



INSTITUTO POLITÉCNICO NACIONAL  
UNIDAD PROFESIONAL INTERDISCIPLINARIA EN INGENIERÍA Y TECNOLOGÍAS  
AVANZADAS



PROTOCOLO  
INGENIERÍA TELEMÁTICA

DTA-PPT-01

TÍTULO DEL PROTOCOLO

Sistema de análisis para la ponderación de usuarios de twitter en temas de tendencia

DATOS DEL PROTOCOLO

Número de Revisión (Primera, segunda, tercera o Protocolo para Registro)	Registro	Semestre	18-1
Número Proyecto Asignado (Número asignado por el profesor de Especialidad)	14	Fecha (Fecha programada)	18/12/17
Confidencialidad (Público o confidencial, incluir documento que lo avale)	Público	Número de Hojas (Cantidad de Hojas del Protocolo)	25
Patrocinador (En caso de existir, incluir el nombre en caso contrario dejar en blanco)			
Número Convenio o Registro (Incluir número de convenio patrocinio o número de proyecto de investigación que patrocina)			

ALUMNO 1

DATOS ALUMNO 1		FIRMA
Nombre del Alumno	Ortiz Romero Héctor Arturo	
Número de boleta	2013640153	
Teléfono		
Correo electrónico	Arturo240791@gmail.com	

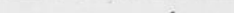
ALUMNO 2

DATOS ALUMNO 2		FIRMA
Nombre del Alumno	Sanchez López Alan Rodrigo	
Número de boleta	2013640297	
Teléfono		
Correo electrónico	thealan01@gmail.com	

ALUMNO 3

DATOS ALUMNO 3		FIRMA
Nombre del Alumno		
Número de boleta		
Teléfono		
Correo electrónico		

DATOS ASESOR 1

Nombre Asesor (Grado Académico)	Dra. Obdulia Pichardo Lagunas				
Academia	Informática	Interno	X	Externo	
Cédula Profesional (Obligatorio)	3532880				
Correo electrónico	aylina@hotmail.com				

VISTO BUENO ASESOR 1

DATOS ASESOR 2

Nombre Asesor (Grado Académico)	M. en C. Bella Citlali Martínez Seis					
Academia	Informática	Interno	X	Externo		
Cédula Profesional (Obligatorio)	6159023					
Correo electrónico	bellims@gmail.com					

VISTO BUENO ASESOR 2

DATOS ASESOR 3

Nombre Asesor (Grado Académico)					
Academia		Interno		Externo	
Cédula Profesional (Obligatorio)					
Correo electrónico					

VISTO BUENO ASESOR 3

NOMBRE DEL PROFESOR TITULAR Lic. Sandra Martínez Solís  
NOMBRE DEL PROFESOR DE ESPECIALIDAD M. en C. Cyntia Eugenia Enríquez Ortiz

(En caso de Asesores Externos, deberá incluir, además de su Cédula Profesional y Currículum Vitae resumido en un archivo anexo al Protocolo)

**AVISO DE PRIVACIDAD** Este aviso de privacidad tiene como finalidad informar a los usuarios de la Unidad Profesional Interdisciplinaria en Ingeniería y Tecnologías Avanzadas (UPITA-IPN) sobre la finalidad, el uso, la recolección, el almacenamiento, la conservación, el acceso, la divulgación, la transferencia, la actualización, la modificación, la eliminación, la destrucción, la portabilidad, la rectificación, la cancelación, la suspensión o limitación de la información, la oposición, el consentimiento, el ejercicio de los derechos de acceso, rectificación, cancelación y oposición, y la responsabilidad de la UPITA-IPN en materia de protección de datos personales. Este aviso de privacidad es de carácter público y está dirigido a todos los usuarios de la UPITA-IPN, tanto internos como externos. El presente aviso de privacidad es parte integrante del Protocolo de Datos Personales, por lo que se debe leer en su totalidad. La UPITA-IPN se reserva el derecho de modificar este aviso de privacidad en cualquier momento sin necesidad de avisar previamente. Los cambios serán publicados en la página web de la UPITA-IPN y en el presente protocolo. La última versión de este aviso de privacidad es la que prevalece.



**Instituto Politécnico Nacional**  
**Unidad Profesional Interdisciplinaria en**  
**Ingeniería y Tecnologías Avanzadas**



*Sistema de análisis para la ponderación de  
usuarios de twitter en temas de tendencia*

**Alumnos:**

*Ortiz Romero Héctor Arturo*

*Sánchez López Alan Rodrigo*

**Asesores:**

Nombre: Dra. Obdulia Pichardo Lagunas Procedencia: UPIITA-IPN Academia de informática
Nombre: M. en C. Bella Citlali Martínez Seis Procedencia: UPIITA-IPN Academia de informática

**Resumen:**

*Este proyecto tiene como finalidad realizar la ponderación y evaluación de la relación existente 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, procesamiento de lenguaje natural, grafos, minería de datos, aprendizaje automático.*

**Versión 3**

**Diciembre 2017**

# Contenido

<b>Introducción .....</b>	<b>4</b>
<b>Planteamiento del problema .....</b>	<b>5</b>
<b>Propuesta de solución.....</b>	<b>6</b>
<b>Alcances (Resultados esperados) .....</b>	<b>8</b>
<b>Objetivo general .....</b>	<b>9</b>
<b>Objetivos específicos .....</b>	<b>9</b>
<b>Estado del arte.....</b>	<b>10</b>
<b>Marco teórico .....</b>	<b>13</b>
<b>Escenario de pruebas.....</b>	<b>20</b>
<b>Cronograma de actividades.....</b>	<b>21</b>
<b>Referencias .....</b>	<b>24</b>

## 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. Pero además de darle un uso tan cotidiano a este servicio, como puede ser el hecho de contactar con amigos o compartir información con el resto del mundo, se puede ir más lejos y sacarle partido a todos esos datos. Gracias a potentes técnicas de minería de datos, extrayendo conocimiento de manera automática o realizando predicciones sobre comportamientos o hechos futuros.

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 esta tiene que escribirse junta y sin espacios.

En el estudio [1] se obtuvo como resultado que en América Latina la comunidad de Twitter está seccionada en roles como usuario cotidiano 62.5%, reporteros 36.3%, Otros (innovadores, consumidores, diarios de información) 1.3%, mientras que los tipos de usuarios son personal 25% y el 75% de tipo profesional, con los datos del usuario se descubrió que el 35% de la comunidad son de profesión artística (actores, cantantes, presentadores, escritores, fotógrafos, etc.); mientras que el 32.5% son periodistas, dejando a 3.8% como medios de comunicación, 1.3% son académicos, el 5% funcionarios públicos, 1.3% futbolistas, 2.5% políticos y solo 18.8% de la comunidad no pudo ser identificada. Además, se puede saber el contexto o el tema del que se habla, con base en ello, se detectó que el 32.5% está en el área de acción de comunicación, el 21.3% en moda/espectáculo, el 15% en arte, 1.3% en ciencias, 1.3% en literatura/comunicación, 7.5% en política, finalmente el 17.5% restante no fue identificado; la mayor cantidad de usuarios son masculinos con el 58.8%, mientras que solo el 15% son mujeres y el 26.3% restante son instituciones, las preferencias de privacidad de los usuarios corresponde al 77.5% que prefieren que sus publicaciones sean anónimas y el 22.5% de forma pública.

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 quienes 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 quienes 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, Este trabajo se enfoca en la relación que existe entre los usuarios y los tuits haciendo un análisis estructural de la red, para lograrlo se realizará un estudio de los tuits relacionados a cada tema de tendencia, se descargarán los tuits en tiempo real por medio de robots programados acorde a las necesidades del estudio hasta tener dos bases de datos de tuits, una base de datos descargada por temas de tendencia y la otra descargara todos los tuits para posteriormente aplicarle análisis de texto, este análisis permitirá saber qué usuario lo emitió, en que momento lo emitió, cuantos seguidores tiene el usuario y qué efecto causó su participación en el tema. De esta forma se identificará en qué temas influyen más los líderes de opinión y también que temas de tendencia se crean de forma espontánea.

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 retuits que obtuvo cada tuit. La extracción se hará de dos formas: por hashtag y por zona. La extracción por hashtag obtiene todos los tuits que contengan un hashtag relacionado a los temas de tendencia en la Ciudad de México, es decir que en este modo se programará al robot para que usen como filtro los hashtags. Esta información se irá almacenando en una base de datos de tipo NoSQL que se denominará base de datos por hashtag y la de zona descargará los tuits por ubicación en la Ciudad de México, sin ningún tipo de filtro en el texto, esto porque hay algunos usuarios que participan en el tema de tendencia

sin usar el hashtag, esta información se irá almacenando en una base de datos de tipo NoSQL que se denominará base de datos por región.

- **Filtrado**

En la etapa de filtrado se aplicarán algoritmos para evaluar la similitud de textos, por ejemplo el algoritmo Knuth-Morris-Pratt o algoritmo de Boyer-Moore, en la base de datos por región con el fin de organizar los tuits que tengan relación con algún tema de tendencia y descartar los tuits que no guarden relación con ningún tema de tendencia en la Ciudad de México.

- **Transformación de datos**

La etapa de transformación de datos se divide en dos módulos, en la primera etapa se obtendrá el nombre del usuario en cada tuit, el número de retuits y los seguidores de esos usuarios. La segunda etapa consistirá en generar subgrafos donde cada tuit será un nodo y las aristas representan las conexiones entre los usuarios. Estos subgrafos serán almacenados en un 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

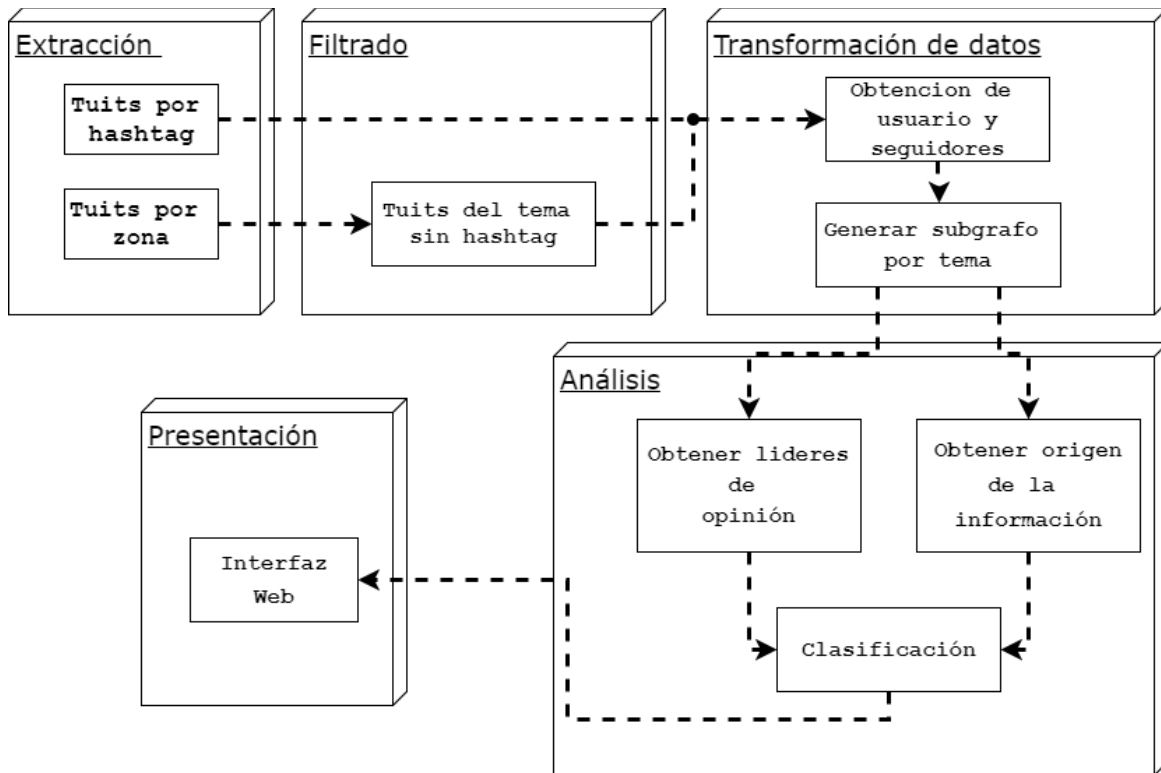
Obtención de líderes de opinión, este proceso se realizará mediante el análisis de la centralidad de cada nodo, se considerará que un nodo es prominente si sus enlaces hacen que este nodo sea particularmente visible para los demás nodos del subgrafo.

Clasificación, consistirá 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 1.



**Diagrama de Bloques 1. Procedimiento de 5 etapas para la propuesta de solución al problema.**

### **Alcances (Resultados esperados)**

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 retuits qué personas se identifican como líderes.

La base de datos a utilizar de tuits descargados será de aproximadamente 6 meses, realizará descargas diarias por medio de robots que almacenarán los datos en una computadora de escritorio, esto debido a que cada robot debe estar trabajando todo el tiempo en la descarga para tener una base de datos completa para futuras pruebas. 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 Megs 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.

## 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 hashtag.
- Implementar un sistema que descargue todos los tuits publicados en la región de la Ciudad de México.
- Pre-procesar 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 a través del número de retuits y respuestas que tenga el tuit de dicho usuario en el tema de tendencia.
- 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.

## 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 retuits, búsquedas o impresiones del hashtag que se busquen en específico, con sólo colocar el hashtag en la búsqueda se obtiene la información que ayudan a las empresas a darle marketing a sus marcas para poder decidir qué hashtags son ideales para promocionar, se conocerá cuántas veces se ha utilizado un determinado hashtag para un mensaje o para un retuit, además muestra el número absoluto de las impresiones que logra el hashtag.

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
Traducciones	Completas	Exportables	Exportables	Exportable	Exportables
Intercambio de Informes	Públicos	Ilimitado	Ilimitado	Ilimitado	Ilimitado
Cantidad de archivos	Ilimitados	Ilimitados	Ilimitado	Ilimitado	Ilimitado
Informe de seguimiento	Ilimitado	Ilimitado	Ilimitado	Ilimitado	Ilimitado

Fuente: «hashtracking,» 2010. [En línea]. Disponible: <https://www.hashtracking.com/>. [Último acceso: 02 11 2017].

En la unidad académica 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, usando 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, haciendo una comparación con cadenas de texto para identificar si el mensaje está relacionado con el tema de tráfico vehicular, identificados 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 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. En 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; se concluye que la 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, cuales son los mensajes más populares y cómo se relacionan los temas de tendencia entre sí, también como ha sido la evolución en el tiempo de los temas de tendencia más importantes.

Tabla 2. Tabla comparativa de los trabajos encontrados y parecidos.

Nombre	Obtener Perfil de	Clasificación por tema	Análisis de sentimientos	Análisis estructural de la	Análisis de la relación
--------	-------------------	------------------------	--------------------------	----------------------------	-------------------------

	usuario			red a través de subgrafos	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	✗	✓	✗	✗	✓
Factores que inciden en la variación de seguidores en los usuarios top20 más vistos en twitter en américa latina y medio oriente	✓	✓	✗	✗	✓
Integración de la red social de Twitter a un ambiente de TV interactiva	✗	✓	✗	✗	✓
Hashttracking	✓	✓	✗	✗	✓
Estudio de tendencias diarias en Twitter	✗	✓	✗	✗	✓
Sistema de análisis para la ponderación de usuarios de twitter en temas de tendencia	✓	✓	✗	✓	✓

## Marco teórico

### *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 hashtag 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 retuis 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, que 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].

### ***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 a 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:

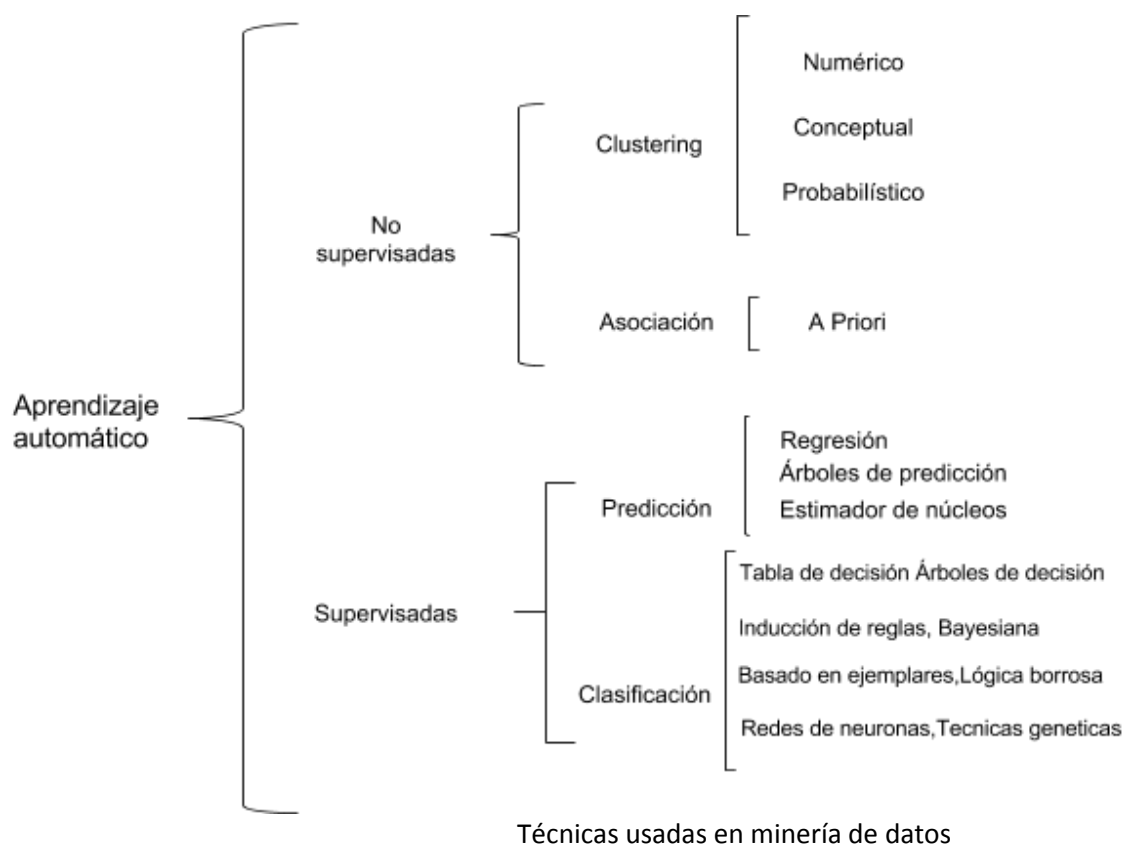
- Algoritmo Knuth-Morris-Pratt (KMP) se basa en usar técnicas de precondicionamiento con autómatas para poder encontrar de manera eficiente la ocurrencia de un patrón  $p$  en una cadena dada con coste en tiempo de preprocesado y de ejecución lineal [12].
- Algoritmo de Boyer-Moore El algoritmo se desliza 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 ventana 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 auto-organización, 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].



**Imagen 1. Técnicas usadas en Minería de Datos [J. García Herrero y J. M. Molina López, 2012, «Técnicas de análisis de datos», Universidad Carlos III de Madrid, Madrid, España.]**

### ***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 co-ocurrencias 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:

- Tabla de decisión: “Este tipo de algoritmos consisten en seleccionar subconjuntos de atributos y calcular su precisión para predecir o clasificar los ejemplos. Una vez seleccionado el mejor de los subconjuntos, la tabla de decisión estará formada por los atributos seleccionados (más la clase),

en la que se insertarán todos los ejemplos de entrenamiento únicamente con el subconjunto de atributos elegido. “[13]

- Árboles de decisión: “El aprendizaje de árboles de decisión está englobado como una metodología del aprendizaje supervisado. La representación que se utiliza para las descripciones del concepto adquirido es el árbol de decisión, que consiste en una representación del conocimiento relativamente simple y que es una de las causas por la que los procedimientos utilizados en su aprendizaje son más sencillos que los de sistemas que utilizan lenguajes de representación más potentes, como redes semánticas, representaciones en lógica de primer orden etc.”[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]

### ***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), Por ejemplo Hadoop.

- Almacenamiento Clave-Valor (sistemas que tienden al procesamiento de transacciones online - OLTP), Por ejemplo: Google BigTable, Amazon Dynamo, Cassandra, Voldemort, HBase.
  - Almacenamiento de Documentos Por ejemplo: CouchDB, MongoDDB, SimpleDB.
  - Sistemas de base de datos Gráficas. Por ejemplo: Neo4j, FlockDB, Pregel.
- [14]

### **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 (edges en inglés) que pueden ser orientados o no. Típicamente, un grafo se representa mediante una serie de puntos (los vértices) conectados por líneas (las aristas).

### **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].

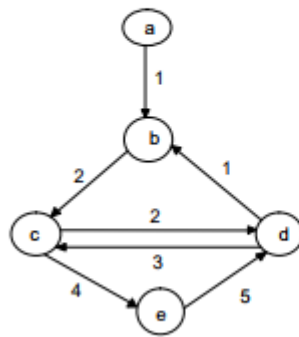
### **Centro de un Grafo**

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 cualesquiera 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]

## El Centro de un Grafo



A	a	b	c	d	e
a	0	1	3	5	7
b	$\infty$	0	2	4	6
c	$\infty$	3	0	2	4
d	$\infty$	1	3	0	7
e	$\infty$	6	8	5	0
max	$\infty$	6	8	5	7

el centro el grafo es d

Imagen 2. Ejemplo de encontrar el Centro de un Grafo por algoritmo de Floyd

## Escenario de pruebas

La base de datos será generada con los tuits descargados por robots durante 6 meses, de esos tuits se debe recuperar solo el texto descartando el contenido multimedia como imágenes videos y de audio.

La interfaz web estará montada en un servidor local, de modo que se podrá acceder a ella por medio de una red LAN.

Los resultados del análisis se presentarán en forma de estadísticas, mostrando que usuarios participaron en los distintos temas de tendencia, en qué temas de tendencia participan más ciertos líderes de opinión, y qué perfil tienen los usuarios considerados líderes de opinión en ciertos temas, por el número de seguidores y retuits que tenga.

Los tuits podrán obtenerse por un corpus de entrenamiento ya existente por ejemplo el corpus del TASS 2015[10], a través de la implementación de librerías de minería de datos para lenguajes como JAVA o Python dependiendo de la información que nos permita obtener cada herramienta.

Para validar nuestro método, se usará el 30% de la información obtenida, así que el 70% de la información se utilizará para entrenar el sistema.

## Cronograma de actividades

**Tabla 3.** Cronograma de actividades para Proyecto Terminal I.

No.	Nombre de la actividad	Objetivos	Resultados esperados	Responsable
1	Identificar los datos que Twitter nos permite descargar con cada tuit.	Identificar la información que nos da cada tuit.	Tener claros los atributos para cada nodo	Arturo Ortiz
2	Investigar acerca de la forma de recolectar los tuits por hashtag y por región.	Tener los conocimientos para generar las bases de datos.	Tener los conocimientos para poder implementar los robots que descarguen los tuits, uno por hashtag y otro por región.	Alan Sánchez
3	Crear un robot (robot de hashtag) capaz de descargar los tuits por hashtag	Crear una herramienta que descargue tuits diariamente por hashtag.	Generar la base de datos de tuits con hashtag donde se almacenen los tuits por intervalo de tiempo y tema de tendencia.	Arturo Ortiz

4	Crear un robot (robot por región) capaz de descargar los tuits por ciudades para generar la base de datos a trabajar.	Generar una herramienta que descargue tuits diariamente por zona.	Generar la base de datos de tuits por región donde se almacenen los tuits de la plataforma en un intervalo de tiempo y por ciudad.	Alan Sánchez
5	Investigar sobre algoritmos de similitud en texto	Identificar el o los algoritmos usando Bayes para clasificar la información.	Resumen de los algoritmos a implementar.	Arturo Ortiz
6	Análisis y selección algoritmos para la obtención de centroides	Encontrar un algoritmo que sea funcional para implementar	Tabla comparativa de los algoritmos investigados	Alan Sánchez
7	Investigar algoritmos para el análisis de difusión por conexión de aristas	Encontrar un algoritmo que sea funcional para analizar el grafo	Tabla comparativa de los algoritmos investigados	Arturo Ortiz
8	Análisis de herramienta para la implementación del sistema	Plantear múltiples enlaces para tener el grafo completamente conectado.	Mantener el grafo lo más estrecho posible para tener una relación entre nodos accesible.	Alan Sánchez
9	Análisis de manejadores de datos no-sql	Tener la estructura de los nodos a utilizar basándonos en los atributos dados por tuit.	Tener en código la estructura de los nodos.	Arturo Ortiz

10	Diseño del sistema	Definir la estructura del sistema	Diagramas UML	Alan Sánchez
			Diagramas de BD	Arturo Ortiz

**Tabla 4.** Cronograma de actividades para Proyecto Terminal II.

No.	Nombre de la actividad	Objetivos	Resultados esperados	Responsable
1	Análisis de texto	Agrupar los tuits por tema de tendencia	Base de datos por cada tema de tendencia	Alan Sánchez
2	Generar perfil de tuit	Obtener nombre de usuario y relación entre usuarios	Tener el perfil para cada tuit	Arturo Ortiz Alan Sánchez
3	Análisis de tendencias	Generar subgrafo por cada tema de tendencia	Conjunto de subgrafos de los temas de tendencia a analizar	Arturo Ortiz
4	Obtener origen de la información	Aplicar un algoritmo para encontrar el origen de la información	Obtener el origen de la información en cada tema analizado	Alan Sánchez
5	Obtener líderes de opinión	Aplicar un algoritmo para obtener los líderes de opinión por cada tendencia	Obtener los líderes de opinión en cada tema analizado	Arturo Ortiz
6	Interfaz web	Presentar los resultados a través de una interfaz web	Desarrollar una interfaz web para la presentación de la información	Alan Sánchez 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. «hashttracking,» 2010. [En línea]. Disponible: <https://www.hashttracking.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 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. A. 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. (2014 abril 29).<<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 línea] Disponible: [https://upcanalisisalgoritmos.wikispaces.com/file/view/BELLMAN-FORD\\_GRUPO\\_5.pdf](https://upcanalisisalgoritmos.wikispaces.com/file/view/BELLMAN-FORD_GRUPO_5.pdf) [Último acceso: 05 11 2017]
18. A. D. Pérez. <<Análisis y Diseño de Algoritmos>> [En línea] Disponible: <http://delta.cs.cinvestav.mx/~adiaz/anadis/Graph2.pdf> [Último acceso: 05 11 2017]