Análisis de Sentimientos en Twitter: Una Implementación sobre Cloudera

Soledad Retamar[1], Lautaro Ramos[1], Natalia Rapesta[1], Juan Pablo Nuñez[1], Patricia Cristaldo[1], Anabella De Battista[1], Norma Herrera[2]

1Dpto. de Sist. de Información, Univ. Tecnológica Nacional, FRCU, Entre Ríos, Argentina

{retamars, ramosl, rapestad, nuñezjp, cristaldop, debattistaa}@frcu.utn.edu.ar

2Dpto. de Informática, Universidad Nacional de San Luis, Argentina

nherrera@unsl.edu.ar

Abstract

Debido al crecimiento exponencial de las fuentes de información disponibles, en la actualidad resulta necesario contar con técnicas y herramientas diferentes a las tradicionales para abordar el procesamiento y análisis de los datos. En este trabajo se presentan las principales herramientas ofrecidas por la distribución de Cloudera del ecosistema Hadoop y el Lenguaje R para implementar un caso de estudio de análisis de sentimiento de tweets sobre la opinión de usuarios de esta red social sobre el proyecto de Ley Antidespidos discutido recientemente en el Congreso de la Nación Argentina.

1. Introducción

En la actualidad se producen a diario grandes volúmenes de datos de diversos tipos (textos, imágenes, audio, videos, entre otros) y desde los más variados orígenes (web, GPS, redes sociales, sensores, etc.). Se prevé que en los próximos años las aplicaciones Internet de las Cosas aumentarán el volumen de datos a un nivel sin precedentes.

En este contexto, surge el término Big Data referido a conjuntos de datos cuyo tamaño supera la capacidad de las herramientas tradicionales de bases de datos de recopilar, almacenar, gestionar y analizar la información. Big Data puede definirse a partir de las siguientes características [1]:

- *Volumen*: órdenes superiores a Terabytes de datos.
- *Variedad*: distintos tipos de datos provenientes de diversas fuentes que pueden organizarse tanto en forma estructurada como no estructurada.
- *Velocidad*: referido a la velocidad de generación de los datos o a la rapidez con la que se generan y procesan los datos.
- *Veracidad*: referido a las incertezas que pueden generar los datos, debido a que se generan en fuentes que pueden no ser confiables o a su baja calidad, lo que dificulta las tareas de análisis.

Las redes sociales son hoy en día una fuente de información en tiempo real inigualable, convirtiéndolas en una herramienta ideal para la realización de encuestas y sondeos

Las redes sociales han convertido a la web en un escenario de interacción social muy popular en el que billones de individuos alrededor del mundo interactúan, comparten, postean y conducen un sinnúmero de actividades diariamente [4]. Las redes sociales se han transformado en interesantes fuentes de informa-

ción para analizar fenómenos sociales, opiniones políticas, posturas éticas en temáticas públicas, opiniones e intereses sobre productos.

Una de las intenciones principales en el análisis de información es saber qué piensa la gente. En este contexto surgen aplicaciones como el análisis de sentimientos en un entorno Big Data, que consiste básicamente en el estudio computacional de la opinión de las personas con el fin de determinar sus actitudes y emociones ante ciertos temas o eventos [2,6].

Debido a que dichos estudios requieren del tratamiento de conjuntos de datos de gran tamaño y múltiples formatos resulta imposible abordarlos con las herramientas y técnicas tradicionales para el análisis de datos. Como solución a esta problemática surgen los sistemas basados en Hadoop, que utilizan almacenamiento distribuido y permiten procesar grandes volúmenes de datos, sean estos estructurados, semiestructurados o no estructurados [3,5].

Junto con la aparición de esta nueva tecnología han surgido gran cantidad de herramientas que facilitan su uso.

En este trabajo se describe, mediante un caso de estudio basado en análisis de sentimientos sobre una base de tweets, las principales herramientas para trabajar con Big Data en un entorno de Cloudera Hadoop y mediante el lenguaje R.

Lo que resta del artículo está organizado de la siguiente manera: en las secciones 2 y 3 se describe el contexto del trabajo, dando una breve reseña de la red social twitter y de la problemática del análisis de sentimientos. En la sección 4 se describen las herramientas utilizadas en nuestra aplicación para luego en la sección 5 detallar el caso de estudio. Finalmente en la sección 6 se presentan las conclusiones y el trabajo futuro.

2. La Red Social Twitter

Twitter se creó en 2006 como una red social en base a contenido con forma de SMS. Desde entonces ha crecido rápidamente ganando mucha popularidad en los últimos años. Según los datos oficiales que figuran en su web [12] son aproximadamente 320 millones de usuarios activos por mes que acceden diariamente para compartir experiencias y opiniones convirtiéndose así en una herramienta ideal para la realización de encuestas y sondeos.

Twitter permite a los usuarios enviar y leer mensajes de texto de hasta 140 caracteres, conocidos como tweets. Este sitio es una gran fuente de información subjetiva en tiempo real ya que estos millones de usuarios comparten opiniones sobre diferentes aspectos de su vida cotidiana [13].

Twitter ha sido usada para una variedad de propósitos en diferentes industrias y situaciones. Por ejemplo, los usuarios pueden encontrar o emitir opiniones sobre un producto o servicio de su interés, las compañías y figuras públicas pueden controlar su reputación en línea, se puede conocer la opinión de cada usuario respecto a implementación de políticas públicas, campañas de difusión, entre otras.

Twitter permite identificar estos temas a través de los denominados *hashtag* o *etiquetas*, que se caracterizan por comenzar con el carácter # y una cadena de caracteres a continuación formada por una o varias palabras concatenadas. Esta etiqueta funciona como metadato y permite que los tweets se añadan a una lista de mensajes con hashtag similares. Esto permite a los usuarios obtener resultados rápidamente sobre un mismo tema.

3. Análisis de Sentimientos

El análisis de sentimientos es una tarea incluida en el ámbito del *Procesamiento de Lenguaje Natural* o *NLP* (*Natural Language Processing*). Su objetivo es encontrar contenido subjetivo en los textos de entrada. El análisis de sentimientos busca extraer opiniones y la polaridad de éstas, mediante la identificación de características que determinen cuán positivo o negativo es el texto.

El análisis de sentimientos ha sido considerado inicialmente como una subdisciplina de clasificación basada en la opinión [7]. La técnica en general puede estar orientada a tres grandes tipos de tareas [8, 9]:

- *Clasificación de sentimientos*: realizar una clasificación de un conjunto de opiniones en tres categorías: positivas, negativas o neutrales. Presenta desafíos adicionales el hecho de que las opiniones se encuentran en múltiples idiomas o provienen de varios dominios, como biología, sociología, etc.

- *Clasificación de subjetividad*: determinar si una oración es subjetiva u objetiva. Una oración objetiva contiene información imparcial, mientras que una oración subjetiva contiene información de carácter personal como opiniones.

- *Resumen de opinión*: permitir extraer las características principales que son compartidas por uno o más documentos y el sentimiento acerca de estas características.

En el análisis de redes sociales, particularmente en análisis de sentimientos a partir del estudio de tweets para obtener su polaridad, existen dos métodos que son los más populares. El primero de ellos utiliza el enfoque del *aprendizaje computacional* (*machine-learning approaches*) [10] y el segundo utiliza *diccionarios léxicos* [11].

El *aprendizaje computacional* analiza la información automáticamente de forma supervisada, basándose en conjuntos de entrenamiento que son utilizados para catalogar al resto de las opiniones encontradas en la web, realizando pruebas y luego validándolas. Las principales técnicas de este método son: Support Vector Machines (SVM), Naive Bayes y Clasificadores de Máxima Entropía. En estas técnicas se utiliza la categoría gramatical de las palabras, la presencia y frecuencia de algunos términos y su composición semántica. Sin embargo, la mayoría de estos métodos son acompañados de algún diccionario que entrega información a priori de los términos para obtener las polaridades respectivas. En algunos casos, estos diccionarios son realizados por personas y en otros se ocupa un sistema semiautomático.

El método de *diccionarios léxicos* se basa en una lista de palabras con un determinado peso y/o categoría emocional. Estos diccionarios presentan principalmente adjetivos, que son los que aportan mayor información al momento de analizar los sentimientos, aunque también incluye verbos, adverbios y sustantivos. La mayor cantidad de versiones de diccionarios está en idioma inglés aunque existen algunas en español, pero en general en versión beta. Éstos permiten determinar si una frase es negativa o positiva dependiendo de la cantidad de palabras presentes en el diccionario y de la fuerza de su sentimiento. Uno de los diccionarios más completos es LIWC [19], presente una versión en inglés bastante completa y una en español que está aún en etapa de desarrollo. En este caso, las palabras son etiquetadas en una categoría determinada y además se les asigna una ponderación.

4. Herramientas y Tecnologías Utilizadas

En esta sección se describen el conjunto de herramientas y librerías utilizadas en el desarrollo de nuestra aplicación.

4.1 Obtención de Tweets

Twitter usa OAuth [http://oauth.net/] para facilitar el acceso a algunas APIs, OAuth es un protocolo abierto que permite la autenticación segura a través de un método simple y estándar para aplicaciones móviles, web o de escritorio.

La API Rest ofrecida por la empresa permite acceder mediante autenticación OAuth a un conjunto de tweets como documentos de JavaScript Object Notation (JSON). Se establece una conexión permanente con los servidores de Twitter mediante una aplicación y se realizan peticiones http con los parámetros de búsqueda deseados.

4.2 Apache Hadoop

Apache Hadoop, creado por Doug Couttin, es un framework que permite el procesamiento de grandes volúmenes de datos a través de clusters, usando un modelo simple de programación. Es un sistema distribuido que utiliza una arquitectura Master-Slave permitiendo pasar de pocos nodos a miles de nodos de forma ágil, ofreciendo cada uno de ellos poder de cómputo y almacenamiento local. Los componentes principales son los siguientes:

- El sistema de almacenamiento distribuido que utiliza fue creado a partir del Google File System (GFS) y es denominado Hadoop Distributed File System (*HDFS*). Este sistema fue preparado para almacenar y trabajar con grandes volúmenes de datos reduciendo la cantidad de E/S en la red para su lectura y escritura. Utiliza la replicación de los datos y la tolerancia a fallas, lo que aporta escalabilidad y disponibilidad. El HDFS se compone principalmente de un NameNode que gestiona los metadatos del sistema de archivos regulando el acceso de los nodos a los datos; y por otro lado los DataNodes que son los responsables de leer y escribir las peticiones de los clientes.

- Hadoop implementa un paradigma computacional llamado *MapReduce* para hacer cálculos, donde la aplicación se divide en pequeños fragmentos de trabajo, cada uno de los cuales se pueden ejecutar y/o volver a ejecutar en cualquier nodo del clúster. Este modelo simplifica el procesamiento en paralelo, abstrayendo al programador de la complejidad que hay en los sistemas distribuidos. Está basado en dos funciones principales: Map transforma el conjunto de datos a un número de pares (*key, value*). Cada uno de estos elementos se encontrará ordenado por su clave, y la función Reduce es usada para combinar los valores en un mismo resultado.

En la actualidad existen numerosas distribuciones basadas en Hadoop, entre las cuales se encuentran: Cloudera, HortonWorks, MapR, Azzure, DataStax, Pivotal, entre otras. La comparación y elección de estas herramientas puede realizarse en base a factores como: herramientas disponibles, productividad, tolerancia a fallos, rendimiento y administración. En [14] se realiza un estudio comparativo entre las principales distribuciones open-source de Big Data – HortonWorks, Cloudera y MapR, concluyendo en que las tres poseen características deseables y otras no, pero si se tiene en cuenta la productividad de la distribución, Cloudera es la escogida.

La distribución de Cloudera (CDH) fue la primera en aparecer en el mercado, combinando Big Data y Hadoop. Incluye el núcleo de Hadoop (HDFS y MapReduce) y diversos proyectos de Apache (HBase, Mahout, Pig, Hive, entre otros), es open-source, aunque cuenta con una interfaz gráfica propietaria, Cloudera Manager, para la administración y gestión de los nodos del clúster Hadoop. Ofrece una instalación a partir de una máquina virtual que contiene un clúster Hadoop, datos y scripts de ejemplo, y herramientas como Hue, Hive, Pig, Solr, entre otras utilidades. Para este trabajo se utilizó esta distribución instalada sobre un nodo simple.

4.3 Hive

Hive es una herramienta para data warehousing que facilita la creación, consulta y administración de grandes volúmenes de datos distribuidos en forma de tablas relacionales [15]. Cuenta con un lenguaje derivado de SQL, llamado Hive QL, que permite realizar las consultas sobre los datos. Por esta misma razón, se dice que Hive lleva las bases de datos relacionales a Hadoop. A su vez, Hive QL está construido sobre MapReduce, de manera que se aprovecha de las características de éste para trabajar con grandes cantidades de datos almacenados en Hadoop.

4.4 Lenguaje R

R es un lenguaje de programación interpretado que se distribuye de forma libre bajo licencia GNU. Inicialmente surgió con el objetivo de permitir realizar análisis estadísticos en ámbitos académicos. Pero en la actualidad es ampliamente utilizado para procesar y analizar grandes conjuntos de datos, dado que provee una considerable variedad de paquetes, desarrollados oficialmente o por la comunidad de usuarios, que implementan diversas funcionalidades para este propósito, como técnicas de data mining, gráficos o interfaces para conectarse con diferentes fuentes de datos [16].

Para el presente trabajo se utilizó una serie de paquetes, adicionales a los incluidos en la instalación de R, para recolectar y pre-procesar los datos sobre los que se realizó el análisis, y para presentar los resultados obtenidos a partir del mismo. Estos paquetes son:

- *twitterR*: brinda acceso a la REST API de twitter incluyendo funciones para recuperar y formatear conjuntos de tweets.
- *ROAuth*: presenta funciones que permiten autenticarse ante la REST API de twitter para poder operar mediante la misma.
- *plyr* y *stringr*: contienen funciones para la manipulación de cadenas de caracteres.
- *tm*: provee diversas funciones para realizar text mining.
- *ggplo2*: provee funciones para la elaboración gráficos de diversos tipos.

5. Caso de Estudio y Metodología

En esta sección se explican y detallan los algoritmos utilizados en un entorno Big Data para realizar el análisis de sentimientos sobre un caso de estudio de Twitter. Para este trabajo se seleccionó el enfoque de análisis de sentimientos basado en diccionarios.

Se presenta la arquitectura seleccionada y las herramientas con las que se realizó cada etapa.

5.1 Caso de Estudio

El caso de estudio se escogió teniendo en cuenta un tema que genere polaridad de opiniones en Twitter. En los últimos meses se debatió en el Congreso de la Nación Argentina el Proyecto de Ley Antidespido que proponía prohibir las cesantías de los contratos en el Estado hasta el 31 de diciembre de 2017 retroactivo al 1 de marzo de 2016. Ante las declaraciones públicas del Jefe de Estado cuestionando dicho proyecto aparecieron una serie de expresiones a través de Twitter muy controversiales y opuestas.

El acceso a la API se realiza utilizando el paquete twitterR, que proporciona una función específica para la descarga de tweets, requiriendo, entre otros, los siguientes parámetros de entrada: la cadena de búsqueda (por ejemplo hashtags, nombres de usuario o palabras clave), la cantidad máxima de tweets a recuperar y el lenguaje en el que se buscan los textos.

```
tweets <- searchTwitter(searchString='#LeyAntiDespidos',
n=1000, lang='es')
```

Los datos se obtienen en formato formato JSON y son organizados de forma tabular utilizando una función también perteneciente al paquete twitterR.

```
twListToDF(tweets)
```

Se recuperaron en esta ocasión 240.000 tweets, que se exportaron a un archivo en formato CSV. Los datos recuperados cuentan con diferentes atributos: un identificador, el texto del tweet, el nombre del usuario que lo publicó, la cantidad de veces que fue marcado como favorito y la cantidad de veces que se compartió (o retweeteo), entre otros. Para el análisis realizado sólo resulta de interés el identificador y el texto de los tweets, por lo cual se conservaron solamente estos atributos.

Fig. 1**.** Palabras y su forma canónica utilizadas en el proceso de normalización

Los textos presentan caracteres, símbolos y palabras, como pronombres, preposiciones y artículos que resultan irrelevantes para el cálculo de polaridad y por lo tanto deben ser eliminados. Para este propósito se desarrolló una función en R que permite eliminar signos de puntuación, caracteres de control, dígitos, nombres de usuario y links a páginas web, haciendo uso de distintas funciones de manipulación de cadenas de caracteres.

Se aplicó además una función del paquete *tm* que permite eliminar un conjunto de palabras; en nuestro caso eliminamos aquellas que no aportan información para el análisis de sentimientos. Dichas palabras son conocidas como *palabras vacías* o *stop words*, y fueron obtenidas a partir de un proyecto con licencia GNU creado en el sitio Google Code que recolecta este tipo de palabras para 29 lenguajes diferentes [17].

La cantidad de comparaciones necesarias para poder normalizar el texto requiere de un gran poder de procesamiento, por ello para llevar a cabo este análisis se utilizaron herramientas que implementen MapReduce aportando las ventajas en la velocidad de cálculo mencionadas anteriormente. Se crearon dos tablas en Hive: *tweets* y *DiccionarioDeLemas*, en las cuales se cargaron los datos de los Tweets y la lista de palabras utilizada para normalizar, quedando almacenados en el sistema de archivos HDFS.

La Figura 1 muestra un conjunto de palabras y su forma canónica utilizadas en el proceso de normalización.

A partir de esto se procesaron los textos de los Tweets mediante el lenguaje Hive QL para reemplazar cada palabra por su correspondiente lema existente en la tabla DiccionarioDeLemas.

En la Figura 2 se ilustra el contenido de la tabla tweets conteniendo las columnas del texto original y el texto resultante una vez normalizado.

5.2 Algoritmo de Análisis de Sentimientos

El análisis de sentimientos propuesto en Cloudera-Hadoop utiliza contadores MapReduce personalizados que presenta una característica interesante y es que permite personalizar el modo de calcular el valor de confianza o positividad.

En este análisis básico se plantea una puntuación total que puede ser considerada como un índice de positividad, que se calcula como el cociente entre la diferencia de cantidad de palabras positivas versus cantidad de negativas y la suma de dichas cantidades.

```
Sentimiento = (positivas - negativas)/(positivas + negativas)
```

Para ejecutar el algoritmo previamente se debe descargar el archivo sentimentAnalysis.tar.gz proporcionado por Cloudera, luego se descomprime para posteriormente acceder desde una terminal al directorio donde se han descomprimido dichos archivos. El

paquete está compuesto por las clases Map, Reduce y MrManager, además por tres archivos que contienen las palabras positivas, negativas y stopwords, el archivo makefile y un directorio donde se almacenan los datos a analizar.

El funcionamiento general del algoritmo consiste en leer cada línea que se ingresa como argumento, en nuestro caso los tweets, y en enviar dichas palabras a los métodos parsePostitive y parseNegative para que cada uno identifique las palabras en sus respectivas listas.

Fig. 2. Tweets originales vs. normalizados

El método parsePositive itera sobre la lista de términos y crea una entrada por cada palabra que determine como positiva. Del mismo modo trabaja el método parseNegative.

Mientras el método Map cuenta las palabras positivas y negativas, el método Reduce reúne los resultados y se los devuelve al método MrManager. Este método no devuelve los resultados de inmediato sino que almacena el resultado en una variable, permitiendo trabajar con los resultados e ingresar comandos por consola.

Para implementar este algoritmo sobre los tweets relevados con el hashtag #LeyAntiDespidos se reemplazaron los archivos de la distribucion CHD poswords.txt, neg-words.txt y stop-words.txt, por sus equivalentes en español.

Se filtró sólo la columna de textos normalizados de los tweets y luego se exportó esta tabla a un archivo de texto:

```
CREATE TABLE texto(text STRING);
INSERT INTO texto SELECT text FROM
tweets_ley;
EXPORT TABLE texto TO
'/user/cloudera/analisis/'
```

Este archivo fue colocado en el directorio de datos del algoritmo *SentimentAnalysis*. Una vez exportada la tabla se ejecutó el algoritmo a través del comando:

```
run makefile
```

En la Figura 3 se muestra el algoritmo en ejecución donde se pueden observar la cantidad de Map y Reduce realizados

Tras la ejecución del algoritmo se pueden verificar las frecuencias de las palabras en archivos dentro del sistema HDFS, y por consola (Figura 4) se visualiza el coeficiente final calculado, que para nuestro caso de estudio es de 38% de Positividad. Esto indicaría que el 38% de las opiniones volcadas en Twitter sobre la ley Anti despidos han sido positivas y el 23% han sido opiniones negativas.

Luego de la ejecución, se analizó el archivo de salida que genera el algoritmo, que contiene las frecuencias de cada palabra (Figura 5) y se observó que un gran porcentaje de las opiniones eran evaluadas como nulidad, y el motivo de esto es porque en los textos de los tweets analizados existen muchas palabras que no se encuentran en los archivos de palabras positivas ni negativas.

```
16/04/22 12:13:11 INFO impl.YarnClientImpl: Submitted application application_1
461016490465_0014
16/04/22 12:13:13 INFO mapreduce.Job: The url to track the job: http://quicksta
rt.cloudera:8088/proxy/application_1461016490465_0014/
16/04/22 12:13:13 INFO mapreduce.Job: Running job: job_1461016490465_0014
16/04/22 12:14:10 INFO mapreduce.Job: Job job_1461016490465_0014 running in uber mode : false
16/04/22 12:14:10 INFO mapreduce.Job:  map 0% reduce 0%
16/04/22 12:17:52 INFO mapreduce.Job:  map 17% reduce 0%
16/04/22 12:17:55 INFO mapreduce.Job:  map 67% reduce 0%
16/04/22 12:17:57 INFO mapreduce.Job:  map 92% reduce 0%
16/04/22 12:18:09 INFO mapreduce.Job:  map 100% reduce 0%
16/04/22 12:19:19 INFO mapreduce.Job:  map 100% reduce 67%
16/04/22 12:19:23 INFO mapreduce.Job:  map 100% reduce 80%
16/04/22 12:19:27 INFO mapreduce.Job:  map 100% reduce 100%
16/04/22 12:19:34 INFO mapreduce.Job: Job job_1461016490465_0014 completed successfully
```

Fig. 3. Algoritmo *Sentiment Analysis* en funcionamiento

```
**********


Sentiment score = (56158 - 90443) / (56158 + 90443)
Sentiment score = -0.23386607


Positivity score = 56158 / (56158 + 90443)
Positivity score = 38%


**********
```

Fig. 4. Resultado de la ejecución del Algoritmo SentimenAnalysis

Fig. 5. Archivo de salida conteniendo la frecuencia de cada palabra encontrada

Las palabras con sus respectivas frecuencias brindan información adicional al análisis de sentimientos, ya que permiten determinar los conceptos principales mencionados junto con una temática en particular o anexar las palabras más usadas al diccionario.

6. Conclusiones y Trabajo Futuro

En este trabajo se abordó un estudio de análisis de sentimientos, tomando como objeto de estudio la red social Twitter, y como caso particular el análisis de tweets sobre la Ley Antidespidos que se trató recientemente en el Congreso de la Nación Argentina.

Respecto al caso de estudio analizado se pudo comprobar que los métodos utilizados son altamente dependientes de los diccionarios con el que se comparan los tweets recolectados. Esto es una desventaja al momento de analizar textos que contienen muchas palabras del lenguaje informal.

Por otro lado, en la técnica utilizada no se consideran intensificadores o palabras que inviertan la polaridad del mensaje, por lo tanto, no se podría asegurar que los resultados obtenidos reflejen realmente la opinión de las personas sobre la temática estudiada. Será necesario abordar algoritmos más complejos y profundizar en el pre-procesamiento de los datos para obtener un mayor grado de validez en los resultados.

En cuanto a las herramientas utilizadas se pudo observar la gran versatilidad que proveen al momento de trabajar con diferentes formatos y estructuras de datos. Como trabajo futuro se plantea el abordaje de algoritmos más complejos de análisis de sentimientos y la posibilidad de hacerlo en tiempo real. También se espera poder enriquecer el diccionario de palabras en español, agregando información relacionada con la clasificación de las palabras y ponderaciones con el objetivo de aportar mayor precisión a estas técnicas.

7. Referencias

1. IBM Big Data & Analytics Hub. http://www.ibmbigdatahub.com/infographic/four-vs-big-data
2. Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2 (2008), no. 1-2, 1–135.
3. The Apache Software Foundation. What is Apache Hadoop? Hadoop, Apache https://hadoop.apache.org (2016). Accedido el 15 de Abril de 2016.
4. Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. Social Media Mining: An Introduction. Cambridge University Press, 2014. Draft version: April 20, 2014
5. Fan Wei and Bifet Albert. Mining big data: Current status, and forecast to the future. SIGKDD Explor. Newsl., 14(2):1-5, apr 2013.
6. Gil Pérez, Borja, et al. TDC (Twitter Data Collection): Creación de una gran base de datos de Tweets. 2014.
7. Fermín L. Cruz, José A. Troyano, Fernando Enriquez, Javier Ortega: Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español. Procesamiento del lenguaje Natural, no 41 (2008), pp. 73-80
8. Martínez Peláez, José Juan (2015), Una plataforma base para Big Data, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, México D.F., México
9. Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 2014.
10. Montesinos, Lucas (2014). Análisis de sentimientos y predicción de eventos en Twitter. Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile. Chile
11. Ronen Feldman, Benjamin Rosenfeld, Roy Bar-Haim, and Moshe Fresko. The stock sonar | sentiment analysis of stocks based on a hybrid approach. 2011.
12. Twitter. https://twitter.com (2016). Accedido el 22 de Abril de 2016
13. Fernández Javi, Gutiérrez Yoan, Gómez José M., Martínez-Barco Patricio, Montoyo Andrés, Muñoz Rafael. Sentiment Analysis of Spanish Tweets Using a Ranking Algorithm and Skipgrams. IV Congreso Español de Informática. Septiembre 2013, Madrid, España
14. Perreau de Pinninck, L. (2015) Análisis y desarrollo de una plataforma Big Data. Universidad Pontificia Comillas, Madrid, España
15. Apache Hive. The Apache Software Foundation. Accedido el: 5 de Abril de 2016. https://cwiki.apache.org/confluence/display/Hive/Home
16. El arte de programar en R: un lenguaje para la estadística. Julio Sergio Santana. Efraín Mateos Farfán. ISBN: 978- 607-9368-15-9. 2014
17. https://code.google.com/archive/p/stop-words/
18. A. Gelbukh, G. Sidorov. Approach to construction of automatic morphological analysis systems for inflective languages with little effort. In: Computational Linguistics and Intelligent Text Processing (CICLing-2003), Lecture Notes in Computer Science, N 2588, Springer-Verlag, 2003, pp. 215–220.
19. Sitio web diccionario LIWC. http://liwc.wpengine.com/ Accedido el 22 de Abril de 2016.