

**ESCUELA POLITÉCNICA NACIONAL**

**FACULTAD DE INGENIERÍA DE SISTEMAS**

**INGENIERÍA EN SISTEMAS INFORMÁTICOS Y DE COMPUTACIÓN**

**PERÍODO ACADÉMICO**: 2017-A

**ASIGNATURA**: Inteligencia de Negocios

**PROFESOR**: Elisa Mena

**ESTUDIANTES:** Avalos Sebastián, Lema Santiago

**Informe de Proyecto Final**

# Introducción

Twitter es una red social que permite expresar opiniones de cualquier ámbito con el mundo, ya sea acerca de salud, moda, política, publicidad, etc. A estos mensajes rápidos se les llama tweets y se requiere realizar un estudio para analizar tendencias de opinión de los mismos. Para lograrlo se usará un analizador de sentimientos que clasifiqué los tweets publicados en Quito en positivos, negativos o neutros.

**Definiciones**

Algunas definiciones que serán necesarias tener en cuenta para el presente documento:

**Base de Datos CouchDB:** Es una base de datos NoSQL orientada a documentos que permite tener una arquitectura escalable con una gran cantidad de información que está representada en formato JSON y usa JavaScript como lenguaje de consulta. (CouchDB, 2017)

**Emoticones:** Es una forma de representar visualmente un objeto, símbolo o emoción.

**Lematizar:** Proceso en el cuál se obtiene el lema de una palabra, es decir su forma básica que representa a todas las formas posibles de esa palabra. (Lematización, 2017)

**Weka:** Software escrito en Java para Aprendizaje de Máquina

**Perceptrón Multicapa:** Red Neuronal usada para la clasificación

**Angular:** Es un framework basado en javascript que se usa para el desarrollo de aplicaciones web con arquitecturs que es principalmente mantenida por Google. (AngularJS, 2017)

# Método

**Adquisición de Datos**

Para este proceso se usó el script en Python entregado en el curso que consta de las siguientes partes importantes:

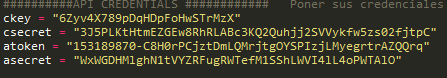


Figura 1. Credenciales Twitter

La figura 1 muestra las credenciales entregadas por Twitter al crearse una cuenta de desarrollador, estás son indispensables para poder empezar a cosechar los tweets.



Figura 2. Coordenadas para la cosecha

La figura 2 nos muestra el código donde empieza la cosecha, se puede observar que se envían como parámetro las coordenadas por las que se quiere filtrar los tweets en este caso se ingresó las correspondientes a la ciudad de Quito

**Preprocesamiento**

Una vez cosechados los tweets se identificó algunos fueron publicados en Colombia o Perú, por lo que se procedió a realizar una vista en la Base de Datos CouchDB que me permita filtrar solo los que sean de Ecuador.



Figura 3. Filtrado de Tweets

En la Figura 3 se puede observar el código usado para realizar este filtrado y adicionalmente se agregó otro filtro para obtener solo los tweets que se han publicado en español ya que para el análisis su usará un clasificador de sentimientos en este idioma

**Procesamiento**

Los tweets tienen contenido basura como: caracteres especiales, links, tags, etc. Estos nos impedirían realizar un análisis correcto del sentimiento del tweet por lo que se procedió con un barrido y limpiado de estos para dejar solo el texto importante que permitirá clasificar el tweet.

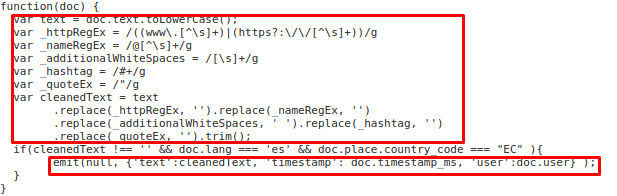


Figura 4. Limpiado del Tweet

Se usó expresiones regulares para realizar el barrido que se puede observar en la Figura 4, con estas se eliminan los links, caracteres especiales y espacios en blanco adicionales. Después el texto limpiado se envía junto con la fecha e información del usuario que será usada después para el análisis de los tweets

Las personas no solo expresan sus opiniones por medio de palabras, sino que también usan emoticones, por lo que estos no fueron removidos de los tweets y también serán usados para la clasificación.

**Análisis**

Pasos que se siguieron para el análisis de sentimientos y la clasificación de los tweets:

1. Usamos un diccionario de palabras que tienen información sobre el tipo (sustantivo, adjetivo o verbo), el nivel de agrado, entre otras.



Figura 5. Diccionario de Palabras

El tipo de palabra esta dado por el atributo “obj” y el nivel de agrado por el atributo “pleasantness”

1. Separar las palabras del tweet

https://fb-s-c-a.akamaihd.net/h-ak-fbx/v/t34.0-12/20864592_1555149461173459_896434336_n.png?oh=e1d1c03bedfb8e49551921719825e70f&oe=59958F60&__gda__=1502990703_f01040f20725076bb4a56521b0e6368e

Figura 6. Separación de palabras

En Python por defecto los strings están codificados en ASCII por lo que se vio la necesida de decodificarlos a utf-8 para no tener problemas con caracteres especiales usados en el español y emoticones. En la figura 7 se muestra el proceso del método que separa las palabras y las decodifica, este método es usado en la Figura 6 para obtener el tweet tokenizado.

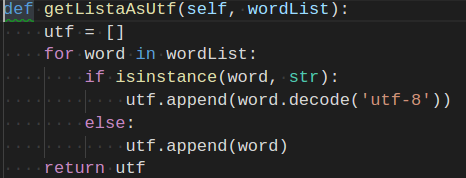


Figura 7. Decodificación a utf-8

1. Lematizar las palabras separadas usando la librería Pattern

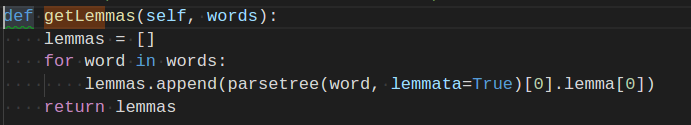


Figura 8. Lematización de palabras

1. Limpiar palabras vacías

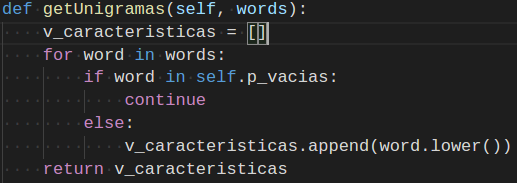


Figura 9. Limpiado de palabras vacías

Las palabras como: a, para, este, quien, hasta, fueron, cuanto, etc. no aportan para el análisis de sentimientos de los tweets, por lo tanto, son removidas.

1. Del conjunto de palabras de un tweet obtener la frecuencia con la que aparecen adjetivos positivos, adjetivos negativos, verbos positivos, verbos negativos, sustantivos positivos y sustantivos negativos. Este será el vector de características y se usará para construir un archivo. arff que permitirá realizar la clasificación por medio de librerías de Weka.

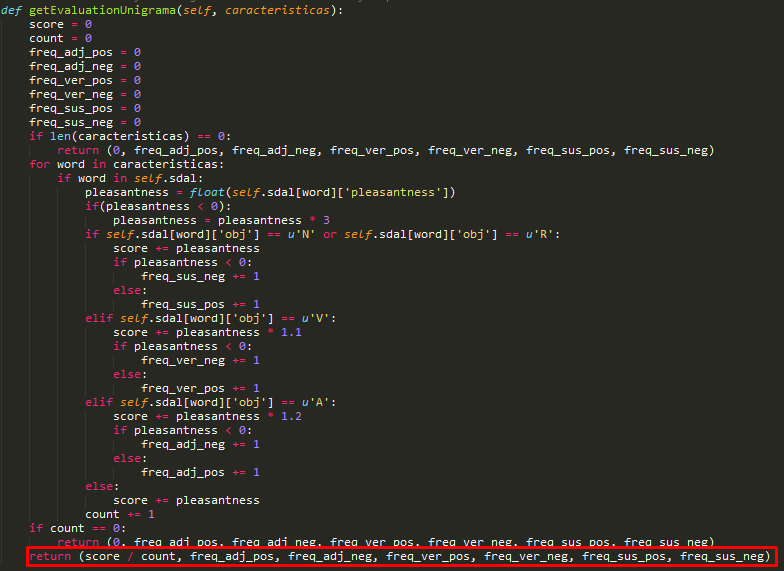


Figura 10. Vector de características

Para sacar la frecuencia simplemente se cuenta la aparición de cada tipo de palabra en el conjunto de palabras provenientes del tweet. En la Figura 10 se puede observar este proceso y se retorna una tupla con el vector de características.

1. Con el archivo. arff se procede a crear el modelo que permita clasificar el sentimiento de los tweets.

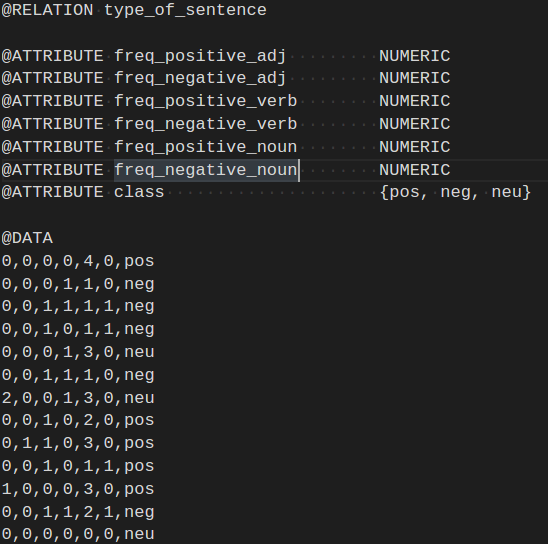


Figura 11. Archivo .arff

La Figura 11 muestra como está estructurado el set de entrenamiento.

1. Usando las librerías de Weka, se crea el modelo de clasificación de sentimientos

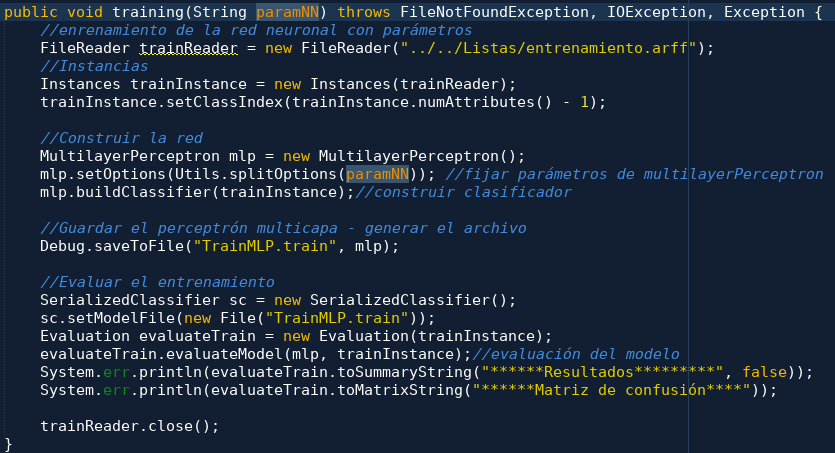


Figura 12. Creación Modelo de Clasificación

En esta parte se lee el archivo .arff de entrenamiento y se usa el clasificador Multiplayer Perceptron para construir el modelo.

1. Evaluación del entrenamiento

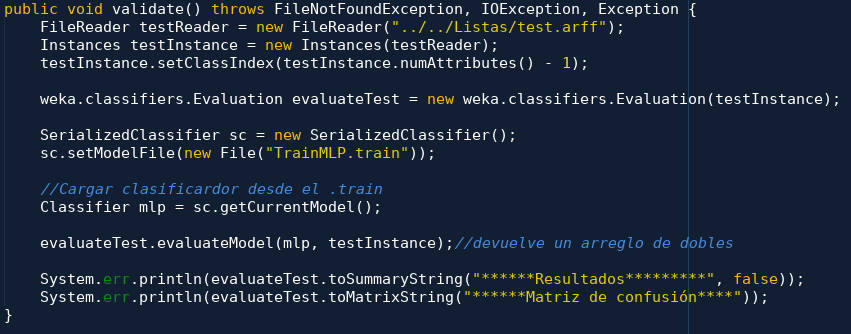


Figura 13. Evaluación

Se realiza una evaluación para determinar la precisión del clasificador y los resultados se pueden observar en la Figura 14

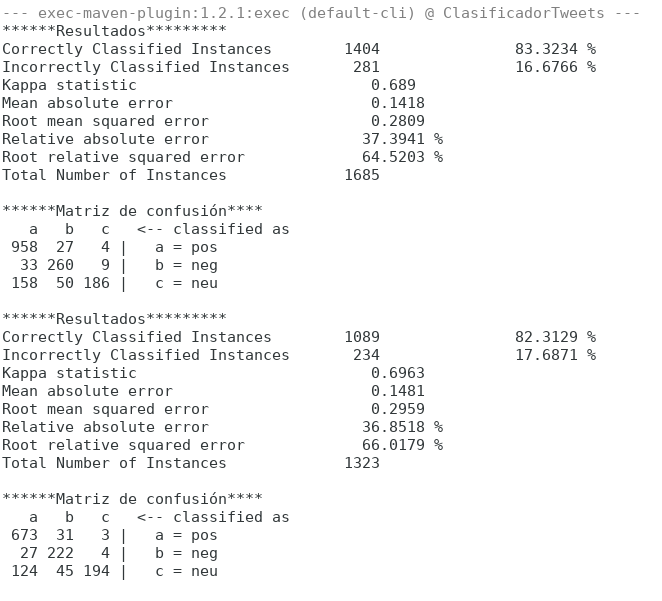


Figura 14. Resultados de la Evaluación

Con este modelo se empieza a clasificar los tweets.

**Presentación**

Para la presentación de la información se siguieron los siguientes pasos:

1. Crear servicios web



Figura 15. Creación de servicios web

Se crearon servicios web que entreguen información sobre los tweets, esto con la finalidad de consumir estos servicios y poder usarlos en nuestro frontend

1. Para el frontend se usó Angular y la librería HiCharts para los gráficos.

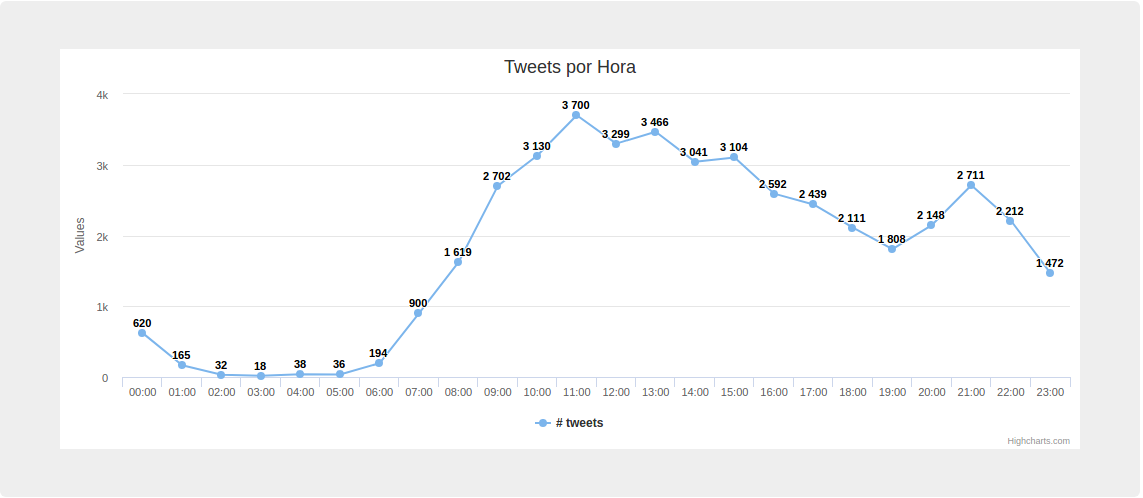


Figura 16. Gráfico de Tweets por Hora

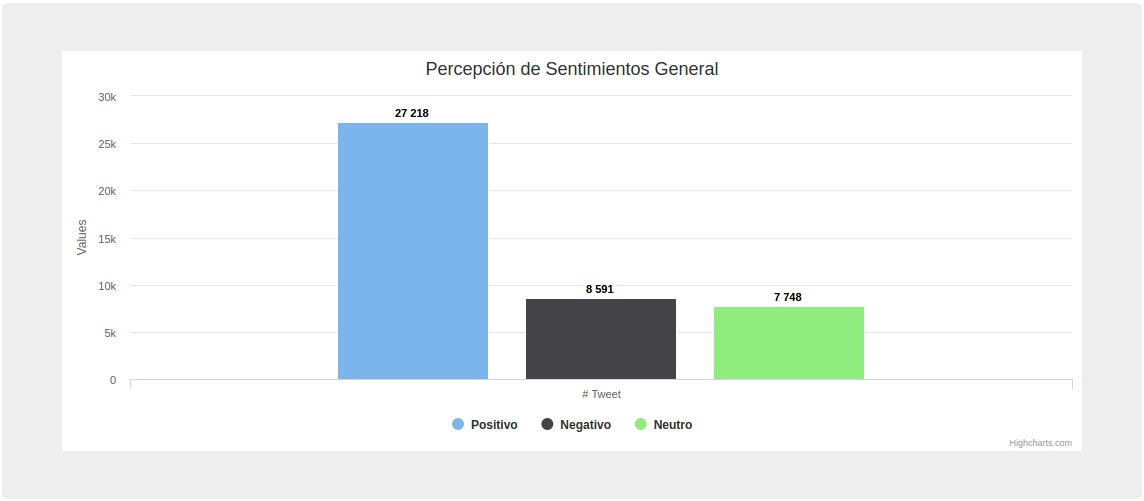


Figura 13. Gráfico de Percepción de Sentimientos General



Figura 14. Análisis de sentimientos en vivo

Se creó una página que muestre el análisis de sentimientos de los tweets en vivo la estructura de esta se muestra en la Figura 14 y la lista de tweets que se muestra sigue aumentando mientras se siguen cosechando y los va clasificando con el modelo creado.

# Resultado

En la Figura 16 se puede observar que la menor cantidad de tweets publicados se encuentran entre las 00:00 y 06:00 horas, a partir de esta hora empiezan a subir el número de tweets hasta llegar al pico más alto a las 11:00 horas con un total de 3700 tweets, con esto se puede decir que la mayor cantidad de tweets publicados se encontraron entre las 11:00 y 13:00 horas, después se nota una caída de la cantidad de tweets hasta las 19:00 horas donde llegan a los 1808 tweets y por último se tiene una última subida a las 23:00 horas con un total de 2711 tweets para de nuevo ir en caída.

Los resultados del clasificador de sentimientos se pueden observar en la Figura 13 donde se indica que 27218 tweets fueron clasificados como positivos, seguido de 8591 tweets negativos y 7748 tweets neutros.

# Conclusiones y trabajo futuro

* Se usó recursos facilitados en el curso para cosechar tweets y almacenarlos en una base de datos noSQL CouchDB.
* Se investigó y se creó un clasificador usando la red neuronal perceptrón multicapa, que analicé los sentimientos de un tweet y los clasifiqué como positivos, negativos o neutros para identificar tendencias de opinión en Quito
* Se identificó que las personas tienden a tweetear y expresarse más a medio día, teniendo la menor afluencia en horas de la madrugada
* Usamos librerías de Weka para entrenar nuestro clasificador con la red neuronal “Perceptrón Multicapa” y como vector de características se eligieron la frecuencia de adjetivos positivos, adjetivos negativos, verbos positivos, verbos negativos, sustantivos positivos y sustantivos negativos que aparecen en cada tweet.

Trabajo a futuro:

* En un futuro se espera poder usar esta información para analizar los tweets negativos y encontrar patrones que permitan poder anticiparse ante un posible suicidio, ya que personas que mencionaron el suicidio o mensajes con una alta negatividad son más propensas a pensar o llevar a cabo un acto de suicidio.
* Se puede usar este clasificador para sacar estadísticas de posiciones de las personas frente a movimientos filtrando los tweets que tengan que ver con política.

**Repositorio**

El código de la aplicación esta subido en un repositorio de Github y se lo puede encontrar con el siguiente link: [https://github.com/sebas1208/sentiment-analizer-bi.git](https://l.messenger.com/l.php?u=https%3A%2F%2Fgithub.com%2Fsebas1208%2Fsentiment-analizer-bi.git&h=ATNZFjgQiJPy1RQrfiK9XY7ZtCLm0nL5A2uehI5HBs2JBxigvrSUiXsAhWwRd0ed9SOpn-c5CFqEUGdeav75-ouzak2tyl1vuaEACKvuGD7Tdc7F8HGaTk_bHor_N2Yh7HubCwJF565VPA)

# Bibliografía

*AngularJS*. (16 de Agosto de 2017). Obtenido de Wikipedia: https://en.wikipedia.org/wiki/AngularJS

*CouchDB*. (16 de Agosto de 2017). Obtenido de Wikipedia: https://en.wikipedia.org/wiki/CouchDB

*Lematización*. (16 de 08 de 2017). Obtenido de Wikipedia: https://es.wikipedia.org/wiki/Lematizaci%C3%B3n