[[1]](#footnote-1)

Diseño e Implementación de un modelo de clasificación de sentimientos utilizando machine learning

Jaguaco Sonia, Miño Jorge

*Abstract*—En el presnte informe se presenta la implementación de un modelo de clasificación de sentimientos utilizando machine learning. Para el análisis de sentimiento se utilizo el procesamiento de lenguaje natural, análisis de texto y lingüística computacional. Se intento crear un clasificador de sentimiento en español utilizando datos extraídos de Twitter para minar opinión pública en la ciudad de Quito. Ademas, se identifica y selecciona las herramientas necesarias para procesar y analizar datos en tiempo real provenientes de Twitter. Los resultados obtenidos se muestran en forma de graficos e imágenes que representan la tendencia de los tweets ya sea, neutral, positiva o negativa.

*Index Terms*—Analisis de Sentimientos, Machine Learning, Clasificador de Sentimientos, Tweetes, Predecir.

# INTRODUCCION

L

as redes sociales han revolucionado la forma en que las personas se comunican. La información disponible de las redes sociales es beneficiosa para el análisis de la opinión del usuario. Asi pues, tenemos la medición de la información sobre un producto nuevo en el mercado, se puede observar la respuesta al cambio en la política o el disfrute de un evento en curso [1]. Analizar manualmente los datos de forma que se pueda tomar decisiones es tedioso y potencialmente costoso.

El análisis de sentimientos es un área que se ocupa de la extracción de opinión del usuario de forma automática. Un ejemplo de un sentimiento positivo es el Procesamiento del lenguaje natural, ya que es normalmente usado de forma automática. Se puede considerar como sentimiento positivo a la siguiente frase, “El Procesamiento del lenguaje natural es muy divertido” alternativamente, un sentimiento negativo es “Es un día horrible, no voy afuera”. En el presente proyecto se evalua como textos objetivos ha la palabra “divertido” alternativamente, un sentimiento negativo es “Es un día horrible, no voy afuera” [1]. Los datos que no expresan ningún sentimiento se considera ha los titulares de noticias, por ejemplo, “Estantes compañía viento planes del sector”.

Hay muchas maneras en que los datos de redes sociales se pueden aprovechar para tener una mejor comprensión de la opinión del usuario tales problemas están en el centro de procesamiento de lenguaje natural (NLP) y la investigación de minería de datos.

En este artículo presentamos el análisis de los sentimientos, con herramientas calificadas, que son capaces de analizar datos de Twitter. Mostramos cómo recolectar automáticamente en una base de datos, cochdb, los datos de tweeter, para el análisis de emociones y de minería de opiniones. Usando el phyton construimos un clasificador de sentimientos, que es capaz de determinar los sentimientos positivos, negativos y objetivas para los ciudadanos de quito.

.”

# Marco Teórico

## CouchDB

## Es una base de datos que abarca completamente la web y guarda sus datos con documentos JSON. Dentro de las cararacteristicas más importantes tenemos las siguientes [1].

## Posee acceso a los documentos con su navegador web, a través de HTTP.

## Consulta, combina y transforma tus documentos con

JavaScript.

* CouchDB funciona bien con aplicaciones web y móviles modernas.
* Puede distribuir sus datos de forma eficiente utilizando la duplicación incremental de CouchDB.
* Soporta configuraciones master-master con detección automática de conflictos.
* CouchDB viene con un conjunto de características, como la transformación de documentos en tiempo real y cambios en tiempo real.
* Posee una consola de administración web, fácil de usar, y directamente.
* CouchDB es altamente disponible y la partición tolerante, pero también es eventualmente consistente y tiene una tolerancia a fallos

## Json

JSON es un formato de texto que facilita el intercambio de datos estructurados entre todos los lenguajes de programación.

Es la sintaxis de llaves, corchetes, comas y comas que es útil en muchos contextos, perfiles y aplicaciones.

JSON fue inspirado por los literales de objeto de JavaScript alias ECMA. Asi pues, no intenta imponer los datos internos de ECMAScript [2].

Se caracteriza por compartir un pequeño subconjunto de las variables textuales de ECMAScript. Ademas posee representaciones con todos los demás lenguajes de programación.

## Phyton

Python es un poderoso lenguaje de programación de computadoras moderno. Tiene algunas similitudes con

Fortran, uno de los primeros lenguajes de programación, pero es mucho más poderoso que Fortran.

Se basa en la indentación como estructura de control. No se esta obligado a definir clases en Python

Python fue desarrollado por Guido van Rossum, y es software libre [3].

## Machine Learning

Una maquina de aprendizaje es cuando un programa de computadora "aprende" de una experiencia A, con respecto a alguna clase de tareas B y con una medida de rendimiento C, si su desempeño en tareas B, medido por C, mejora con la experiencia A. Es decir, se considera que una maquina aprende de experiencias y cada vez, va perfeccionando su rendimeiento en base a sus experiencias obtenidas [4].

## ElasticSearch

Elasticsearch es una búsqueda y análisis en tiempo real, distribuidos y de código abierto. Elasticsearch es de código abierto, desarrollado en Java y utilizado por muchas organizaciones grandes alrededor del mundo. Está licenciado

bajo la licencia de la versión 2.0 de Apache [5].

*D. Tweeter*

Es una línea de noticias y redes sociales de servicio donde los usuarios publican e interactúan con los mensajes, "tweets", restringida a 140 caracteres. Los usuarios registrados pueden publicar tweets, pero los que no están registrados pueden leerlos [7].

# Metodologia

La metodologia utilizada en este proyecto es experimental, la cual permite poner en práctica conocimientos y técnicas adquiridas en clase y en investigacion. Asi el desarrollo del proyecto se describe a continuación.

## Recoleccion de Tweets

Para la recolección de tweets se utilozo un script desarrollado en phyton y posteriormente se almaceno en una base de datos CouchDB.

Tenemos que crearnos una cuenta como desarrollador para poder utilizar el recolector de tweets.

## Filtrado de Datos

Se filtran los datos con el propósito de obtener los datosunicamente de la ciudad de quito, ya que, en la recolección de tweetes, se observo que se tenia información de Colombia, entre otras. Para esto se creo vistas en la base de datos CouchDB.

## Procesamiento de los Tweets

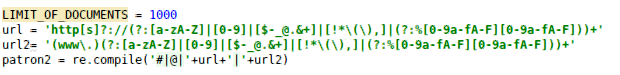
Una vez que se obtiene únicamente los tweets de Quito, se procede al procesamiento de la información recolectada, es decir, se limpia la data de los emoticones, likes, etc. Esto se hace para cada tweet, con el fin de analizar la opinión pública.

## Analisis de los Tweets

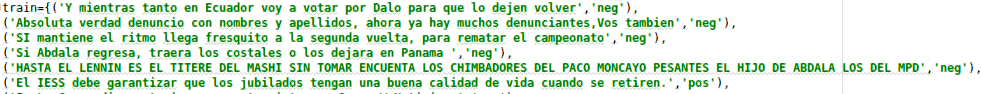
Para el analisis de los Tweets filtrados y procesados, se pasa ha analizar cada texto, de los respectivos tweets. Esto se hace con el fin de minar la opinión pública que es lo que se desea obtener. En este punto se ha utilizado Elasticsearch ya que nos permitio, realizar búsquedas sobre el texto de cada tweet.

Se analiza cada uno de los tweets para determinar el estado anímico de la opinión pública. Esto permitirá determinar si el estado anímico de las personas es positivo, nuetral o negativo.

Utilizamos expresiones regulares para filtrar el contenido



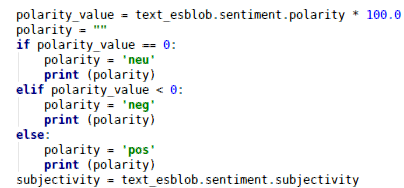
Creamos un train para poder implementar los sentimientos



Posterior a esto limpiamos cada tweet con el siguiente metodo esto nos permite eliminar todas las expresiones como paginas web.



Posterior a esto procedemos a ponerle sentimiento a cada tweet utilizando



Posterior a esto guardamos en la vista el tweet limpio y el sentimiento.



Para visualizar los datos filtrados según sentimiento lo realizamos en vistas

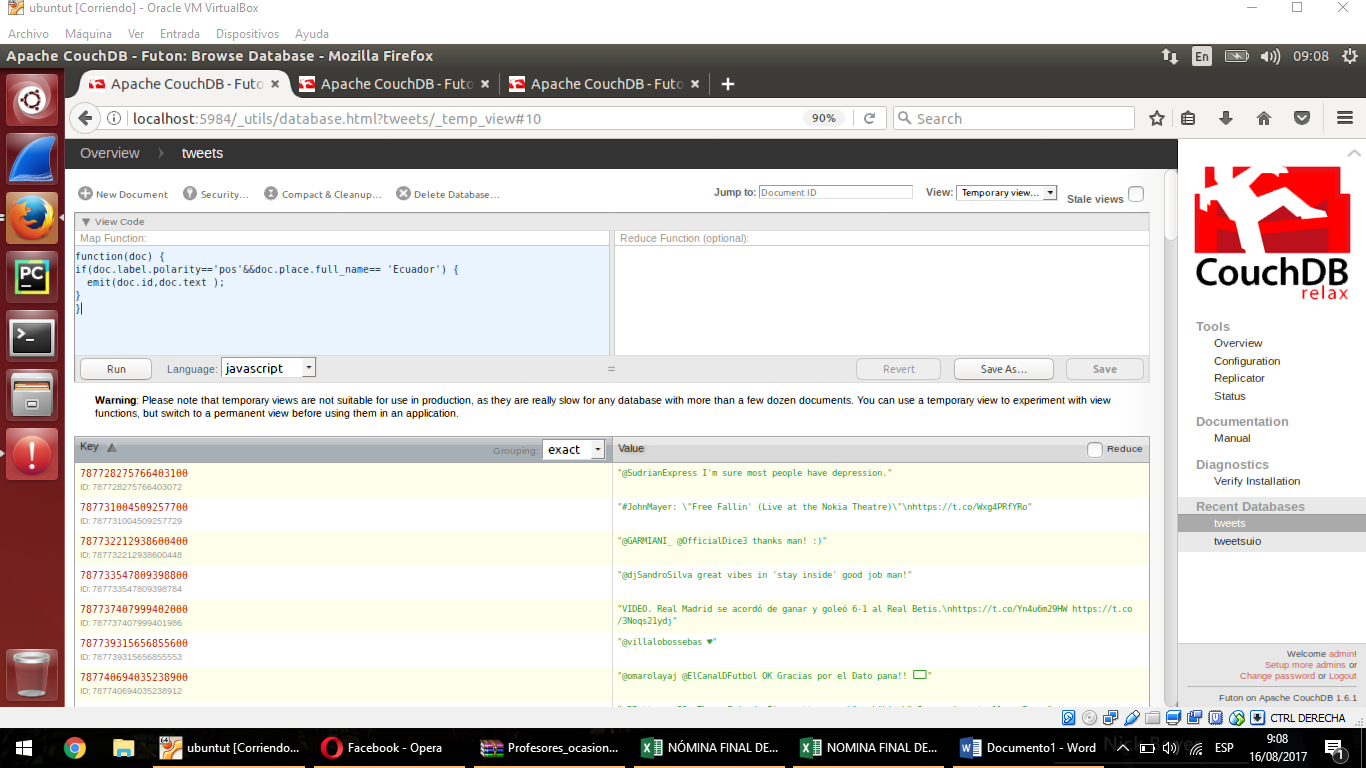
Datos filtrados por Ecuador y por sentimientos pos

function(doc) {

if(doc.label.polarity=='pos'&&doc.place.full\_name== 'Ecuador') {

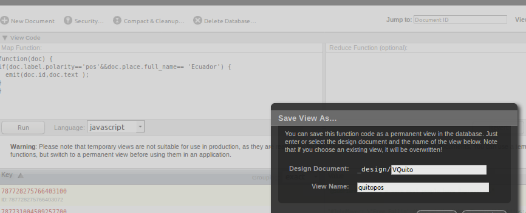
emit(doc.id,doc.text );

}}



*En la Fot1. Se muestra la función creada para filtrar los tweets.*

* Creacion de la vista creada para sentimeintos positivos



*En la Fot2. Se muestra la vista para filtrar los sentimientos positivos.*

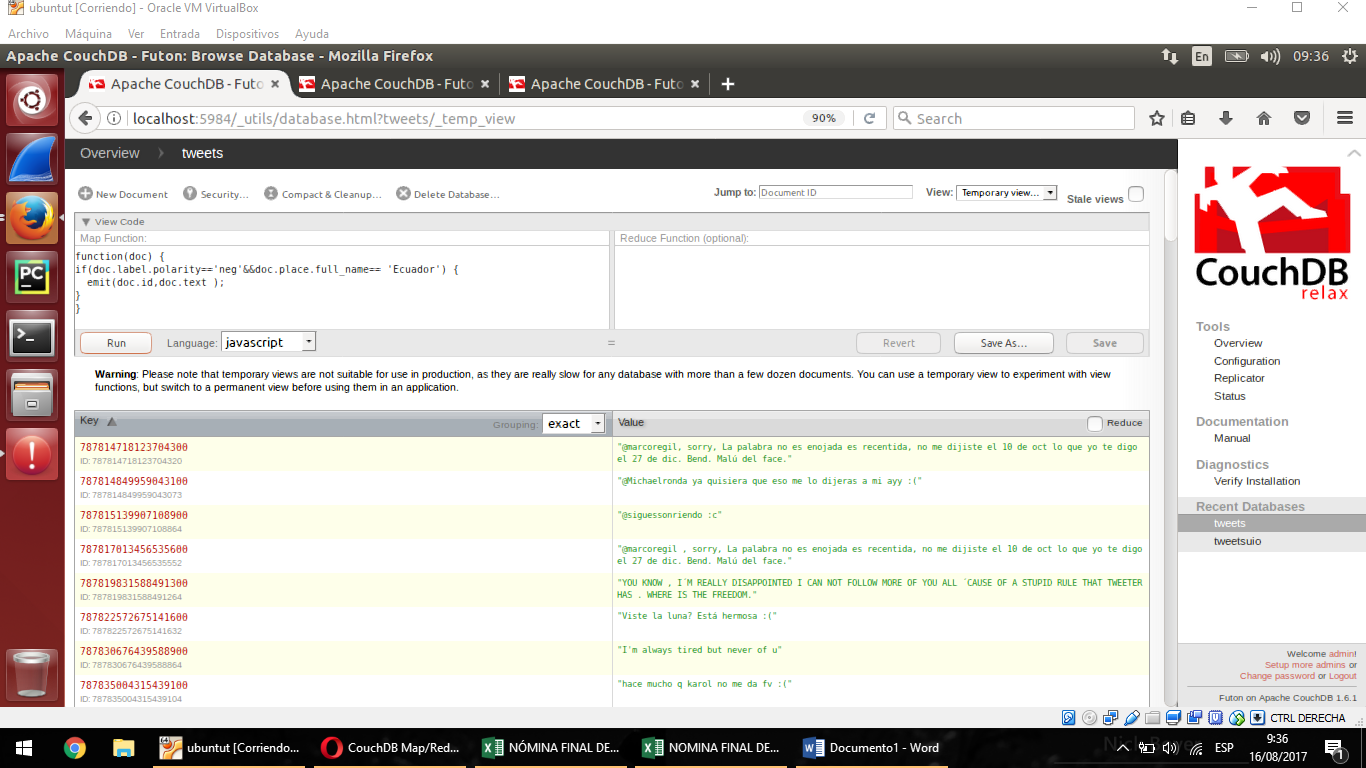
* Vista para sentimientos negativos, la función q se utiliza es la siguiente.

function(doc) {

if(doc.label.polarity=='neg'&&doc.place.full\_name== 'Ecuador') {

emit(doc.id,doc.text );

}}



*En la Fot3. Se muestra la función para los sentimientos negativos.*

* Vista para Sentimientos Neutrales

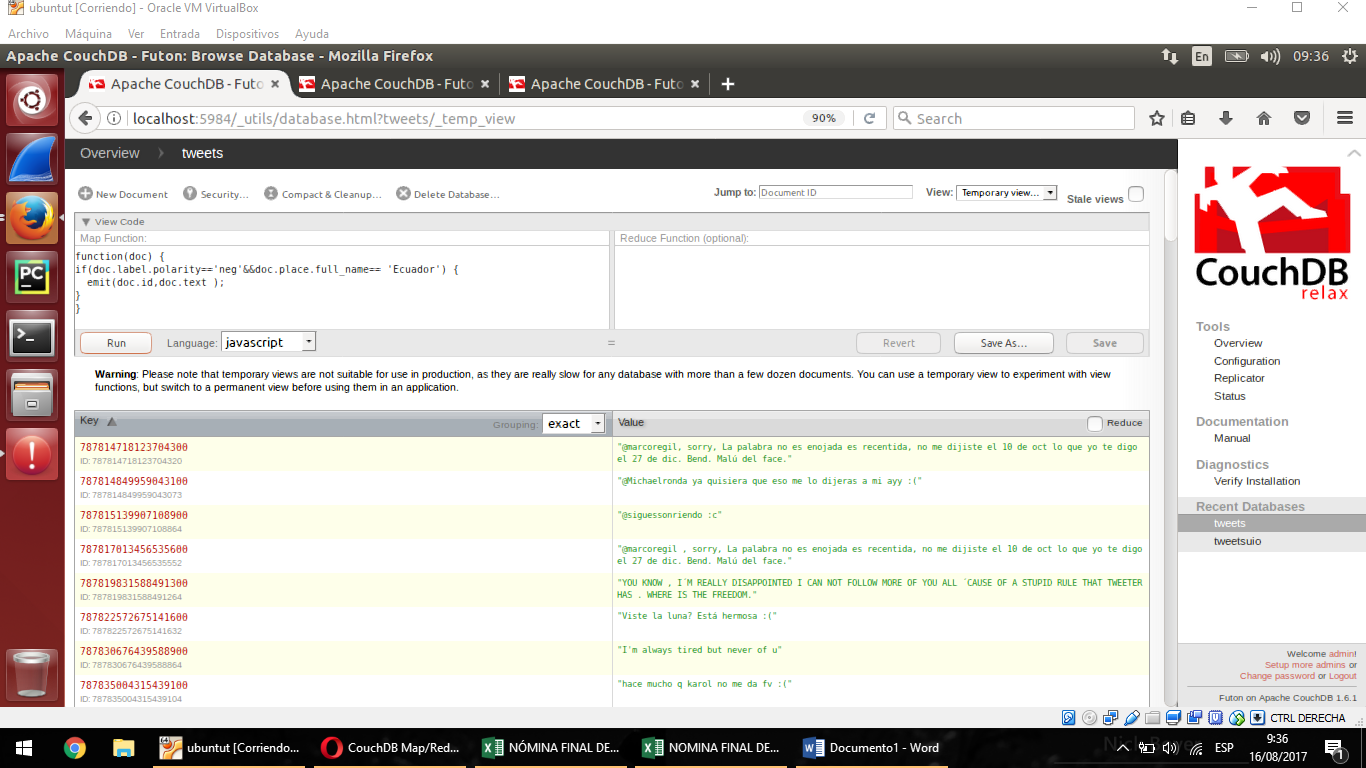
Ahora generamos la vista de los tweets filtrados neutralmente

function(doc) {

if(doc.label.polarity=='neu'&&doc.place.country\_code=='EC') {

emit(doc.id,doc.text );

}}



*En la Fot4. Se muestra la función para los sentimientos neutrales.*

Para el análisis de datos se utilizo los siguientes

|  |  |
| --- | --- |
| Datos |  |
| Sentimiento | # Tweets |
| neu | 12361 |
| pos | 596 |
| neg | 241 |

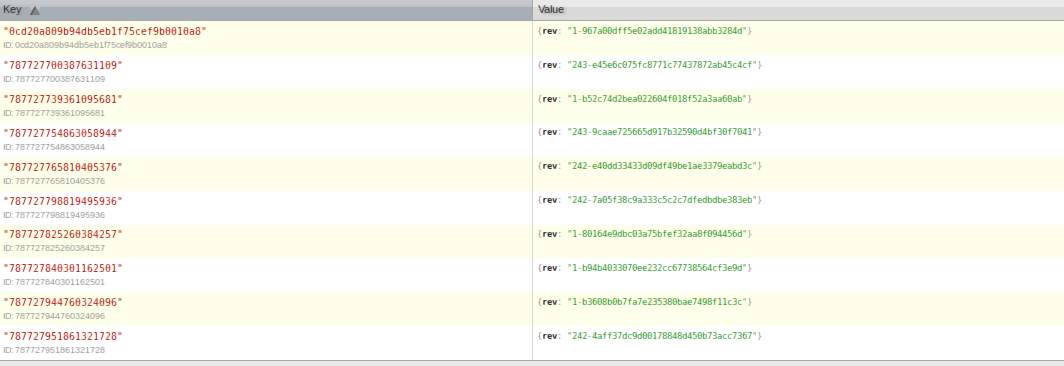
## Presentacion

Para saber cuál es la tendencia pública ya sea, neutral, positiva o negativa. Se realizo una visualización en graficos representativos, tablas que facilitaran entender fácilmente el resultado obtenido. Par esto se utiliza la herramienta.

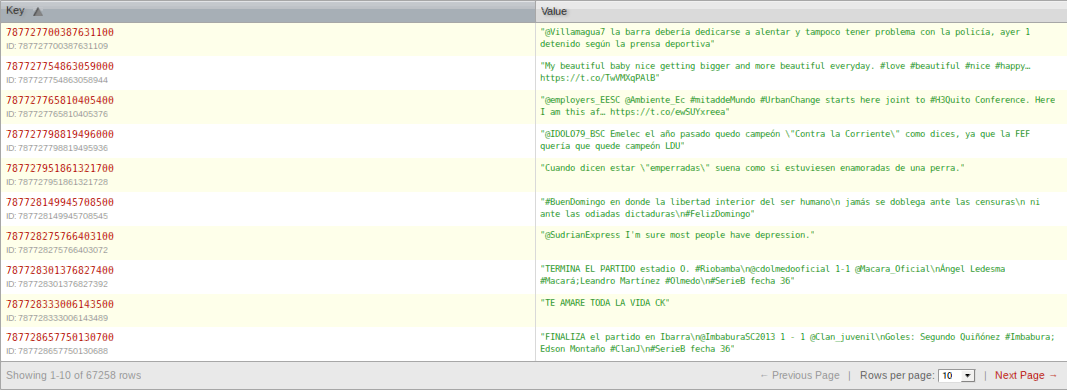
# Analisis de Resultados

## Los resultados obtenidos muestran una gran cantidad de datos recolectados, esto permite evidenciar que la canticad de tweets publicados diariamente es alto, esto sirve de mejor manera para poder tener más información que calsificar. Sin embargo, una gran cantidad de datos no garantiza una clasificación óptima ya que se debe tomar en cuenta la coherencia y concisión de los datos.

## Los Datos Recolectados

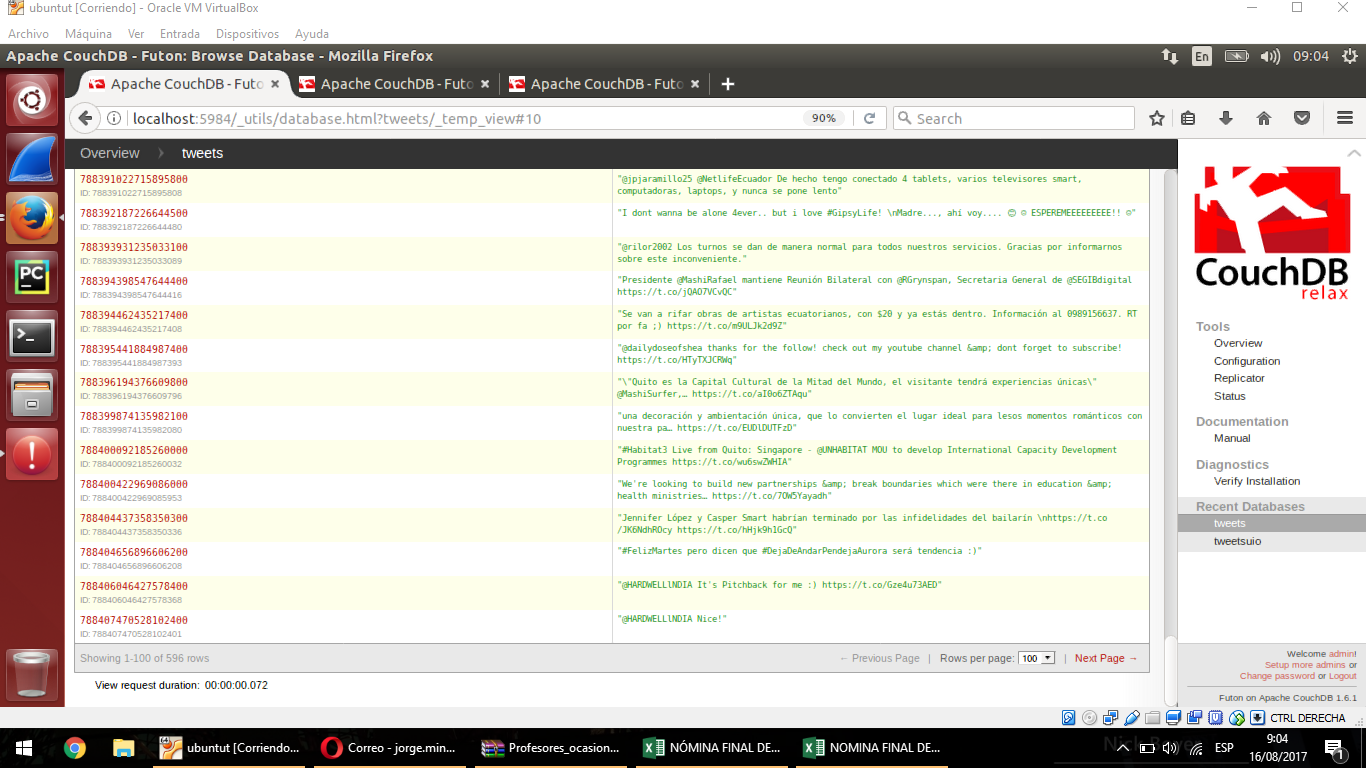


En la imagen se presenta el total de tweets recolectados.



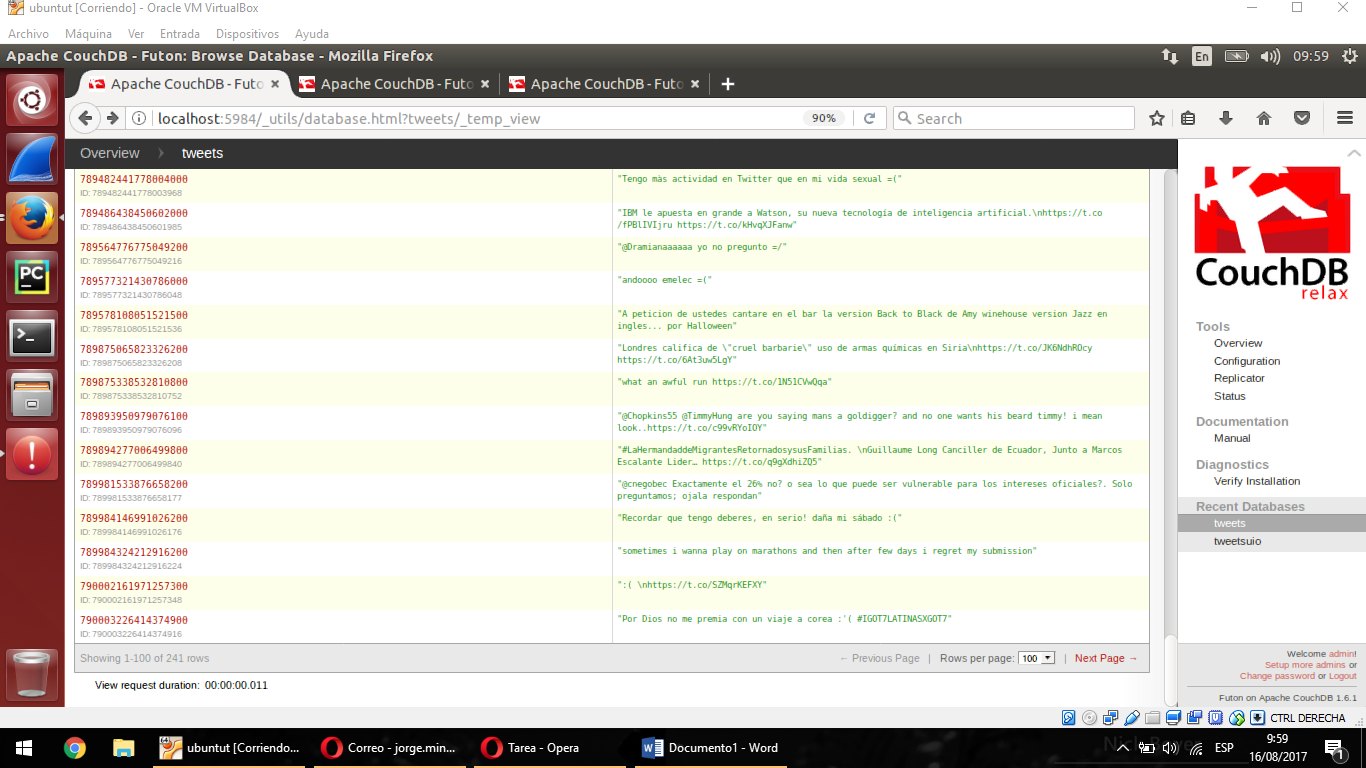
En el filtrado se encontró para cada característica que se desea implementar, se necesita crea una vista, los resultados obtenidos demuestran que el filtrado brinda una canticad de datos totales ya sea neutral, positivo, negativo.

* Datos totales filtrados positivamente.

**

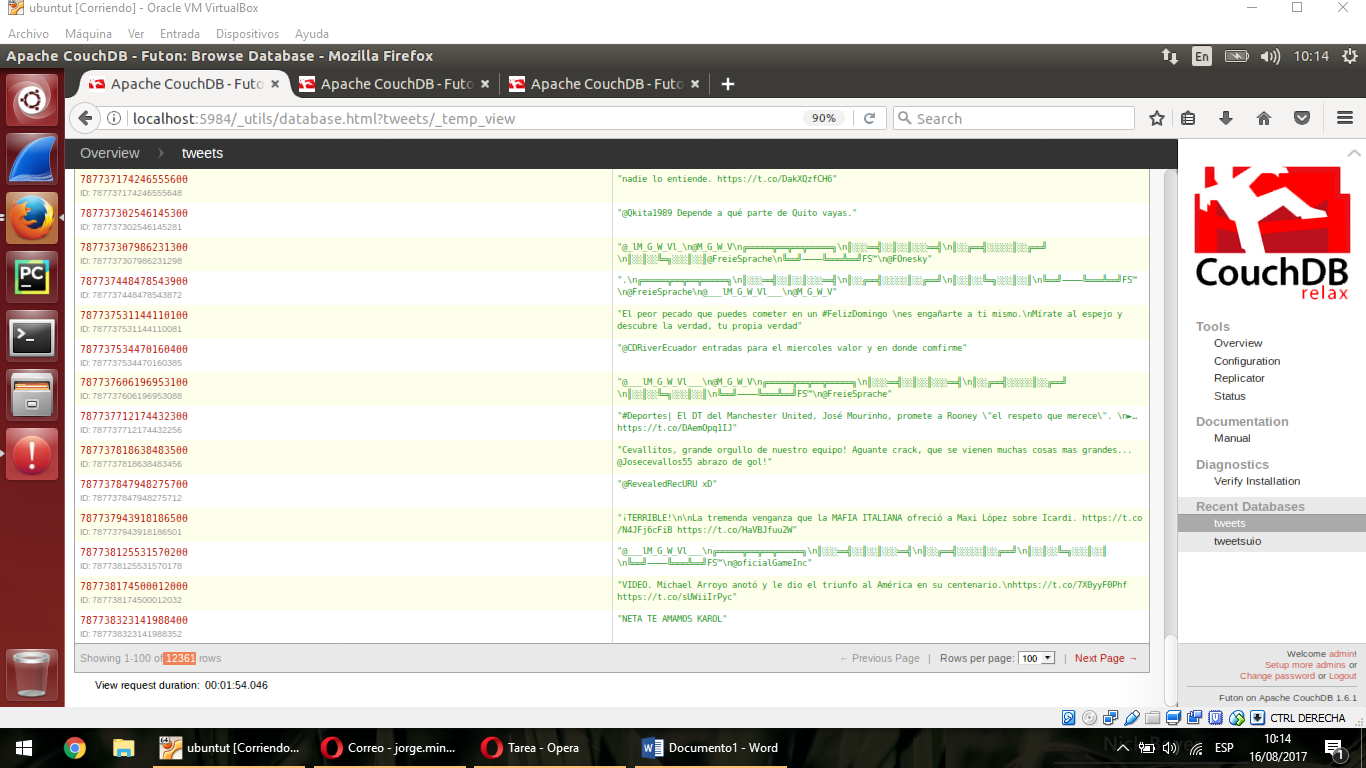
*La imagen muestra que la cantidad de datos filtrtados positivamente son 594.*

* Datos totales filtrados negativamente



*La imagen muestra la cantidad de datos filtrados negativamente son 241*

* *Datos tatales fitrados neutralmente.*

**

*La imagen muestra la cantidad de datos filtrados neutral mente*

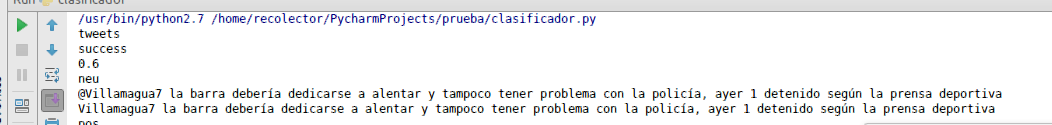


*En la siguiente imagen se vizualiza el codigo para subir los tweets al Elasticsearch lastimosamente no pudimos analizarlos*

# Graficacion resultados

**Presición**

La presicion obtenida fue del 0.6



Los Resultados obtenidos en el anaisis de datos muestra la siguiente grafica.

|  |  |
| --- | --- |
| Datos |  |
| Sentimiento | # Tweets |
| neu | 12361 |
| pos | 596 |
| neg | 241 |

Lo que permite evidenciar que la tendencia en las publicaciones de Twiter tiende a ser neutral, después positivo y por último negativo. Esto garantiza que las publicaciones en twiter pueden determinar el estado anímico de una persona al momento de publicar un tweeter.

* Los resultados obtenidos nos han facilitados conocer la condición de estado anímico que se encuentra la ciudad de Quito.

# conclusiones

* Se observo que Twitter ofrece una oportunidad sin precedentes para crear y emplear teorías y tecnologías que buscan y minan sentimientos. El trabajo presentado en este documento especifica un enfoque para el análisis de opiniones en los datos de Twitter.
* Se logra determinar cual es la tendencia del publico de la ciudad de quito, al momento de tweetear. Donde se demostró que la gramn mayoría tiende ha postear neutralmente.
* Para la clasificación de los datos se determino que es importante contar con un conocimiento coherente y consiso tanto endata como en lo que se desea obtener.
* El api de Tweeter facilita obtener información que ayuda ha conocer muchastendencias deacuerdo a lo que se desee conocer, esto facilita mucho trabajo que manualmente seria muy trabajoso y con un alto costo económico.
* Se observo que el presesamiento de los datos es una partae muy importante ya que, si no tiene un procesamiento óptimo todos los resultados podrían ser equivocados.

# Discusión

En el análisis de tweets concluimos que las personas en Quito tweetean de forma mas neutra en función de nuestro análisis debido a que la recolección de datos se verifico que eran datos informativos.

Tambien obtuvimos muchos tweets neutrales debido a nuestro clasificador tendría que complementarse con mas condiciones de sentimientos dentro de cada oración y asi obtener un mejor análisis.

Se debe tomar en cuenta que los tweets analizados fueron enfocados a los sentimientos en general y no a un área especifica si hubieras filtrado los tweets y descartar los que informaban cosas de deportes se hubiera obenido mejores resultados, también hay que tomar en cuenta que durante la clasificación se limpio primero el tweet para luego posterior se analizado esto tiende a quitar cierta

# Trabajos futuros

Como trabajos futuros se implementaría una mayor cantidad de training e implementar una nueva técnica de clasificación como la valoración de cada parte de la oración y en función de esto determinar el sentimiento del tweet.

# Enlace repositorio

<https://github.com/jorgemino/Proyecto_Tweets_clasificador_Mino_Jaguaco>

# Bibliografía

|  |  |
| --- | --- |
| [1] | Clasificación de Patrones (Duda, Hart) |
| [2] | Aprendizaje de la máquina (Tom Mitchell) |
| [3] | Introducción al Aprendizaje Automático (Ethem Alpaydin) |
| [4] | Redes Neuronales (Simon Haykin) |
| [5] | Aprendizaje de máquinas: un enfoque algorítmico (Marsland) |
| [6] | J Learning scikit-learn: Machine Learning in Python. |
| [7] | Programming ArcGIS 10.1 with Python Cookbook |
| [8] | Building Machine Learning Systems with Python |

|  |  |
| --- | --- |
| [9] | Miño Jorge, «GitHub,» 222, 22 2 2. [En línea]. Available: https://github.com/jorgemino/Proyecto\_Tweets\_clasificador\_Mino\_Jaguaco. [Último acceso: 16 8 2017]. |

1. [↑](#footnote-ref-1)