**ASIGNATURA:** Inteligencia de negocio

**TÍTULO:** Informe de proyecto final

**FECHA DE ENTREGA:** 20/02/2018

**PERIODO ACADÉMICO:** 2017B

**INTEGRANTES:**

Cisneros Paul

Macías Gabriel

**CLASIFICADOR CONSULTA POPULAR**

**Contenido**

[Introducción 1](#_Toc506900824)

[Machine learning 1](#_Toc506900825)

[Análisis de sentimiento 2](#_Toc506900826)

[Método 2](#_Toc506900827)

[Adquisición de datos 2](#_Toc506900828)

[Preprocesamiento 2](#_Toc506900829)

[Procesamiento 3](#_Toc506900830)

[Aprendizaje supervisado 3](#_Toc506900831)

[Expresiones regulares 3](#_Toc506900832)

[Clasificador 3](#_Toc506900833)

[Análisis 3](#_Toc506900834)

[TextBlob 3](#_Toc506900835)

[Discovery 3](#_Toc506900836)

[Presentación 5](#_Toc506900837)

[25trends 5](#_Toc506900838)

[Twitter 5](#_Toc506900839)

# 

# Introducción

Machine learning (aprendizaje de maquina) es parte de la inteligencia artificial, para otorgarles la capacidad de aprender por sus propios medios, esto quiere decir que no necesitados estar programando consecutivamente.

El aprendizaje de maquina se toma como un gran aporte para el análisis estadístico por el uso de las computadoras, una de os mayores aplicativos es el realizar predicciones computacionales de datos previamente obtenidos de alguna fuente determinada.

Se puede decir que el aprendizaje de maquina está fuertemente relacionado también con la optimización matemática, esto se debe a que aporta con métodos, teoría y dominios de las aplicaciones para este campo.

El aprendizaje de maquina se combina con la extracción de datos, donde está más centrado en el análisis exploratorio de datos y se conoce como aprendizaje no supervisado.   
Machine Learning también se puede supervisar y utilizar para aprender y determinar perfiles de comportamiento básicos para diferentes entidades y luego encontrar anomalías importantes.

Análisis de sentimiento es el campo de estudio que analiza las opiniones, sentimientos, evaluaciones, actitudes de las personas, y emociones hacia entidades tales como productos, servicios, organizaciones, individuos, problemas, eventos, temas y sus atributos.

El análisis de opinión y la minería de opinión se centra en las opiniones que expresan o implican sentimientos positivos o negativos.

Se aplicará machine learning en la reacción de grandes masas en un sistema geográfico determinado mediante un análisis eficiente y eficaz que tome en cuenta los sentimientos para analizar los datos recogidos de la red social Twitter y obtener como resultado la inclinación política en general.

En el trabajo que se realizó se utilizó la información generada por los usuarios de la red social Twitter, para lograr realizar un estimado del pensamiento en masa de los votantes, donde la unidad fundamental del trabajo fueros los tweets especialmente el texto de cada uno de ellos. Las pautas del trabajo realizado fueras las de tener en cuenta únicamente los tweets relacionados a la consulta popular y claramente que sean del idioma español, además contamos con la clasificación de los tweets dentro de las 3 ciudades mas importantes del Ecuador a nuestra consideración (Quito, Guayaquil, Cuenca); Los tweets fuente primordial para este trabajo pasan un trabajo arduo y selectivo donde se les prepara y deja listo para ser etiquetados en función a que si están a favor de la consulta popular o en contra de ella.

El impacto de un clasificador y analizador de sentimientos es muy importante porque abarca temas relacionados entre sí como la sociología, las finanzas y para entender el comportamiento de las personas. Las aplicaciones a nivel de empresa van desde tener conocer a que mercados abordar y explotar, hasta pronosticar el comportamiento de las personas después de que se hayan regulado o cambiado alguna ley.

Existen varias herramientas cuando se trata del análisis de sentimientos tanto como libres o sean pagadas dentro de la investigación hacemos caso a Discovey servicio de Watson-IBM que seria el asistente virtual logramos encontrar preguntas que jamás nos habríamos planteado o llegar a relacionar los resultados y las preguntas a información concreta por ejemplo lograr hacer que después de leer un tweet la maquina sea capaz de reconocer a que pregunta del referéndum hace referencia el texto, esto ayudaría a los analistas a adelantarse a los resultados y poder tomar decisiones convenientes para su trabajado.

# Método

## Adquisición de datos

Los datos para analizar serán cosechados de Twitter que se publicaron en las ciudades más importantes del país. Para esto se ocupará CouchDb como gestor de base de datos que por medio de un script realizado en Python que almacena en tiempo real.

Las coordenadas que usar son las siguientes:

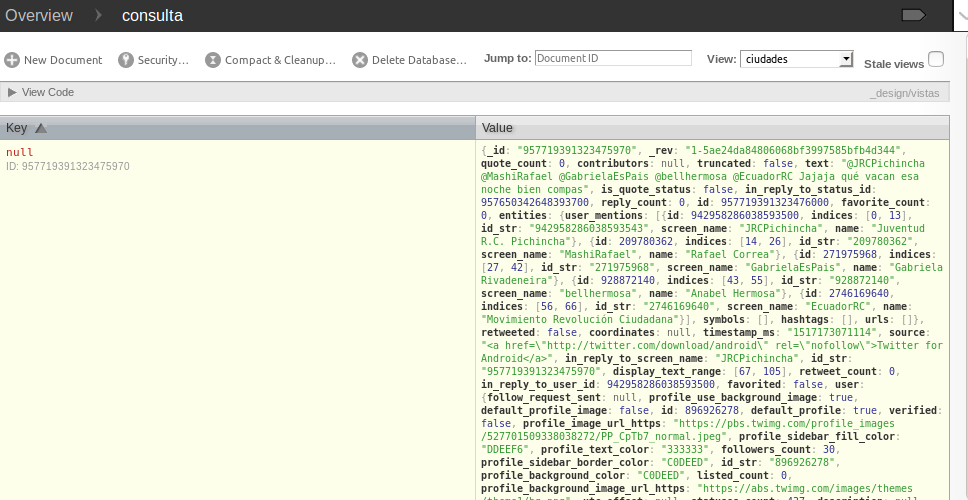
Quito: -78.332361, -0.371299, -78.546594, 0.020773

Guayaquil: -79.850098-2.282060, -80.078751, -2.026808

Cuenca: -78.927926, -2.922271, -79.056500, -2.851636

El periodo del cual los datos se obtuvieron es de la semana del cual se realizó la consulta popular del 01/02/2018 hasta 07/02/218. En el cual se recolecto los tweets creados por los usuarios que diariamente crean en el transcurso de los días.

El cosechador se creó con el código de Python con el fin de capturar los tweets dependiendo de las coordenadas especificadas para cada ciudad en especifico en la que se almacenaba cada una en la base de datos noSQL llamada CouchDb la cual íbamos trabajando a lo largo del semestre en la materia de inteligencia de negocios.

Realizando un seguimiento a los Tweets notamos que el 45% de los tweets almacenados son de la ciudad de Quito capital del Ecuador y la ciudad que utiliza mucho twitter como red social, en segundo lugar tenemos a la ciudad de Guayaquil con una cantidad también considerable de tweets de 30%, completando el podio esta Cuenca la cual llegaría 15% de la base de datos, el porcentaje restante dispone de ciudades del ecuador que el cosechador almaceno porque están en el rango de las coordenadas establecidas. La estructura de un Tweet viene en forma de documento de la cual se puede obtener la siguiente información:

{

"\_id": "957734014022029318", // el identificador del tweet para no tener repetidos en la base de datos

"text": "Aparte de la Consulta cuyos resultados no son suyos este es el\nResumen de 9 meses del gobierno ROBOlucionario e ile… https://t.co/gZ3MmUyihN" , // El texto para realizar el análisis que al comienzo se recibe algunos datos que no son alfanuméricos. Texto que debe ser acorde a la consulta

},

"created\_at": "Thu Feb 1 21:55:57 +0000 2018", La fecha es importante para revisar que se encuentre en tiempos de consulta.

"place": {

"country": "Ecuador",

"name": "Quito"

}

}, La coordenada es muy importante porque lo que se buscara es la tendencia de tweets en las 3 ciudades principales.

El código se encuentra tanto en el github como en el aula virtual.

https://github.com/paulsteban/ProyectoBI2017B/invitations

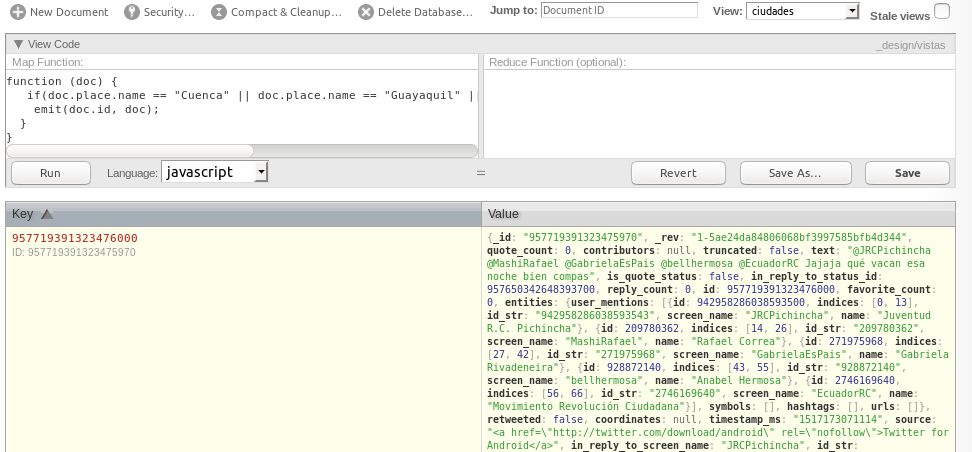
## Preprocesamiento

Para filtrar los datos obtenidos de Quito, Guayaquil y Cuenca se lo pude realizar directamente en CouchDb.

El código que se muestra a continuación fue usado para realizar vistas, las cuales permiten filtrar el texto, y obtener solo los que son en español.

El código que se muestra a continuación fue usado para realizar vistas, las cuales permiten filtrar el texto, y obtener solo los que son en español y de la ciudad que necesitemos obtener gracias a los campos lang que es para el idioma del tweet este es “es” de español como place.name donde se ingresa el nombre de la ciudad a buscar, aprovechamos también la función emit para etiquetar el texto con su identificador como a su vez usar expresiones regulares para eliminar los caracteres que no sean alfanuméricos como un problemas con las tildes de algunos tweets.

Para filtrar los datos por ciudad, debido a que CouchDb no necesariamente recoge datos específicamente de las coordenadas indicadas, por lo tanto, usamos los siguientes códigos para cada ciudad.



function (doc) {

if(doc.place.name == "Quito" && doc.lang == "es”){

emit(doc.\_id, {text: doc.text.replace(/[í]/gi, 'i').replace(/[^\w\s#.!]/gi, '')});

}

}

Para la vista de Guayaquil

function (doc) {

if(doc.place.name == "Guayaquil" && doc.lang == "es”){

emit(doc.\_id, {text: doc.text.replace(/[í]/gi, 'i').replace(/[^\w\s#.!]/gi, '')});

}

}

Para la vista de Cuenca

function (doc) {

if(doc.place.name == "Cuenca" && doc.lang == "es”){

emit(doc.\_id, {text: doc.text.replace(/[í]/gi, 'i').replace(/[^\w\s#.!]/gi, '')});

}

}

## Procesamiento

Aquí vamos a obtener los tweets que tengan como tema la consulta popular, esto realizar con el análisis de el texto que tiene cada tweet, para posteriormente ser clasificados como apoyo para el SI o el NO de dicha popular.

Los tweets no son una fuente confiable de información cuando se trata de análisis de sentimientos, sarcasmo de las personas para expresar sus pensamientos son cosas que la maquina no puede entender a simple vista, así mismo usar los hastag de los tweets o frases comunes para lograr distinguir los textos de la consulta popular es poco efectivo ya que note el patrón que se estaba formando de que se estaba sesgando los datos y tomando más a la conveniencia más que por análisis riguroso, los tweets de la consulta apoyando al no eran los mas difíciles de clasificar por la complejidad de los textos que estaban enriquecidos de mucha información tal que no se podía encontrar algo en común con los demás.

Dado a eso se llego a cernir los tweets uno por uno para guardar en una archivo json solo los tweets relacionados a la consulta

Aprendizaje supervisado

Para esto vamos a realizarlo con una técnica de aprendizaje supervisado, el código a usar fue realizado en Python.

Expresiones regulares

Como siguiente paso vamos a ocupar expresiones regulares para remover todos los caracteres especiales que tenemos en el texto de cada tweet.

## Análisis

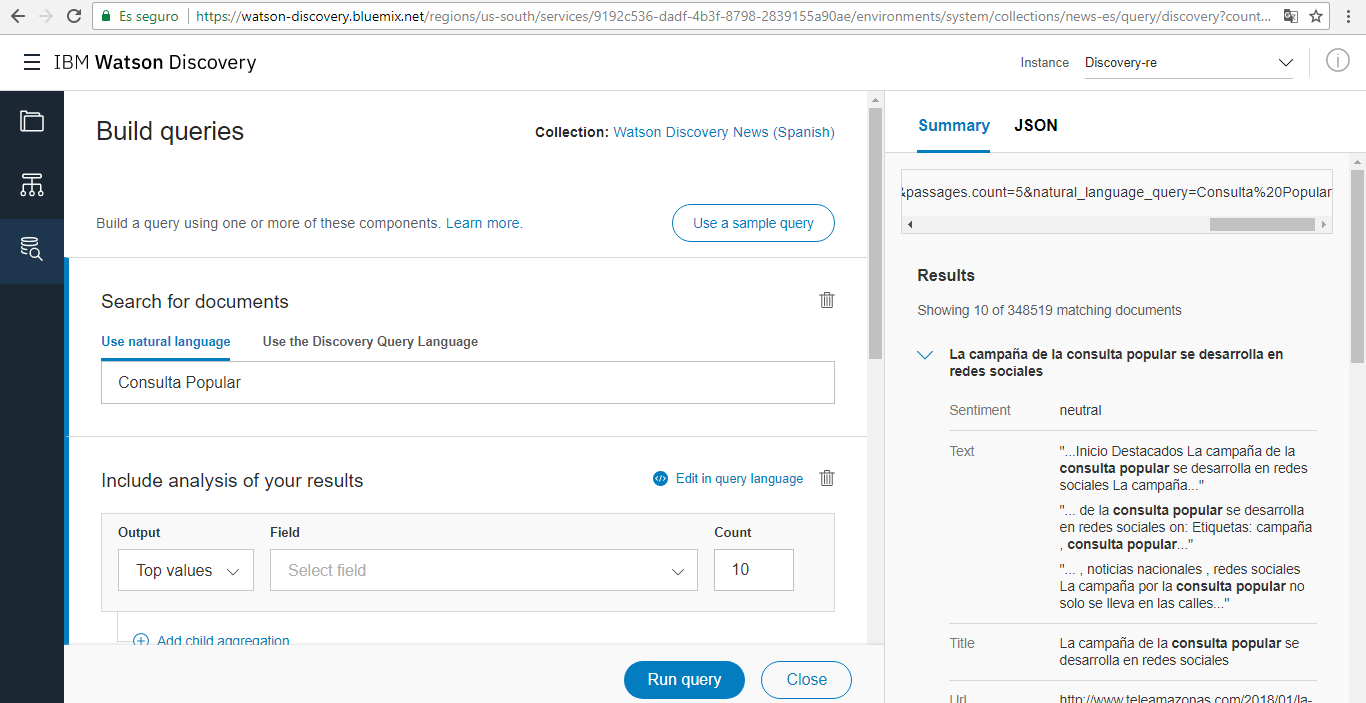
Discovery

Para hablar de Discovery primero debemos hacer una pequeña explicación de lo que es Watson.

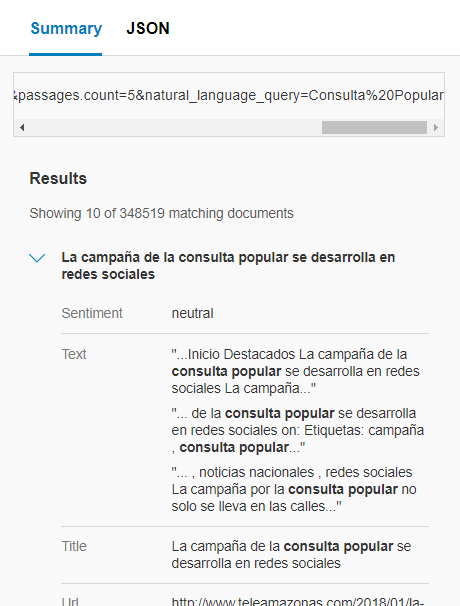
Watson principalmente fue creado para responder preguntas hechas en lenguaje natural, este proyecto que usa inteligencia artificial fue realizado por IBM. Este proyecto es extremadamente grande y ahora otorga herramientas para el apoyo del desarrollo de sistemas basados en inteligencia artificial, un resumen de lo que ofrece es:

* Conversation: Crea y despliega agentes virtuales en una variedad de canales, incluidos dispositivos móviles, plataformas de mensajería e incluso robots que son capaces de entablar una conversación.
* Language Translator: Watson usa la inteligencia artificial para traducir palabras y frases, pero intentando adaptarse a la jerga que se utilice en el ambiente al que se está desarrollando.
* Personality Insights: Su tarea es predecir las características, necesidades y valores de la personalidad a través del texto escrito. Comprenda los hábitos y preferencias de las personas a nivel individual y a gran escala.
* Visual Recognition: el reconocimiento visual etiqueta, clasifica y entrena de manera rápida y precisa el contenido visual mediante el aprendizaje de maquina o machine learning.
* Discovery: Es una herramienta que se encarga de buscar respuestas, monitorear tendencias y patrones que se obtienen de una base de datos obtenida de redes sociales u otras fuentes.

En la página web de Discovery podemos ingresar una palabra para buscar deseamos analizar y las palabras claves que facilitaran el análisis de los datos.



Se obtendrá como resultado todo lo referente a la consulta popular



Las utilidades de Discovery son muchas ya que nos ayuda a comprender con documentos mas a fondo sobre el tema que esperamos comprender. Y mas aun entrenar a la maquina con temas a fines a la consulta por ejemplo si se realizaría un estudio de la pregunta 6 de eliminar la ley de plusvalía Discovery podría ver la tendencia de los resultados y adelantarse para informar a la empresa donde trabaja podría adelantarse al favorable ambiente que se crearía y puedan desarrollar mas proyectos, así como el otro lado de la moneda si se prever que se van a comenzar mas proyectos y construcciones una empresa de cemento podría aumentar la producción ya que los compradores estarían dispuesto a comprar mas

## Presentación

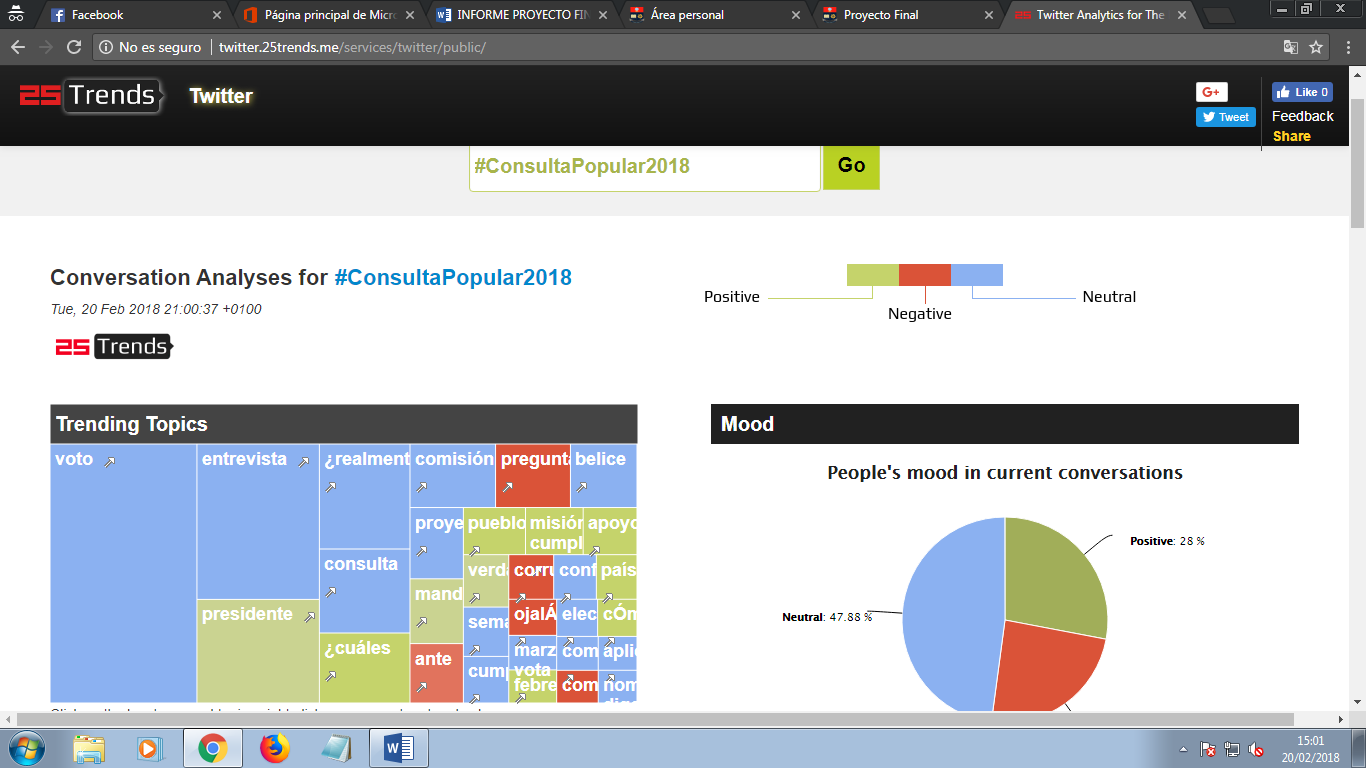
Existen varias herramientas que permiten el análisis de los datos, mediante gráficos y tablas:

25trends

Es un servicio de análisis de datos para redes sociales todo de forma automático. Esta herramienta otorga la oportunidad de analizar las siguientes redes sociales:

* Facebook: permite el análisis de las páginas de Facebook según una palabra clave.
* Instagram: analiza usuarios y etiquetas de Instagram mediante un @ o #.
* YouTube: puede analizar los comentarios de los videos y los clasifica según su alcance.
* Google: analiza páginas y etiquetas.

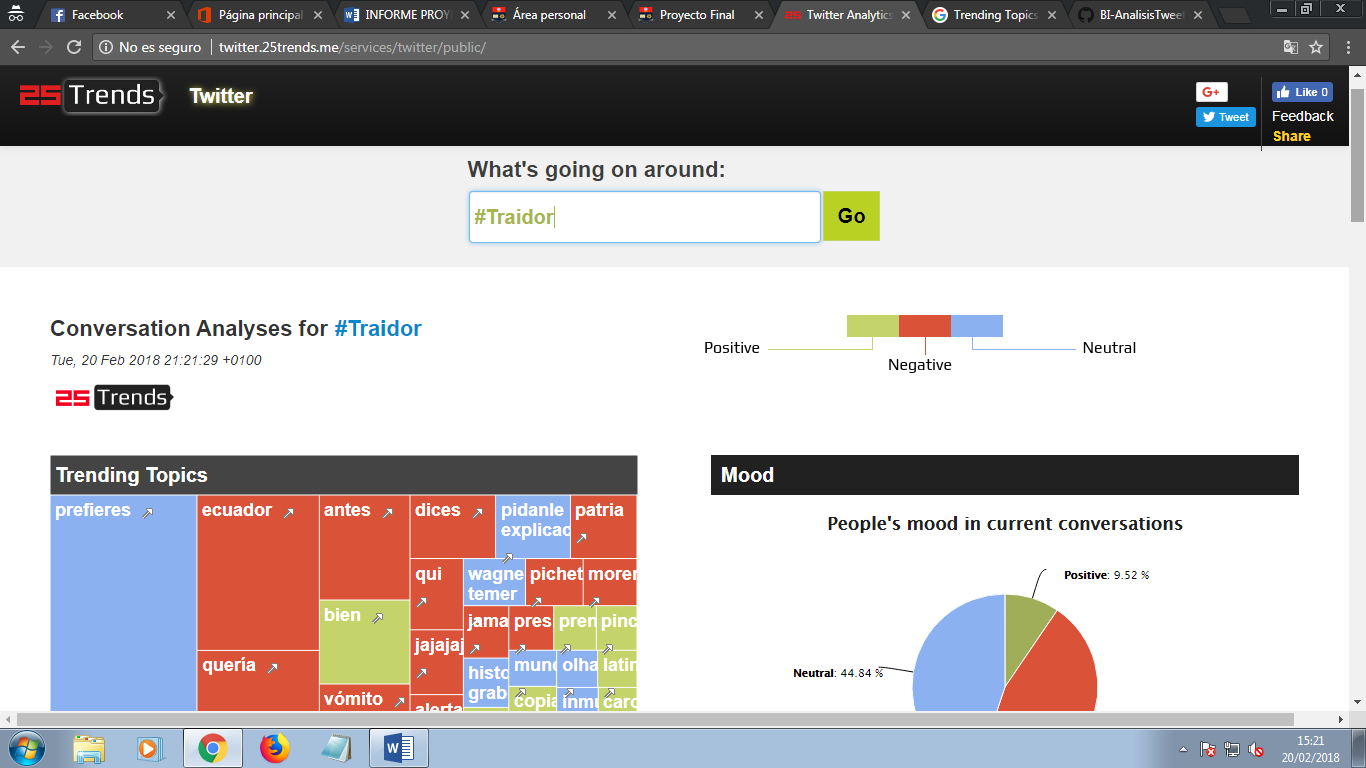
Presentado un hastag #ConsultaPopular2018



**Grafico ConsultaPopular2018**

El grafico representa una clasificación entre popular, tanto como neutral, y de negativo

En la página muestra Tendencia donde se ha usado el hastag #ConsultaPopular2018



La facilidad de encontrar Los hastag con más tendencia y también las combinaciones de ellos que se crean es una forma fácil de encontrar los tweets relacionados con la consulta como un aproximado del análisis de sentimientos, como también poder ver en que twets se utilizo y cuantas veces se retweet y algunos usuarios que lo usaron.

Twitter

En Twitter se puede analizar conversaciones en tiempo real con # o @. Esta herramienta también puede describir una tendencia histórica.

**Conclusiones Y Trabajo Futuro**

El analisis de datos en twitter no nos garantiza que los resultados obtenidos sean iguales a los reales debido a que en twitter existen varios falsos positivos, pues que hay mucha gente que publica en twitter varias veces, y tambien falsos negativos tomando en cuenta que existen muhas poersonas que nunca en su vida han utilizado esta red social.

Los tweets son un arma de doble filo las personas no suelen ser claras al momento de escribir y es dificil interpretar el sentimientos de ello scon disvovery le educo la maquina para que conosco lo que es consulta popular