

ESCUELA POLITECNICA NACIONAL

FACULTAD DE INGENIERIA DE SISTEMAS

*INGENIERIA EN SISTEMAS INFORMATICOS Y DE COMPUTACION*

**PROYECTO FINAL**

**FECHA: 23/02/2016**

**NOMBRES: Jorge Miño, Marlon Vernaza, Luis Villalta**

**MATERIA: Inteligencia de Negocios**

**PROFESOR: Ing. Elisa Mena**

**Proyecto Final**

**Objetivos:**

Implementar e investigar el funcionamiento de un clasificador de sentimientos utilizando los algoritmos de aprendizaje vistos en clase y los datos recolectados de Twitter desde octubre del 2016 para identificar tendencias de opinión sobre los candidatos presidenciales.

**Objetivos Específicos:**

* Crear un clasificador con alta precisión “accuracy” utilizando datos extraídos de Twitter para minar opinión pública acerca de candidatos presidenciales.
* Identificar y seleccionar las herramientas necesarias para procesar y analizar datos en tiempo real provenientes de Twitter

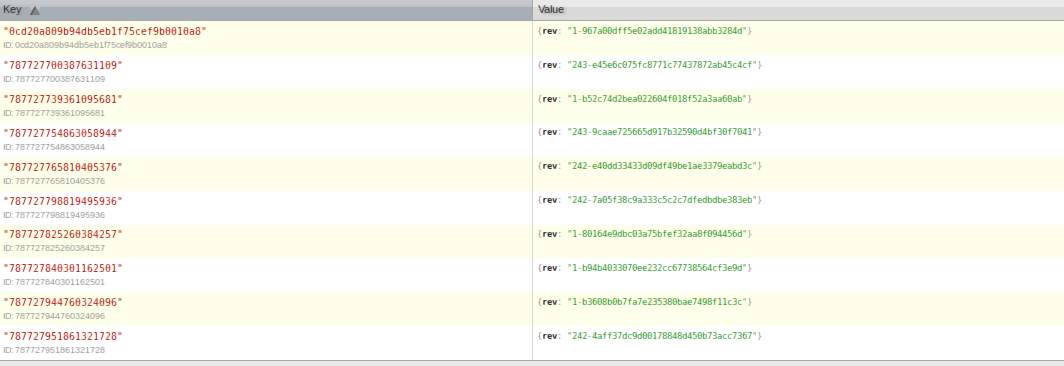
Introducción

Debido a la aparición de nuevas herramientas hoy en día se puede realizar un análisis de las tendencias políticas de maneras diferentes a las cuales se realizaba en el pasado, en el presente proyecto se pretende realizar el análisis de estas tendencias con la utilización de herramientas tales como CouchDB en donde se almacenara datos recolectados de la red social twitter, estos datos posteriormente se indexaran a la herramienta Elasticsearch la cual se la utilizara para realizar búsquedas y minar la opinión pública, esto se hace con el objetivo de conocer cuál es la opinión publica de los candidatos presidenciales, posteriormente a este paso se realizara gráficas para mostrar los resultados obtenidos para este paso se utilizara la herramienta Kibana.

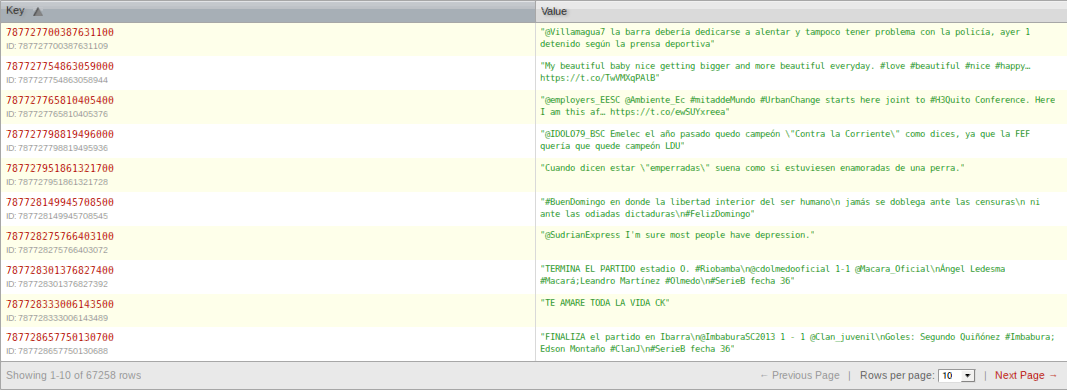
**Método**

1. Primero tuvimos que realizar la adquisición de los datos para lo cual utilizamos un cosechador de tweets donde se establecieron algunas coordenadas necesarias para capturar estos tweets que son correspodientes a la ciudad de Quito. Para esto se utilizó un script codificado en Python y una base de datos de CouchDB.
2. Ahora necesitamos que estos tweets se filtren para solo salgan los correspondientes a la cuidad de Quito ya que en la recolección de tweets se obtiene información que proviene otro otros países como Colombia o Perú; entonces para eso es necesario utilizar vistas en la base de datos de CouchDB.
3. Cuando ya se tengan solo los tweets de Quito, es necesario procesar el cambio de texto de cada tweet ya que con ese procesamiento se podrá analizar la opinión pública. Pero este procesamiento de texto va involucrar la remoción de algunos caracteres especiales, links, tags, etc.
4. A continuación se tendrá que realizar un análisis sobre el texto de cada uno de los tweets, para poder minar opinión pública, para esto utilizaremos la herramienta Elasticsearch para permitir realizar búsquedas sobre el texto de los tweets.
5. Por último se debe analizar cada uno de los tweets para cada candidato o partido al que pertenezca esto permitirá identificar cual es la tendencia de la opinión pública. Para saber cuál es la tendencia se deberá hacer una presentación de gráficos, tablas para poder comprender fácilmente los resultados que se van a presentar para esto se necesitara la herramienta Kibana.

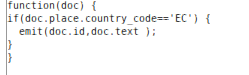
**Resultados**



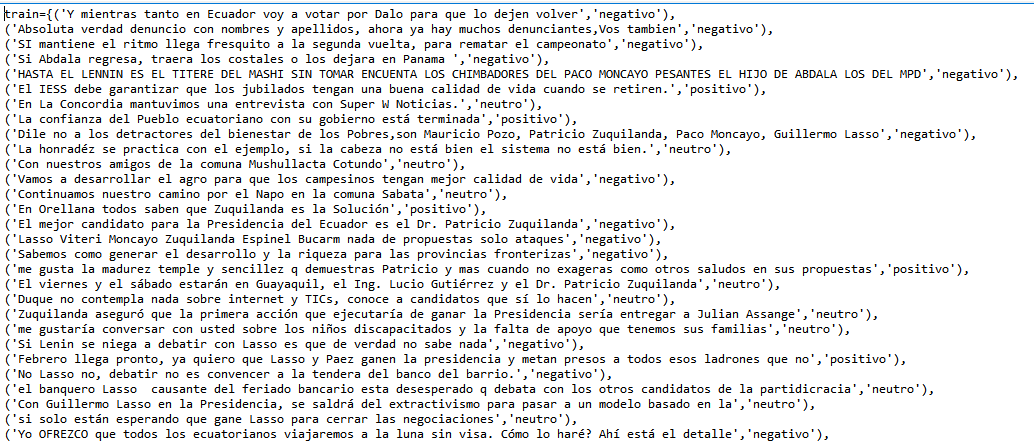
En la siguiente imagen se puede ver los tweets totales



En la siguiente imagen se puede ver los tweets filtrados



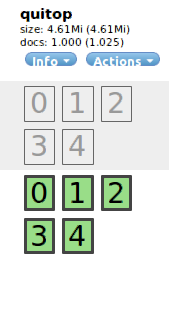
En la siguiente imagen se puede ver el código filtrado



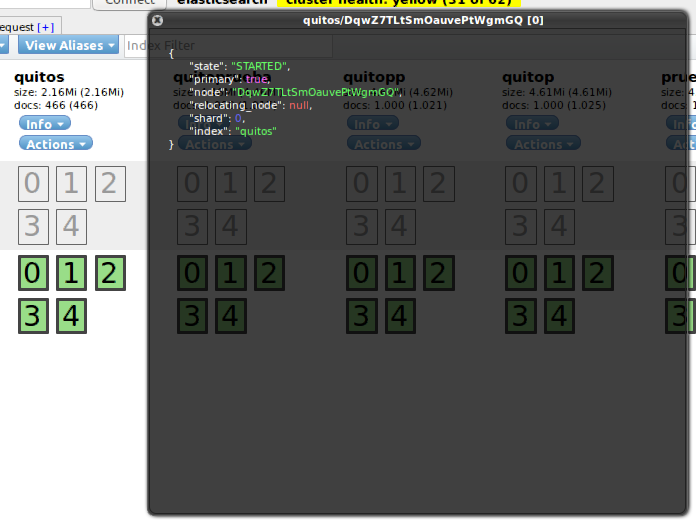
En la siguiente imagen se muestra el training



En la siguiente imagen se vizualiza el codigo para subir los tweets al Elasticsearch



En la siguiente imagen la vista del Elasticsearch



En la siguiente imagen se muestra la indexación de la base de datos de Couchdb en el Elasticsearch.

**Discusión**

La precisión mediante un sistema de análisis de sentimiento es, en un principio, cuánto está coincidiendo con juicios humanos en este caso con la opinión pública.

El aumento de los medios sociales tales como blogs y redes sociales ha ido alimentando el interés en el análisis de sentimientos, por ejemplo con la propagación de críticas, calificaciones, recomendaciones y algunas otras formas diferentes de expresión, entonces la opinión pública se ha convertido en una especie de divisa para negocios que buscan comercializar sus productos, identificar nuevas oportunidades y manejar sus reputaciones.

También podemos decir que mediante el análisis del sentimiento, se aspira entender, en primer lugar, con que guarda relación el texto que se analiza. Por ejemplo, interesa saber si la frase dada trata sobre el concepto que interesa analizar. Una vez que el sistema tiene claro sobre que es una opinión, el siguiente objetivo será conocer el sentido de esa opinión (positivo negativo o neutro, básicamente).

Los beneficios del análisis de sentimiento son numerosos y muy importantes. Por ejemplo la destreza de extraer información de datos de las redes sociales es una práctica que ya están adoptando empresas a nivel mundial.

**Conclusiones y trabajo futuro**

* Concluyo que con este proyecto final hemos abordado la compleja tarea de explicar el contexto del análisis de sentimientos y algunos conceptos relacionados con esta área.
* Concluyo también el análisis de sentimiento es una herramienta muy poderosa, con la cual se puede obtener una gran cantidad de información apreciable para la empresa u organización.
* Además puedo concluir que como los datos provienen de fuentes como Twitter, Facebook o blogs, los cuales progresan en volúmenes inmanejables la minería de datos juega un papel relevante en cualquier proceso, ya que permite el almacenamiento y procesamiento de grandes cantidades de datos.
* Para trabajos futuros es difícil especular sobre como un sistema relativamente aprendiz evolucionara en un futuro, pero existe una afirmación muy general en la cual se dice que el análisis de sentimientos necesita ir más allá de una escala unidimensional positiva-negativa.
* Entonces podemos decir que en un futuro no muy lejano, para capturar y razonar de manera más efectiva el amplio rango de emociones que los humanos expresan por escrito, se necesitara una escala multidimensional más sofisticada