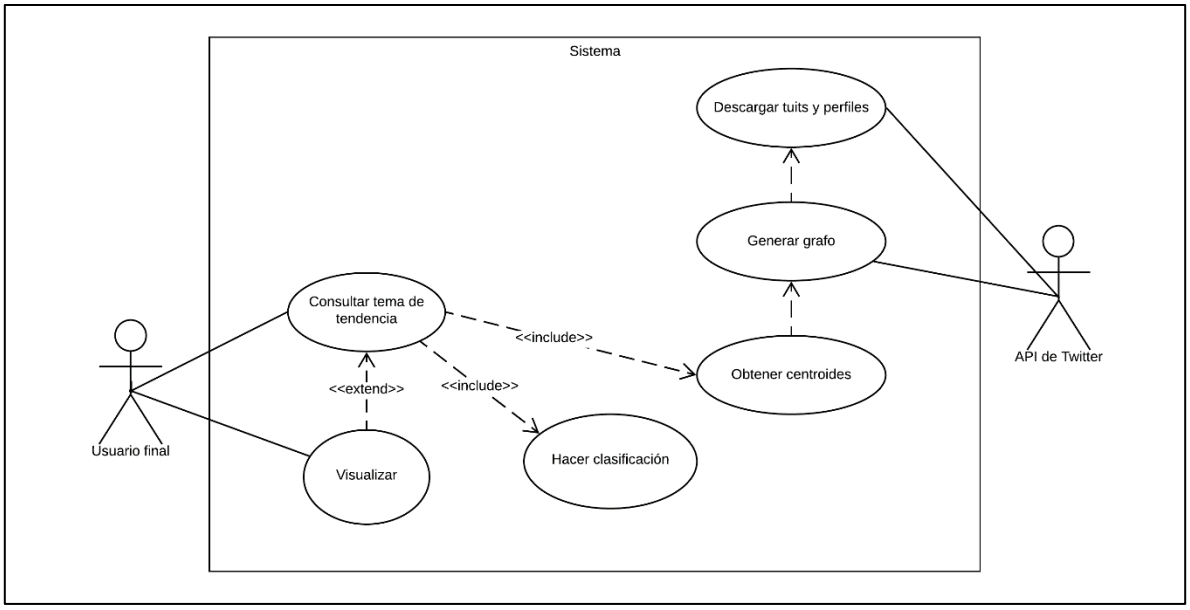


Ilustración 7. Diagrama caso de uso

Tabla 4. Caso de uso "Consultar tema de tendencia"

Nombre:	Consultar tema de tendencia
Autores:	Arturo Ortiz y Alan Sánchez
Fecha:	22 de octubre del 2018
Descripción: Permite al usuario seleccionar el tema de interés en la página web.	
Actores: Usuario final.	
Precondiciones: El usuario debe abrir página web del sistema.	
Flujo Normal: 1.-El actor introduce la página web. 2.-Cuando se carga la página se mostrarán los temas de tendencia de análisis. 3.-El actor selecciona un tema de tendencia que sea de su interés 4.-Se envía la información al caso “Hacer clasificación” y se espera respuesta. 5.-Recibida la información del caso “Hacer clasificación” se solicita el proceso de “Visualizar” y se espera respuesta. 6.-Con las respuestas se le muestra al actor los resultados de una manera gráfica en la página web.	
Flujo Alternativo: 1.-Si no hay respuesta del caso “hacer clasificación” se avisa al actor y se guarda un registro del intento. 2.-Se comprueba que los datos recibidos por el proceso “Visualizar” sean correctos, si no lo son, se avisa al actor y se da un tiempo para volver a intentarlo.	

Tabla 5 tabla caso de uso "Hacer clasificación"

Nombre:	Hacer clasificación
Autores:	Arturo Ortiz y Alan Sánchez
Fecha:	22 de octubre del 2018
Descripción: Realiza la clasificación de los temas de tendencia por tópicos	
Actores: Caso de uso “Consultar tema de tendencia”.	
Precondiciones: Se debe haber realizado el análisis de los tuits por cada tema de tendencia.	
Flujo Normal: 1. Obtener título del tema de tendencia. 2. Mediante algoritmos de análisis de textos se asignará una etiqueta al tema de tendencia. 3. De acuerdo con la etiqueta asignada se asignará a una categoría. 4. De no existir una categoría se podrá decidir si crear una nueva categoría o dejar el tema en la categoría sin clasificar. Esto se definirá en la etapa de entrenamiento del sistema.	
Flujo Alternativo: 1. Si el título del tema de tendencia no se puede reconocer se solicitará nuevamente. 2. Si aún no es reconocible se enviara un informe al administrador y el tema pasara a la sección sin clasificar.	

Tabla 6 Tabla caso de uso "Visualizar"

Nombre:	Visualizar
Autores:	Arturo Ortiz y Alan Sánchez
Fecha:	22 de octubre del 2018
Descripción: Permite al usuario visualizar los resultados obtenidos de nuestro análisis del tema de interés en la página web.	
Actores: Usuario final.	
Precondiciones: El usuario debe abrir página web del sistema.	
Flujo Normal: Se reciben datos del caso “Consultar tema de tendencia”. Se realiza una conexión con la base de datos. Se realiza l búsqueda de la información a partir de los datos obtenidos del caso “Consultar tema de tendencia”. Se presenta el grafo del tema de tendencia, el o los líderes de opinión en ese tema y el origen de la información.	
Flujo Alternativo: 1.-Si no hay respuesta de la base de datos se avisa al actor y se guarda un registro del intento. 2.-Se comprueba que los datos recibidos por el proceso “Consultar tema de tendencia” sean correctos, si no lo son, se avisa al actor y se da un tiempo para volver a intentarlo.	

Tabla 7 Tabla de caso de uso "Descargar tuits y perfiles"

Nombre:	Descargar tuits y perfiles
Autores:	Arturo Ortiz y Alan Sánchez
Fecha:	22 de octubre del 2018
Descripción: Realiza la descarga de los tuits relacionados con un tema de tendencia y el perfil de los usuarios que participaron en el.	
Actores: Objeto del sistema	
Precondiciones: Debe haberse autenticado en el API de Twitter. Debe tener un tema de tendencia para la búsqueda.	
Flujo Normal: Establecer la zona geográfica para la búsqueda y descarga. Buscar y guardar, los tuits relacionados con el tema de tendencia en cuestión. Obtener una lista de los usuarios que participaron en el tema a partir los tuits guardados. Obtener el perfil de cada usuario que se tiene en la lista anterior.	
Flujo Alternativo: Si no hay respuesta del servidor de Twitter se espera 15 minutos y se vuelve a intentar 3 veces más en intervalos de 1 minuto. Después de esperar si aún no hay respuesta se manda un aviso al administrador para verificar que no existan otros errores que el sistema no puede corregir por él mismo.	

Tabla 8 Tabla de caso de uso "Generar grafo"

Nombre:	Generar grafo
Autores:	Arturo Ortiz y Alan Sánchez
Fecha:	22 de octubre del 2018
Descripción: Genera el grafo a partir de los tuits y lista de usuarios obtenidos en el caso “Descargar tuits y perfiles”	
Actores: Objeto del sistema	
Precondiciones: Necesita la lista de tuits y usuarios que participaron en cada tema de tendencia.	
Flujo Normal: Realiza una conexión con el gestor Neo4j. Genera los nodos a partir de los tuits. Genera las conexiones a partir de los *retweets,* favoritos y lista de seguidores (cada uno es un tipo de arista distinto).	
Flujo Alternativo: Si existiera error en la conexión con el gestor de la base de datos, se realizarán 2 intentos más de conexión en intervalos de 1 minuto. De no poderse realizar la conexión con el gestor Neo4j se enviará un reporte del fallo al administrador del sistema.	

Tabla 9 Tabla de caso de uso "Obtener centroides y origen de la información"

Nombre:	Obtener centroides y origen de la información
Autores:	Arturo Ortiz y Alan Sánchez
Fecha:	22 de octubre del 2018
Descripción: Realiza un análisis sobre cada grafo para obtener los líderes de opinión y el origen de la información de cada tema de tendencia.	
Actores: Objeto del sistema	
Precondiciones: Requiere de una conexión con el gestor Neo4j y que existan en él grafos de los temas de tendencia.	
Flujo Normal: Realiza un análisis mediante algoritmos de recorrido en grafos para encontrar a los nodos más relevantes dentro del tema de tendencia. A partir de los nodos más relevantes obtiene al o los usuarios que serán clasificados como líder de opinión dentro de ese grafo. Mediante algoritmos de difusión en grafos se quiere encontrar el origen de la información en cada tema de tendencia.	
Flujo Alternativo: De no obtenerse un resultado convincente se recorrerá el grafo usando otros algoritmos.	

Análisis de Herramientas

Twitter

API de Twitter

Una API (*Application Programming Interface*) es un conjunto de reglas (código) y especificaciones que las aplicaciones pueden seguir para comunicarse entre ellas: sirviendo de interfaz entre programas diferentes de la misma manera en que la interfaz de usuario facilita la interacción humano-software. En la página oficial para desarrolladores se comenta "Twitter es lo qué está pasando en el mundo y sobre lo que las personas están hablando en este momento. Puedes acceder a Twitter en la Web o desde tu dispositivo móvil. Para compartir información en Twitter de la forma más amplia posible, también les proporcionamos a las empresas, los desarrolladores y los usuarios acceso programático a los datos de Twitter mediante nuestras API [18].

Hay 5 categorías en las que se dividen las API:

1. Cuentas y usuarios: permite programar el perfil y configuración de una cuenta, silenciar o bloquear usuarios, administrar usuarios y seguidores, solicitar información sobre actividad de una cuenta.

2. Tweets y respuestas: Permite acceder a los tweets de toda la plataforma de forma global o con una búsqueda usando palabras clave específicas, también la publicación de nuevos tweets desde cuentas específicas de usuarios.

3. Mensajes directos: otorga la información y acceso a las conversaciones por mensajes directos a usuarios que dieron permiso de forma explícita a una aplicación específica, esto para que los desarrolladores puedan crear, probar y analizar conversaciones de *chatbots* y hacer la experiencia de diálogo más real y fluida.

4. Anuncios: Esta API permite a los desarrolladores crear algoritmos para crear y administrar campañas de anuncios de una forma automática usando tuits públicos para temas o

intereses específicos logrando que las empresas cumplan con sus objetivos de llegar a las audiencias deseadas.

5. Herramientas y SDK del editor: Estas API permiten integrar las cronologías de Twitter, el botón de compartir y otros contenidos de Twitter en las páginas web, como cuando se hace alguna publicación y de forma automática los clientes puedan compartir la información fácilmente.

En la plataforma para desarrolladores de Twitter proporciona herramientas para crear campañas orientadas a diferentes resultados, se divide por Productos de datos, API de anuncios y Herramientas para editores que hacen que sea más fácil para los desarrolladores solicitar y aprovechar la información de Twitter.

Productos de datos

1. API estándar: Son gratuitas, excelentes para comenzar, son usadas para probar integración de códigos para proyectos, crear soluciones de análisis o estudios, se puede publicar contenido y obtener datos no disponibles en grandes volúmenes.

2. API Premium: Ofrecen acceso escalable a los datos de Twitter para los desarrolladores que buscan crecer sus códigos o experimentar e innovar con la información. Los desarrolladores que usan estas APIS son los que necesitan mayor cantidad de datos para sus fines o buscan enriquecer sus códigos con información más detallada.

3. API de empresa: éstas se ofrecen el nivel más alto de acceso y confiabilidad para quienes necesitan la mayor cantidad de información que ofrecen los datos que están en Twitter. Con acceso más confiable, paquetes de costos personalizados a medida del uso o contratos anuales. Se ofrecen los servicios de administración de cuenta dedicados y soporte técnico.

❖ API de anuncios

1. API de anuncios: Ofrece una forma programada de integrarse con la plataforma de anuncios de Twitter en la cual orienta como

poder hacer publicaciones programadas o automatizadas usando sus propias herramientas para administrar y optimizar campañas publicitarias, ofreciendo informes, creatividades y audiencias personalizadas por parte de Twitter.

2. MoPub: Esta plataforma es una monetización para editores y desarrolladores de aplicaciones móviles. Es usada para la mediación de red y el intercambio programático móvil, ayuda a los editores a aumentar los ingresos publicitarios mientras se mantiene un control de la información.

❖ Herramientas de editor y SDK

1. Web: Estas API sirven para poner contenido y actividad social en una página web, permiten visualizar tuits, impulsar acciones sociales y administrar medios.

2. AMP en Twitter: AMP (*Accelerated Mobile Pages*) es como lo dice su nombre un acelerador de páginas móviles, esto hace que la experiencia de los lectores sea más armónica o por lo tanto hace que se involucren crecidamente, provocando que el rendimiento de las publicaciones en Twitter aumente.

Librerías para la administración de la API de Twitter.

Buscando en la web de desarrolladores de Twitter se encuentran las diferentes librerías que existen para hacer uso de la API de Twitter. Se encontraron muchas opciones en distintos lenguajes de programación, se decidió reducir las opciones a 3 lenguajes, C++, Java y Python, debido a que estos lenguajes son los más usados en UPIITA. Una vez seleccionados estos lenguajes nos dirigimos a los sitios web de cada una de las librerías para revisar en su documentación la forma de hacer uso de cada una de ellas además de las funciones que nos proporcionaban para obtener información de Twitter.

Los resultados de esta investigación se presentan en la siguiente tabla comparativa.

Tabla 10 Tabla comparativa de las librerías que administran las API´s de Twitter

Lenguaje	Librería	Métodos para obtener Trends	Métodos de búsqueda	Métodos de Usuarios	Métodos de *timeline*
C++	Twitcurl	Tiene distintos métodos que permiten obtener los *trends* en un día específico, de una semana o los actuales	Permite realizar búsqueda por frases además de guardarlas búsquedas	Nos permite ver la lista de seguidores y a quien sigue el usuario	Permite ver el *timeline* de un usuario
Java	Twitter4J	Trendsresources (mediante http)	Searchresource(mediante http)	Usersresources(mediante http)	Timelinesresources(mediante http)
Python	Tweepy	Se pueden obtener los *trends* en un periodo de tiempo especificado por el usuario	Realiza la búsqueda de tweets por frase y en un periodo de tiempo específico	Permite acceder a la información de un usuario siempre que sea pública	Accede al *timeline* de un usuario y devuelve los resultados en formato *json*
	Python Twitter Tools	Tiene una clase que se encarga del manejo de los *trends*	Realiza búsqueda por hashtag	Tiene una clase que se encarga de obtener información de usuarios	Devuelve el *timeline* de un usuario en formato *json*
	Birdy	Se pueden obtener los *trends* en un periodo de tiempo especificado por el usuario	Realiza la búsqueda de tuits por frase y en un periodo de tiempo específico	Permite acceder a la información de un usuario siempre que sea pública	Accede al *timeline* de un usuario y devuelve los resultados en formato *json*

Se puede ver que todas las librerías son muy parecidas y esto es por que acceden a la API de Twitter así que se decidió usar la librería *tweepy* debido a que cuenta con mayor documentación y su comunidad es más grande, además también tomamos en cuenta que Python es un lenguaje muy usado en la comunidad de análisis de datos en redes sociales.

Descarga y análisis de datos descargados de Twitter

Para realizar la descarga de información de Twitter primero se tiene que registrar la aplicación en la web de Twitter para poder acceder a la API de Twitter mediante claves que nos proporciona. Una vez que se tienen las claves de acceso en el script se crea una función que recibe como parámetro esas claves y nos devuelve un objeto que nos permitirá hacer uso de las diferentes funciones que ofrece el API de Twitter.

Para la descarga de tuits se hace uso de la API de búsqueda que nos permite realizar descargas de información a partir de palabras, es por eso se generó una función para encontrar los temas de tendencia en la Ciudad de México, además de esto la función de descarga necesita una coordenada geográfica y un radio para buscar tuits en una zona bien delimitada. Debido a que Twitter establece un límite de descargas por aplicación, se decidió dividir la zona de descargas en 14 subzonas, para poder capturar la mayor cantidad de tuits sin pérdidas por el límite que establece Twitter. En la ilustración 8 se muestra un circulo con centro en el centro histórico de la Ciudad de México y un radio de 35 Km, que es la zona elegida para el estudio de las tendencias, y dentro de ese círculo las 14 zonas que se generaron para disminuir la perdida de información.

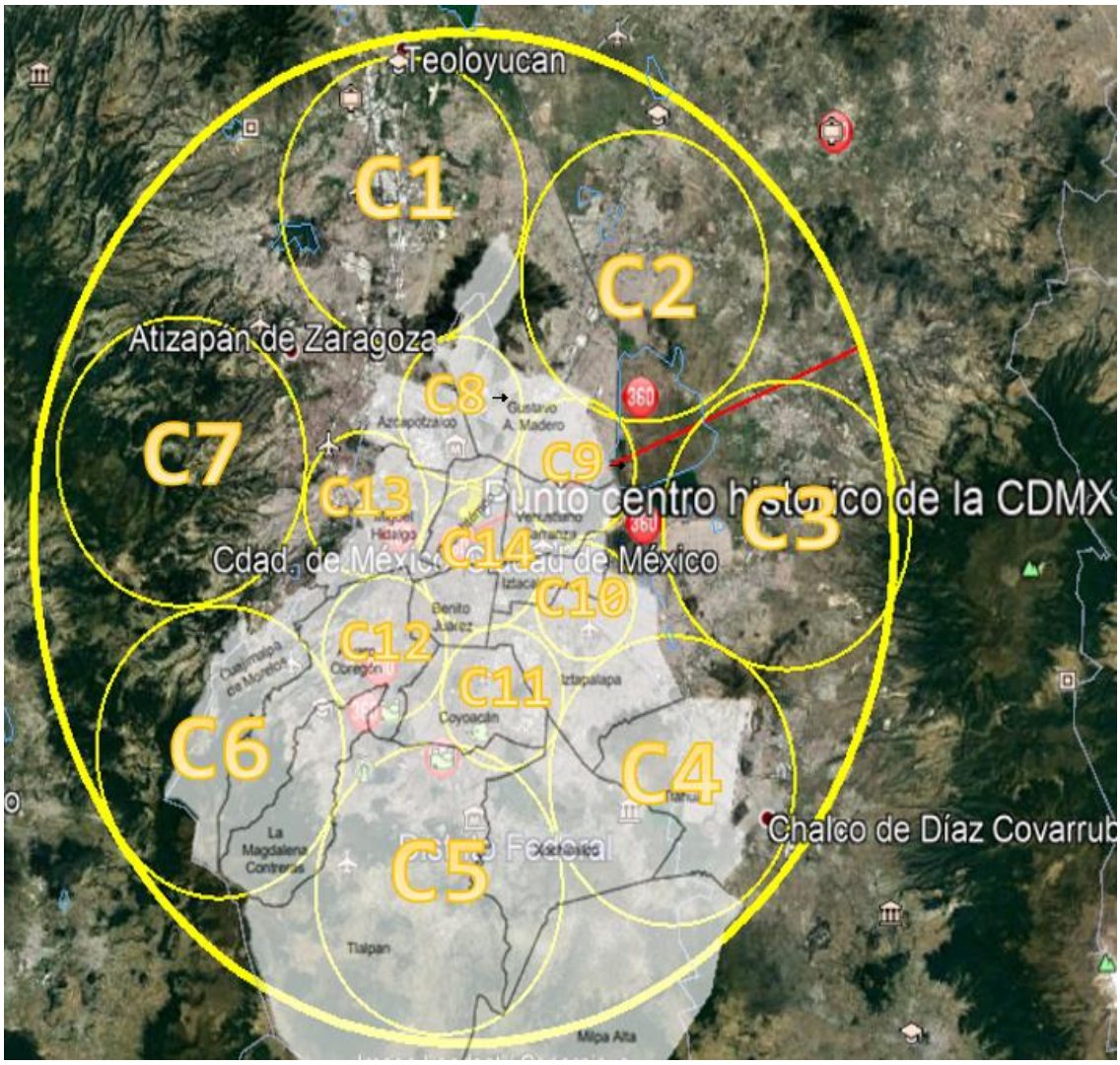


Ilustración 8 Mapa de la zona a estudiar

De la información que obtenemos de Twitter solamente nos interesa guardar los siguientes campos:

- *created_at*: Indica la fecha en la que se emitió el tweet.
- id. (tuit): Es un identificador del tweet que nos servirá para posteriormente eliminar elementos repetidos.
- texto: Es el contenido del tuit.
- *users_mentions*: Indica si dentro del tuit se mencionó a algún usuario.

- *in_reply_to_status_id*: Indica si el tuit actual es respuesta de algún otro emitido con anterioridad.

- *in_reply_to_usr_id*: En caso de que el tuit actual sea respuesta nos indica el id del usuario al que se le respondió.

- *user*:

    - *user_id*: Identificador del usuario que emitió el tuit.

    - *screen_name*: Es el nombre con el que se identifica al usuario dentro de la red social.

    - *followers_count*: Indica el número de seguidores que tiene el usuario.

- *retweet_count*: Indica el número de *retweets* que ha tenido el tweet actual.

- *favorite_count*: En caso de que el tweet actual haya sido marcado como favorito nos muestra la cuenta de favoritos

- *favorited*: Es un valor booleano dependiendo si el tweet ha sido marcado o no como favorito.

- *retweeted*: Es un valor booleano en caso de que el tuit tenga o no *retweets*.

- *zona*: indica a que subzona pertenece el tuit descargado.

Algoritmos para análisis de texto

Los datos en general son un producto básico, no existe una forma ideal para procesarlo, su valor es cuestionable debido a que no es un valor generalizado. La ciencia de datos es un campo multidisciplinario cuyo objetivo es extraer valor de los datos en todas o la mayoría de sus formas. En el campo de la ciencia de datos es por medio de un proceso. Eso no quiere decir que sea mecánico, cuando se profundiza en las etapas de procesamiento de datos, desde la reducción de las fuentes donde se obtienen los datos y la limpieza de los datos ya guardados hasta el aprendizaje automático para finalmente mostrar o visualizar la información se verá que los únicos pasos implican transformar los datos sin procesar en una visión.

De acuerdo con una publicación de IBM “Introducción a la ciencia de datos” [19], en una encuesta del 2016 se descubrió que los científicos de datos dedican el 80% de su tiempo a recolectar, limpiar y preparar datos para su uso en el aprendizaje automático. El 20% restante gasta datos de minería o modelado mediante el uso de algoritmos de aprendizaje automático. Debido a esto, mientras más alto sea el nivel de datos se pueden clasificar los datos en tres categorías: Estructurado, semiestructurado y no estructurado (figura 1). Los datos estructurados son datos altamente organizados que existen dentro de un contenedor como puede ser una base de datos o hasta un archivo de valores separados por comas (CSV). Los datos no estructurados carecen de cualquier estructura de contenido (por ejemplo, una secuencia de audio o texto en lenguaje natural). En el medio se encuentran los datos de semiestructura, que pueden incluir metadatos o datos que pueden procesarse más fácilmente que los no estructurados usando etiquetado semántico.

La tarea de similitud textual se encarga de comparar textos para conocer el parecido entre ellos. Los métodos de similitud textual han tenido aplicaciones muy importantes durante muchos años en los campos de lenguaje natural. La variante en estos métodos son los que calculan el grado de similitud en textos cortos; una tarea un poco más complicada que la original por la poca información que se maneja para poder obtener el parecido de los textos.

Para la detección de similitud textual hay dos modelos muy populares, el Modelo Geométrico y el modelo de teoría de conjuntos.

- Modelo Geométrico

  Este modelo representa objetos en un espacio métrico $< X, \delta >$ logrando manipular el espacio geométrico de manera algebraica, donde cada objeto es representado por un punto $\times \in X$ y $\delta(a,b)$ es la distancia entre un objeto a y un objeto b. Esta distancia es una función sobre X que retorna un número no negativo, donde, entre menos distancia se puede interpretar como mayor similitud. En un espacio geométrico, para todos los puntos $a, b, c \in X$ se cumplen los axiomas de una métrica: minimalidad, simetría e inequidad triangular:

Minimalidad: $\delta(a,b) \geq \delta(a,a) = 0$

Simetría: $\delta(a,b) = \delta(b,a)$

Inequidad triangular: $\delta(a,b) + \delta(b,c) \geq \delta(a,c)$

•Teoría de conjuntos

El modelo geométrico ha sido criticado por encontrar a objetos como similares cuando no lo son. Por esta razón se utiliza la Teoría de conjuntos. El modelo se basa en las características de los objetos a comparar. La similitud $sim\ (a,b)$ se calcula a partir de una función que coteja los elementos que comparten tanto a, como b y tomando en cuenta, también, los elementos que solo están en a y en b, es decir:

$$Sim(a,b) = f(a \cap b, a - b, b - a)$$

De esta forma, se toman en cuenta las características que comparten dos objetos, pero, también se toman en cuenta, directamente, los elementos que hacen que se distinga a "a" de "b".

La notación en adelante a usar es:

- $\sum$ denota algún conjunto finito ordenado de caracteres

- $\sum^*$ el conjunto de cadenas formadas por la concatenación de cero o más caracteres de $\sum$.

- A y B denotan dos cadenas de texto de longitud n y m definidas sobre $\sum^*$, donde $n \geq m$.

- a_i representa algún carácter de A para $1 \leq i \leq n$

- b_j representa algún carácter de B para $1 \leq j \leq m$

La distancia entre dos cadenas, no se expresa exactamente de la similitud entre ellas ya que una magnitud puede calcularse a partir

de la otra. Una distancia real $d \in [0,1]$, donde 0 indica que ambas cadenas son idénticas y 1 que no tienen ni un solo carácter en común, equivalente a una similitud $s= 1-d$, lo que quiere decir que a mayor distancia menor es la similitud y viceversa.

•Algoritmos basadas en caracteres

Estas funciones de similitud consideran cada cadena de caracteres como una secuencia ininterrumpida de caracteres.

La Distancia de edición entre dos cadenas de texto *A* y *B* se basa en el conjunto mínimo de operaciones de edición necesarias para transformar *A* en *B* (o viceversa). Las operaciones de edición permitidas son eliminación, inserción y sustitución de un carácter.

En este modelo original, cada operación tiene costo unitario, siendo mencionada distancia de Levenshtein. Más tarde Needleman y Wunsch lo modificaron para permitir operaciones de edición con distinto costo, permitiendo modelar errores ortográficos y tipográficos comunes, por ejemplo, en español es frecuente encontrar la letra *"n"* usada como *"m"* siendo el caso de siempre y siempre. Teniendo así que asignar un costo de sustitución menor al par de caracteres *"n"* y *"m"* que a otros dos sin relación alguna.

Los modelos anteriores al ser absolutos y no relativos tienen una desventaja, carecen de normalización por lo que tres errores son más significativos entre dos cadenas de texto de longitud cuatro que entre dos de longitud 20. La solución a esto es la primera técnica de normalización de Yujiang y Bo que satisface la desigualdad triangular (la distancia directa para ir del punto *"x"* al *"z"* nunca es mayor que aquella para ir primero del punto *"x"* al *"y"* y después del *"y"* al *"z"*). [20]

Algoritmos de difusión en grafos y búsqueda de centroides en grafos dirigidos

Algoritmo de Prim del árbol de expansión

El algoritmo parte del problema de que se quiere transferir eficientemente un elemento de información a todos y cada uno de los elementos de la red. La solución a este problema radica en la construcción de un árbol de expansión de ponderación mínima. Formalmente definimos el árbol de expansión mínimo T para un grafo $G=(V,E)$ como sigue. T es un subconjunto acíclico de E que conecta todos los vértices de V. Se minimiza la suma de las ponderaciones de las aristas de T.

El algoritmo de Prim encuentra el árbol de expansión de coste mínimo. Es un algoritmo clásico que se incluye dentro de la categoría de voraces. El punto de partida es un grafo $G=(V,A)$ modelo red y sea $c_{i,j} \geq 0$ el peso o coste asociado al arco (i,j). Si se supone $V = \{1,2,3,...,n\}$ el algoritmo arranca asignando un vértice inicial al conjunto W, por ejemplo el vértice 1. $W = \{1\}$. A partir del vértice inicial el árbol de expansión crece, añadiendo en cada iteración otro vértice v de V-W tal que si u es un vértice de W, el arco (u,v) es el más corto de entre todos los arcos que tienen un vértice u en W y otro v en V-W. El proceso termina cuando $V=W$. Se observa que en todo momento el conjunto de nodos que forma W constituye una componente conexa sin ciclos. [21]

Algoritmo de caminos mínimos

El algoritmo de caminos mínimos es un algoritmo para la determinación del camino más corto, dado un vértice origen, hacia el resto de los vértices en un grafo que tienen pesos en cada arista. La idea de este algoritmo consiste en ir explorando todos los caminos más cortos que parten del vértice origen y que llevan a todos los demás vértices; cuando se obtiene el camino más corto desde el vértice origen hasta el resto de los vértices que componen el grafo, el algoritmo se detiene, No funciona con grafos que contengan aristas de coste negativo. [21]

Teniendo un grafo dirigido ponderado de *N* nodos no aislados, sea *X* el nodo inicial. Un vector *D* de tamaño *N* guardará al final del algoritmo las distancias desde *X* hasta el resto de los nodos.

1.- Inicializar todas las distancias en *D* con un valor infinito relativo, ya que son desconocidas al principio, exceptuando la de *X*, que se coloca en 0 dado que para que ese mismo nodo vaya a si mismo seria 0.

2.- Se iguala a con *X* para indicar que es el nodo actual que trabajar.

3.-Se corren todos los nodos adyacentes de a, excepto los nodos marcados. Serán nodos no identificados V_i.

4.- En el nodo actual, se calcula la distancia tentativa desde dicho nodo, hasta sus vecinos adyacentes, con la fórmula:

$$dt(V_i) = Da + d(a, V_i)$$

Por lo que la distancia tentativa del nodo V_i es la distancia que actualmente tiene el nodo en el vector *D* más la distancia desde dicho nodo '*a*' hasta el nodo V_i. Si la distancia es menor que la distancia almacenada en el vector, se actualiza entonces el vector con la distancia tentativa $dt(V_i) < D(V_i) \text{---} > D(V_i) = dt(V_i)$

5.- Ese nodo se marca como recorrido.

6.- Se toma como próximo nodo actual, el de menor valor en *D* (almacenando los valores en una cola de prioridad) y se regresa al paso 3 para seguir con el ciclo hasta que se hayan marcado todos los nodos como recorridos.

Algoritmo de Floyd

El problema se puede resolver por medio del algoritmo de Dijkstra, aplicándolo a cada uno de los vértices, pero hay otra alternativa más directa, que es el algoritmo de Floyd. Sea *G* un grafo dirigido valorado, *G= (V, A)*. Se supone que los vértices están numerados de 1 a *n*; la matriz *A* es en este caso es la matriz de pesos, de tal forma que todo arco (i,j) tiene asociado un peso $c_{ij} \geq 0$; si no existe arco (i,j) se supone que $c_{ij} = \infty$. Ahora se quiere encontrar la matriz *D* de *nxn* elementos tal que cada elemento *D(i,j)* contenga el coste mínimo de los caminos que van del vértice i al vértice al vértice j. El proceso que sigue el algoritmo de Floyd tiene los mismos pasos que el algoritmo de Warshall para encontrar la matriz de caminos. Se generan iterativamente la secuencia de matrices $D_o, D_1, D_2, \ldots, D_k, \ldots, D_n$ cuyos elementos tienen el significado: $D_0[i,j] = C[i,j]$ coste (peso) del arco de *i* a *j*. Para todo *k* =1, 2, …, *n*, $D_k[i,j]$ contiene la longitud de un camino mínimo para ir del vértice *i* al vértice *j* usando los vértices 1,2,…,*k* . Para calcular $D_k[i,j]$ basta con

observar que el camino mínimo para ir del vértice i al vértice j o bien no usa el vértice *k* en cuyo caso. $D_k[i,j] = D_k\text{-}1[i,j]$ o bien lo usa en cuyo caso $D_k[i,j] = D_k\text{-}1[i,k] + D_k\text{-}1[i,k]$. Es decir $D_k[i,j]=$ minimo($D_k\text{-}1[i,j]$, $D_k\text{-}1[i,k]+D_k\text{-}1[k,j]$). La matriz D_n será la matriz de caminos mínimos del grafo.

Al igual que se hace en el algoritmo de Dijkstra por cada vértice se guardar el índice del último vértice que ha conseguido que el camino sea mínimo del *i* al *j* en caso de que el camino sea directo tiene un cero. Para ello se usa una matriz de vértices predecesores *P*. [21]

Algoritmo K-Means.

En el algoritmo K-Means proceso de agrupamiento es particional, es decir, forma subgrupos a partir de un grupo más general, inicialmente se determina el número de grupos *K* que se desean formar y se eligen los centroides. Para determinar los centroides iniciales, hay dos alternativas: la primera es tomar de forma aleatoria *K* objetos como centroides y la segunda es tomar los primeros *K* objetos del conjunto de objetos que se esté analizando. En el siguiente diagrama de flujo se explica el algoritmo. [21]

Herramientas para bases de datos

En principio se descartó la opción de usar una base SQL porque el esquema SQL es muy rígido y nuestro modelo de datos necesita ser flexible ya que no todos los elementos comparten las mismas características. Mientras que las bases de datos NoSQL ofrecen alto rendimiento en volúmenes grandes de información y se adaptan varios esquemas.

Base de datos orientada a grafos (Neo4j): En este tipo de bases de datos, la información se representa como nodos de un grafo y sus relaciones con las aristas, de manera que se puede hacer uso de la teoría de grafos para recorrerla. Esta estructura nos permite modelar todo tipo de escenarios definidos por entidades y relaciones entre sí: una red social, motores de recomendaciones en tiempo real, etc.

Los beneficios de utilizar una base de datos de estas características son:

- Flexibilidad: los datos no tienen que mantener una estructura rígida, los nodos pueden ser cada uno de un tipo diferente, se pueden añadir con facilidad atributos adicionales.

- Búsqueda: podemos hacer búsquedas muy rápidas basadas en las relaciones establecidas entre los nodos, por ejemplo: *"¿cuándo se han hecho amigas dos personas?"*.

- Indexación: las bases de datos de grafo se indexan de forma natural por sus relaciones, proporcionando un acceso más rápido en comparación a cómo resuelve una base de datos relacional las *foreign keys* entre diferentes tablas. [22]

Capítulo 5 Diseño

A continuación, se muestra el diseño del sistema a través del diagramado UML, compuesto por diagramas de clases, actividades y secuencia.

Diagrama de clases

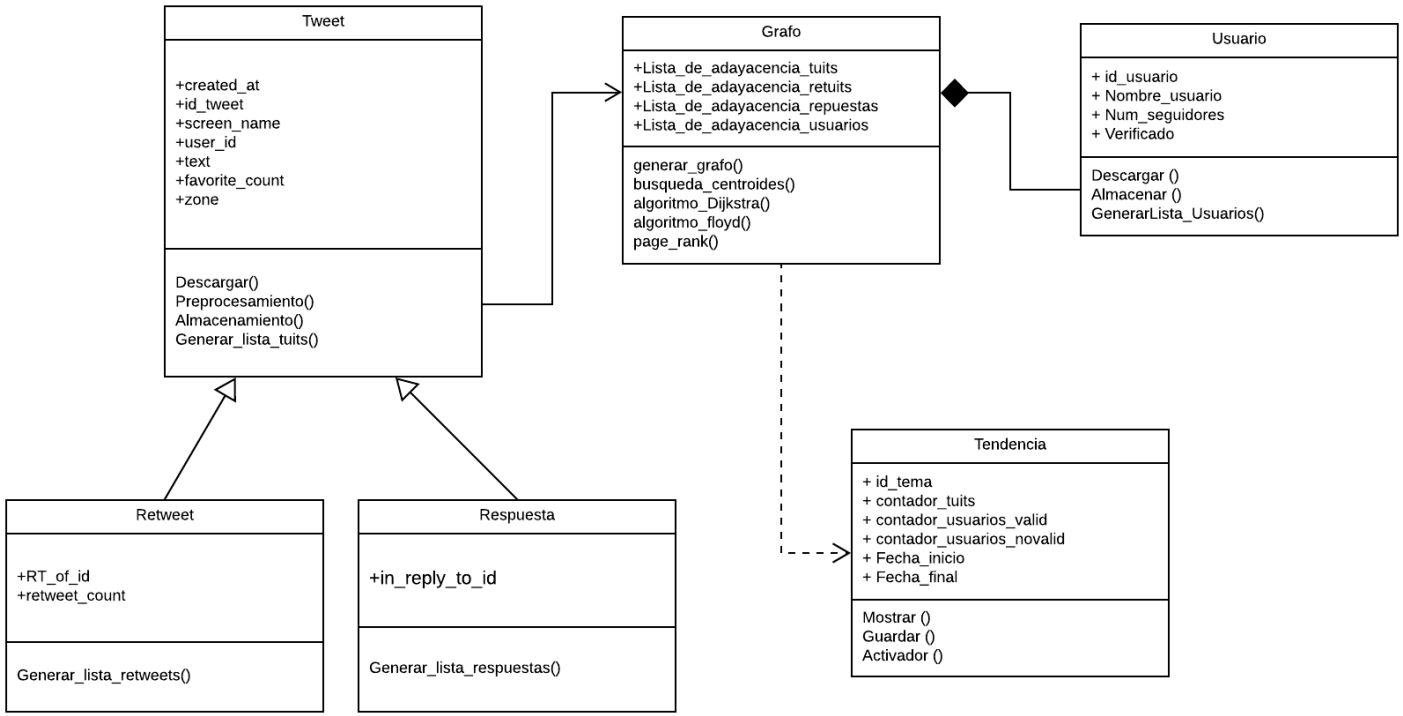


Ilustración 9 Diagrama de clases

Diagrama de Actividades

El sistema realizará, por día, la descarga de tuits relacionados con los temas de tendencia en la Ciudad de México. Una vez terminada la descarga de los tuits de todos los temas de tendencia en la ciudad, se aplicará un preprocesamiento del texto para facilitar el manejo de los datos, posteriormente se generan listas de adyacencia para usuarios y tuits con las que se construirán la base de datos, y además servirán para aplicar los algoritmos correspondientes para encontrar líderes de opinión y origen de la información en los temas de tendencia, en la Figura 4 se muestra el diagrama de actividades correspondiente a este sistema.

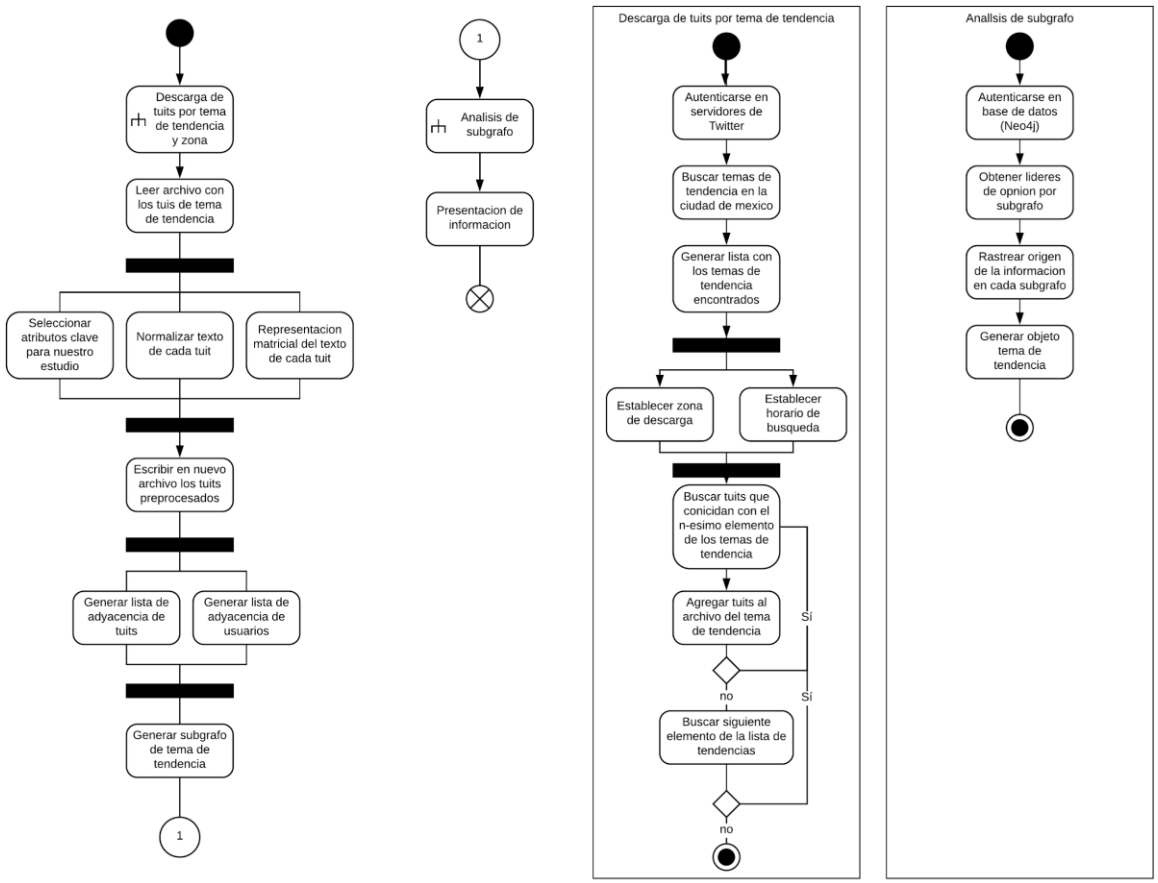


Ilustración 10 Diagrama de actividades parte del sistema

El usuario final accederá a una interfaz web que servirá para presentar los resultados del análisis realizado por nuestro sistema a los temas de tendencia en la Ciudad de México, la figura 5 muestra el diagrama de actividades para la interfaz web.

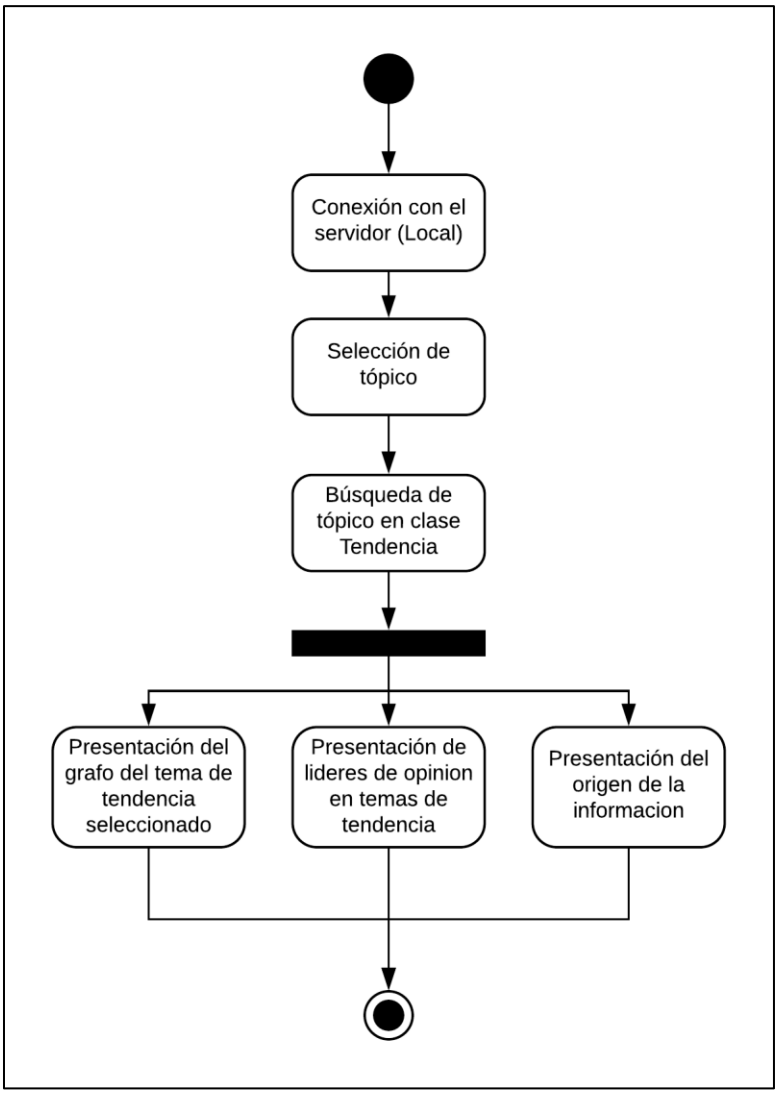


Ilustración 11 Diagrama de actividades modulo web

Diagrama de secuencia

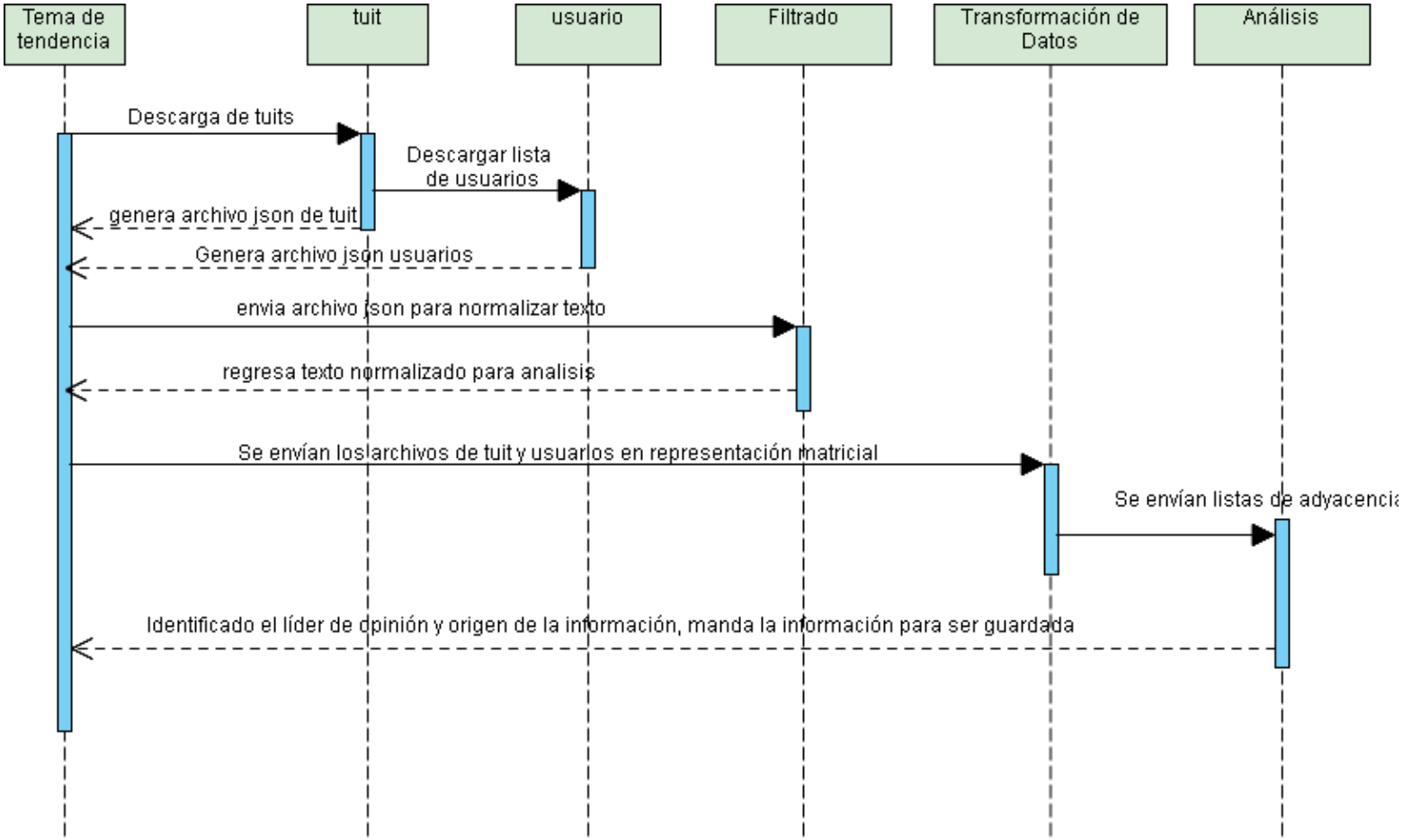


Ilustración 12 Diagrama de actividades parte del sistema

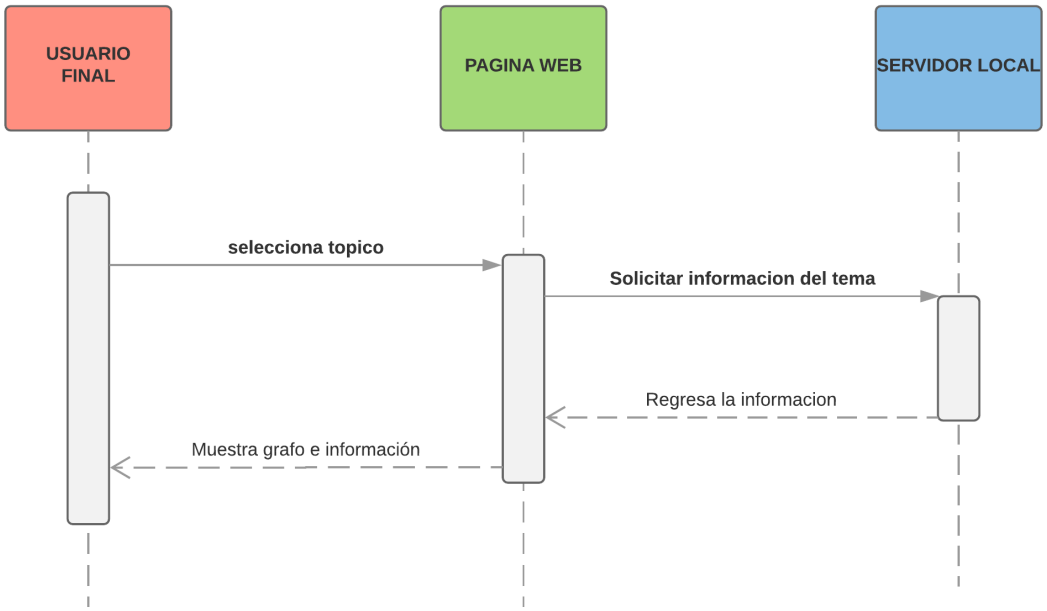


Ilustración 13 Diagrama de actividades parte web

Diseño de la base de datos

La base de datos que se usará en el sistema es una base de datos NoSQL orientada a grafos. En este tipo de bases de datos, la información se representa como nodos de un grafo y sus relaciones con las aristas, de manera que se puede hacer uso de la teoría de grafos para recorrerla. Este tipo de bases de datos son más flexibles en el contenido ya que cada nodo puede tener una estructura diferente y pueden existir distintos tipos de aristas para representar distintos tipos de relaciones entre los nodos.

En la base de datos hay tres tipos de nodos:

- Nodo azul: representa un tuit
- Nodo rosa: representa un *retweet*
- Nodo amarillo: representa una respuesta

Además, también tendremos tres tipos de aristas que indican cómo se relacionan los nodos:

- Arista *Follows*: Indica que el usuario que emitió ese *tweet* es seguidor del nodo al que apunta.
- Arista *Retweet_to*: Indica que ese nodo es un *retweet* (copia) del nodo al que apunta.
- Arista *Anwser_to*: Indica que ese nodo es una respuesta del nodo al que apunta.

La base de datos no pone ninguna restricción en cuanto a las relaciones, de modo que los nodos se interconectarán de acuerdo con los datos obtenidos desde Twitter. La ilustración 8 muestra un ejemplo de la base de datos, al ser un ejemplo se muestra un grafo pequeño y ordenado, sin embargo, un caso real estará compuesto por miles de nodos y cada nodo puede estar conectado a otros cientos o miles.

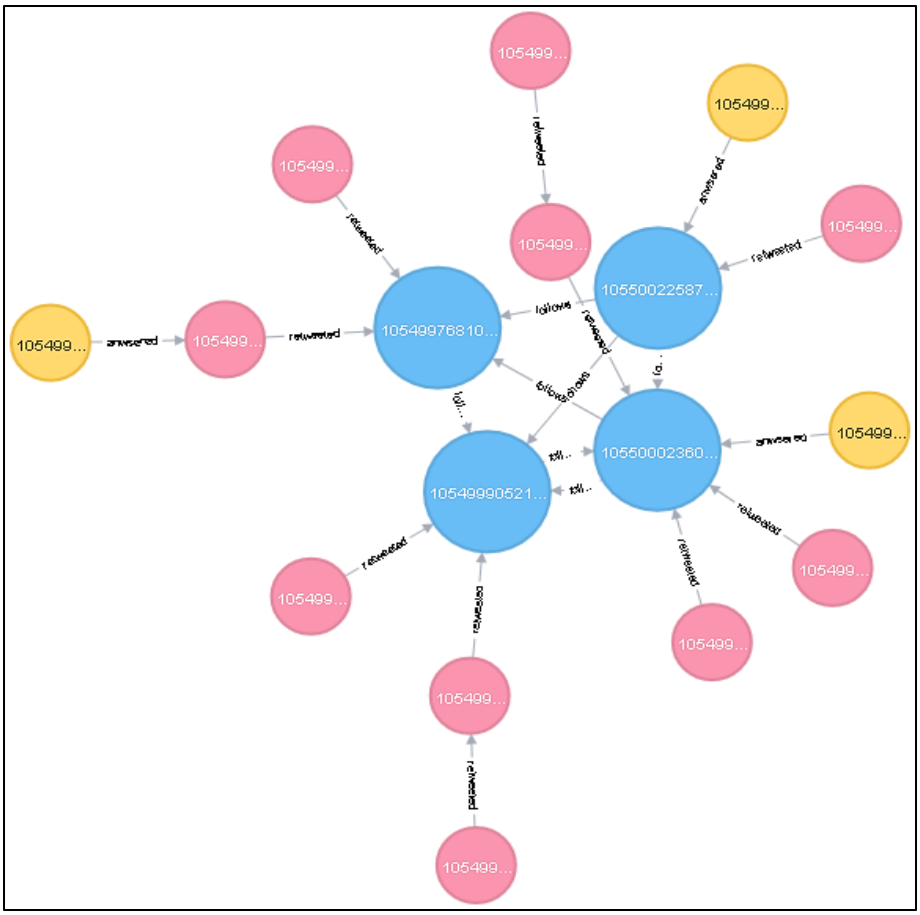


Ilustración 14 Base de datos en grafo

En la tabla 5 se muestra un ejemplo del contenido de un nodo, esta información resulta de filtrar los campos clave, para este estudio,

del archivo json que nos da Twitter con su API de descarga. Estos campos clave nos ayudarán a asignarle un peso a cada nodo para que el sistema defina cuales son los nodos con mayor influencia dentro del grafo.

Tabla 11. Contenido de un nodo


```
{"tweet_id": "1054997435662123008",

  "user_id": 2946980076,

  "created_at": "Wed Oct 24 07:26:04 +0000 2018",

   "text": "RT @mario_delgado1: La #ConsultaNAICM es el inicio de una nueva forma de tomar las decisiones más importantes del país; es la vocación demo…",

  "RT_id": "1054937160086806528",

  "favorite_count": 0,

  "retweet_count": 124

  "zone": "C1",}
```


Conclusiones

Para poder hacer uso de la API de Twitter se tiene que realizar un registro de aplicación, donde se detalle el uso que se le dará a la información obtenida de la red social, aun teniendo algunas limitaciones en el servicio gratuito ha sido suficiente para comenzar el proyecto. Se utilizarán las categorías “Cuentas y usuarios” y “Tweets” para obtener la información de los usuarios con respecto a sus publicaciones y para acceder a los tuits públicos disponibles en la plataforma Twitter, se filtrarán por dos parámetros, localización y tema de tendencia para realizar una base de datos con la información a analizar en un futuro.

Para obtener esa información se diseñó he implementó un sistema para descargar tuits. Este sistema se desarrolló en Python usando la librería *tweepy* que permite gestionar el API de Twitter de manera más sencilla.

Como se mencionó en la sección Propuesta de solución, con las primeras descargas en tiempo real se notó que la información obtenida no sería suficiente para el estudio. Por lo que se decidió descargar los tuits por zona y por tema de tendencia. Realizando las descargas de esta forma la información obtenida es suficiente para cumplir con los objetivos planteados. Algunas herramientas ya fueron probadas para la realización de este reporte. Sin embargo, el reto principal está en usar los algoritmos necesarios para identificar los líderes de opinión y las fuentes de información.

Para encontrar la similitud entre los textos se tendrán que usar diferentes algoritmos, debido a que existen 6 posibles situaciones que se deben considerar para tener una mayor certeza de que el resultado es correcto, no se puede escoger un único algoritmo a utilizar debido a estas problemáticas.

La herramienta de administración para la base de datos será Neo4j. Esta herramienta permite que la estructura de la información se pueda representar de mejor manera en forma de grafo. Los algoritmos de teoría de grafos que mejor se adaptan a nuestro esquema son el de Dijkstra y de Prim, ya que el de Dijkstra nos ayuda para encontrar el peso de la ruta entre pares de nodos y el de Prim nos ayuda a encontrar información de cómo se difunde un mensaje en el grafo. La base de datos que se diseñó está orientada a grafos, estableciendo el contenido de cada nodo como:

- Relación con su lista de seguidores
- Conteo de *retweets*
- Conteo de favoritos
- Es respuesta o no
- Fecha y hora de publicación

Se diseño el sistema apoyándonos de diagramas UML, como diagramas de casos de uso para definir las acciones que hará el usuario final. Diagrama de clases para definir las clases, objetos y las interacciones entre cada uno de ellos dentro del sistema. Diagrama de actividades que nos da información más detallada de las acciones que tiene que realizar el sistema para lograr los resultados esperados. Diagrama de secuencia que nos muestra cómo van a interactuar los objetos dentro del sistema para mostrar los resultados finales.

La herramienta BeeWare es la más completa para programar una página Web debido a su alto grado de compatibilidad con los lenguajes de programación y también por su compatibilidad con Python, además que la herramienta está dividida en módulos, lo que hace fácil un desarrollo al no tener que usar toda la herramienta de no ser necesario.

Para representar el grafo de forma gráfica usaremos las herramientas NetworkX y Graphviz ya que ofrecen muchas opciones para el dibujo de un grafo y las podemos adaptar a la necesidades que surgan con la base de datos.

Cronograma de actividades Proyecto terminal II

Tabla 12. Cronograma de actividades PT2

No.	Nombre de la actividad	Objetivos	Resultados esperados	Responsable
1	Limpieza de texto	Implementar un sistema que realice la extracción de atributos clave de cada tuit en los temas de tendencia.	Sistema para la extracción de atributos clave.	Alan Sánchez
2	Normalización de texto	Implementar un sistema que normalice el texto de cada tuit, además de pasar el texto a su representación matricial.	Sistema para la normalización de texto.	Arturo Ortiz
3	Información de usuarios	Implementar un sistema que obtenga la lista de seguidores de cada usuario y genere un archivo con los usuarios que participaron en cada tema de tendencia.	Sistema para la obtención de lista de seguidores y usuarios que participaron en cada tema.	Alan Sánchez
4	Generar subgrafos	Implementar un sistema que genere un subgrafo con todos los tuits descargados de cada tema de tendencia.	Sistema para generar subgrafos.	Alan Sánchez Arturo Ortiz
5	Obtención de líderes de opinión	Implementar un algoritmo que recorra cada subgrafo y encuentre el o los líderes de opinión en cada uno de los temas de tendencia y guarde los resultados.	Sistema para encontrar al o los líderes de opinión en cada tema de tendencia.	Alan Sánchez
6	Origen de la información	Implementar un sistema que recorra cada subgrafo para encontrar el origen de la información y guarde los resultados.	Sistema para encontrar el origen de la información en cada tema.	Arturo Ortiz
7	Conjuntar los sistemas	Conjuntar los sistemas antes implementados para que realicen sus funciones de forma automática.	Sistema conjunto que realice las tareas de forma automática.	Alan Sánchez Arturo Ortiz
8	Interfaz web	Implementar un sistema web para presentar la información obtenida del análisis.	Interfaz web.	Arturo Ortiz

Referencias

1. Carlos Arcila Calderón, Elias Said Hung, Diciembre (2012). Factores que inciden en la variación de seguidores en los usuarios top20 más vistos en twitter en américa latina y medio oriente. Venezuela. Revista Interciencia 2012 vol. 37.
2. Notimex. (16 de 03 de 2016). Twitter tiene 35.3 millones de usuarios en México. El universal.
3. A.L. Ortiz, O.E Pérez, E. Vargas, Estudio de tendencias diarias en twitter,Trabajo de fin del Grado de Ingeniería Informática, Madrid,España: Universidad Complutense Madrid, 2014-2015.
4. B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, Short text classification in twitter to improve information filtering, in Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval, 2010
5. Bautista Ruiz Agustín, Julio (2012). Extracción y clasificación de información del tráfico vehicular a través de Twitter y su visualización en Google Maps. Ciudad de México. UPIITA.
6. Saul Padilla Maya, Diciembre (2013). Integración de la red social de Twitter a un ambiente de TV interactiva. Ciudad de México, UPIITA.
7. H. Librado, Enero (2016),Análisis automático de opiniones de productos en redes sociales, Ciudad de México, CIC.
8. «hashtracking,» 2010. [En línea]. Disponible: https://www.hashtracking.com/. [Último acceso: 02 11 2017].
9. Castañeda, L. & Gutiérrez, I. (2010). Redes sociales y otros tejidos online para conectar personas. Aprendizaje con Redes Sociales. Tejidos educativos en los nuevos entornos. Sevilla: MAD Eduforma
10. Martínez-Cámara, E., García-Cumbreras, M.A., Villena-Román, J., & García-Morera, J. (2016). TASS 2015 - The The Evolution of the Spanish Opinion Mining Systems. Procesamiento del Lenguaje Natural, 56. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/issue/view/218.
11. C. Pérez López y D. Santín González, Minería de datos técnicas y herramientas, Madrid,España: Thomson, 2007.
12. Wirth, N., Fagoaga, J. C. V., & Vieyra, G. Q. (1987). Algoritmos y estructuras de datos (No. 04; QA76. 6, W56.). Prentice-Hall Hispanoamericana.

13. J. García Herrero y J. M. Molina López,2012 , «Técnicas de análisis de datos,» Universidad Carlos III de Madrid, Madrid, España.
14. Martín, S. Chávez , N. Rodríguez, A. Valenzuela y M. Murazzo, «Bases de Datos NoSql en Cloud Computing,» de XV Workshop de Investigadores en Ciencias de la Computación, Paraná,Argentina, 2013.
15. <<Algoritmos de Búsqueda en Anchura (BFS) y Búsqueda en Profundidad (DFS) >> [En línea]. Disponible: http://www.bibliadelprogramador.com/2014/04/algoritmos-de-busqueda-en-anchura-bfs-y.html [Último acceso: 05 11 2017]
16. J. G. A. Figueroa. (2012 marzo 19)<<CAMINO MÁS CORTO: ALGORITMO DE DIJKSTRA>> [En línea]. Disponible: https://jariasf.wordpress.com/2012/03/19/camino-mas-corto-algoritmo-de-dijkstra/ [Último acceso: 05 11 2017]
17. J. A. MONDRAGÓN, L. A. HERRERA, C. F. FERNANDEZ. (2010 noviembre 23)<<ALGORITMO DE BELLMAN- FORD>> [En linea] Disponible: https://upcanalisisalgoritmos.wikispaces.com/file/view/BELLMAN-FORD_GRUPO_5.pdf [Último acceso: 05 11 2017]
18. D. Pérez. <<Análisis y Diseño de Algoritmos>> [En linea] Disponible: http://delta.cs.cinvestav.mx/~adiaz/anadis/Graph2.pdf [Último acceso: 05 11 2017]
19. *Twiiter* <<Información sobre las API de Twitter>> 2018 [En línea] https://help.twitter.com/es/rules-and-policies/twitter-api [Ultimo acceso: 30/08/2018]
20. M. T. Jones, «IBM *Developer*,» IBM, 01 FEBRERO 2018. [En línea]. Disponible: https://www.ibm.com/developerworks/library/ba-intro-data-science-1/. [Último acceso: septiembre 2018].
21. M. C. V. y. A. S. M.D. Riba, «Aplicación de la Teoría de Conjuntos Borrosos»
22. M. F. A. L. S. G. I. Z. M. Luis Joyanes Aguilar, Estructuras de datos en C, Madrid: McGraw-Hill, 2005
23. Neo4j, «Neo4j,» 2018. [En línea]. Disponible: https://neo4j.com/developer/get-started/. [Último acceso: septiembre 2018].
24. <<Followerwonk>> 2018 [En linea] Disponible: https://followerwonk.com/analyze [Ultimo acceso: 24/08/2018]
25. <<Audiense>> [En linea] Disponible: https://audiense.com/solutions/ [Ultimo acceso: 25/08/2018]
26. <<Edelman intelligence>> 2018 [En linea] Disponible: https://www.edelmanintelligence.com/ [Ultimo acceso: 26/08/2018]

27. <<Topsy>> 2013 [En línea] Disponible: http://topsy.thisisthebrigade.com/ [Ultimo acceso: 29/08/2018]
28. <<Lithium>> 2018 [En línea] Disponible: https://www.lithium.com/company/pricing/ [Ultimo acceso: 29/08/2018]
29. << ¿Por qué los líderes de opinión influyen ahora más que nunca? >> 2011 [En línea] Disponible: https://www.puromarketing.com/42/11072/lideres-opinion-influyen-ahora-nunca.html [Ultimo acceso: 23/11/2018]
30. << ¿Cuánto contenido se genera por minuto en Twitter y Facebook? >> 2014 [En línea] Disponible: https://www.luismaram.com/cuanto-contenido-se-genera-por-minuto-en-twitter-y-facebook/ [Ultimo acceso: 23/11/2018]
31. << ¿Cuáles son las redes sociales más usadas en México? >> 2017 [En línea] Disponible: https://lifeandstyle.mx/tech/2017/05/30/cuales-son-las-redes-sociales-mas-usadas-en-mexico [Ultimo acceso: 24/11/2018]
32. << ¿Cuánto cuesta un "trending topic" ?: la investigación de la BBC que revela cómo hacen las empresas para manipular "hashtags" y crear tendencias en Twitter >> 2018 [En línea] Disponible: https://www.publimetro.com.mx/mx/bbc-mundo/2018/03/05/cuanto-cuesta-un-trending-topic-la-investigacion-de-la-bbc-que-revela-como-hacen-las-empresas-para-manipular-hashtags-y-crear-tendencias-en-twitter.html [Ultimo acceso: 24/11/2018]
33. << ¿Qué es MapReduce? >> 2012 [En línea] Disponible: https://blogs.solidq.com/es/big-data/que-es-mapreduce/ [Ultimo acceso: 25/11/2018]
34. << ¿Qué es una base de datos clave-valor? >>2018 [En línea] Disponible: https://aws.amazon.com/es/nosql/key-value/ [Ultimo acceso: 25/11/2018]
35. << ¿Qué es una base de datos de gráficos? >>2018 [En línea] Disponible: https://aws.amazon.com/es/nosql/graph/ [Ultimo acceso: 25/11/2018]
36. T. W. V. a. R. L. Davies, «Accelerating the Diffusion of innovation using opinion leaders» JSTOR, 1999.