International Journal of Innovation and Applied Studies
ISSN 2028-9324 Vol. 16 No. 4 Jun. 2016, pp. 814-819

http://www.ijias.issr-journals.org/

Análisis de sentimientos usando el API de Twitter

[Feelings analysis using the Twitter API]

Gary Reyes Zambrano, Jonathan Reyes Tomalá, and Wellington Aroca Albiño

Facultad de Ciencias Matemáticas y Físicas,
Universidad de Guayaquil,
Guayaquil, Ecuador

ABSTRACT: In this paper the process for analyzing feelings described using applications such as: Hadoop in version 2.3.2, and the facilities provided by the API (Application Programming Interface, for its acronym in English) from Twitter for extraction and information processing (Tweets) of the University of Guayaquil - Ecuador. So you can evaluate the information obtained by running different scripts, containing algorithms required for the analysis of feelings and determine if it is a positive, negative or neutral comment. Thus obtain the final result information to help determine the feelings of users of the account of the University of Guayaquil. This information is very helpful to support the decision making process.

KEYWORDS: Sentiment analysis, Twitter, Guayaquil´s University, Application Programming Interface.

RESUMEN: En la presente investigación se describe el proceso de análisis de sentimiento utilizando las aplicaciones Hadoop en su versión 2.3.2, así como las facilidades que provee la API de Twitter para la extracción y procesamiento de la información (tweets) de la Universidad de Guayaquil en Ecuador. De esta manera es posible evaluar la información obtenida mediante la ejecución de diferentes scripts, los cuales contienen los algoritmos requeridos para el análisis de sentimientos y así poder determinar si los comentarios son positivos, negativos o neutrales. De esta manera se obtiene la información resultante para determinar el sentir de los usuarios en la Universidad de Guayaquil. Esta información es de vital importancia para el proceso de toma de decisiones.

PALABRAS-CLAVE: Análisis de sentimientos, Twitter, Universidad de Guayaquil, API.

1 INTRODUCCIÓN

Las redes sociales en la actualidad constituyen un gran banco de información social. Millones de personas introducen información de sus vidas en cada una de ellas y la comparten con miles de usuarios. Una de las redes sociales más populares es Twitter, en ella se puede hallar gran cantidad de información referente a cualquier tema en particular que se desea conocer. De manera general este gran banco de información, que constituye las redes sociales, puede ser explotado con el objetivo de obtener información variada tanto del entorno como de los usuarios. Entre los diferentes tipos de información que se puede consultar se encuentra la información tecnológica. En varias ocasiones se requiere de información sobre algún tema específico y realizar la recolección de esta mediante las redes sociales resultaría lento e ineficiente. La clasificación de la información resultaría aún más difícil, por lo que separar los aspectos positivos y negativos requiere de un gran esfuerzo. Para disminuir esta situación se crea la técnica de procesamiento de información masiva denominada análisis de sentimientos.

Corresponding Author: Gary Reyes Zambrano

El análisis de sentimientos, también definido como minería de opción, es el procesamiento del lenguaje natural para identificar y extraer información subjetiva, información basada en el estado de ánimo de cada individuo. El análisis de sentimientos busca determinar la actitud del interlocutor con respecto a un tema específico o la polaridad contextual general de un documento. En específico busca conocer si lo que escribe el interlocutor es positivo o negativo, su impacto sobre el tema.

La determinación el concepto en sí del análisis de sentimientos realizado permite lograr determinar si los Tweets extraídos sobre un tema específico contenían información positiva o negativa para su posterior clasificación y almacenamiento. Esto propicia la obtención de información directamente de los usuarios finales. En la actualidad, este tipo de procesamiento, está teniendo gran aplicación debido a la creciente competencia en el mercado, pues de ello depende el éxito o fracaso de las muchas empresas.

Por la relevancia del tema la presente investigación tiene como objetivo realizar un análisis de sentimientos a la cuenta de Twitter de la Universidad de Guayaquil. Con el estudio se desea revelar el índice de positivismos del que dispone actualmente la misma para determinar su situación actual. El presente artículo se organiza de la siguiente manera: en la sección 2 se evalúan los requerimientos necesarios para realizar el análisis, en la sección 3 se desarrolla la investigación, en la sección 4 se analizan los resultados y en la sección 5 se exponen las conclusiones y el trabajo futuro.

2 REQUERIMIENTOS

Como condición inicial para llevar a cabo la presente investigación se requiere de equipamiento con gran capacidad de almacenamiento. Para realizar el procesamiento del análisis de sentimiento se necesita de una gran capacidad de procesamiento, utilizando como micro un Intel i7 con 6 GB como mínimo de memoria RAM. Esto se debe a que los aplicativos que se requerirán demandaran de gran cantidad de recursos de hardware y se desea evitar cualquier tipo de retraso o inconveniente que pueda afectar el proceso de análisis de sentimientos. Adicionalmente se requerirá de una serie de aplicativos los cuales se listan a continuación:

- Un software virtualizador para lo cual se recomienda la utilización de VirtualBox en su última versión.
- Como aplicativo principal se recomienda la utilización de Hortonworks sandbox para VirtualBox en su versión 2.3.2.

Figura 1. Logo de aplicativo Hortonworks sandbox

- Se recomienda la utilización de Hive ODBC Driver for HDP 2.3 (v2.0.5), que al igual que el anterior es un aplicativo de Hotonworks. Este permitirá establecer la conexión con la base datos.
- Se recomienda la utilización de WinSCP 5.7.6 para facilitar la conexión y manipulación de archivos.

Figura 2. Logo de WinSCP.

3 DESARROLLO

Para el desarrollo de la investigación se implementó una API en Twitter denominada ProEle como se muestra en la figura 3.

Figura 3. API creada en Twitter.

ProEle permitirá la conexión y extracción de la información contenida en los Tweets. Esta información es esencial pues constituye la entrada del proceso de análisis de sentimientos. De cada tweet analizado se extrae la llave del consumidor, el secreto del consumidos, el toquen de acceso y el secreto del toquen de acceso como se muestra en la figura 4. Estos datos servirán como códigos de acceso asignados a la API para configurar los aplicativos y proceder a la extracción de la información requerida.

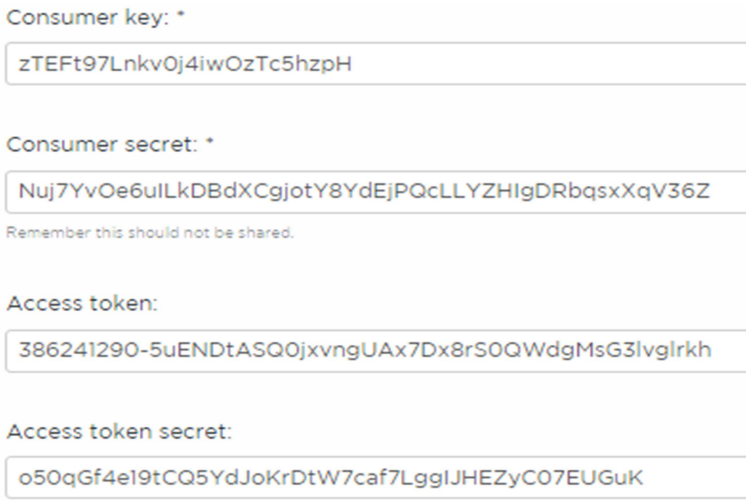

Figura 4. Detalle de información proporcionada por ProEle.

Luego de obtener los códigos de acceso se procederá a configurar FLUME (aplicativo incluido en Hortonworks sandbox). La configuración de este aplicativo se realiza en el archivo de configuración flume.conf. En él se detalla la información antes recolectada por ProEle pues la ejecución de FLUME posibilitará que la información viaje primero a este archivo para iniciar la extracción los parámetros antes delineados. Adicionalmente se podrá extraer información que se detalla y es requerida de los Tweets sobre el tema. Como limitación se presenta que la cantidad de Tweets a extraer se limitará a 1000 para acelerar el proceso de recolección de información y serán almacenados como muestra la figura 5:

```
TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks = HDFS

TwitterAgent.sources.Twitter.type = org.apache.flume.source.twitter.TwitterSource
TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sources.Twitter.consumerKey = zTEFt97Lnkv0j4iwOsTc5hspH
TwitterAgent.sources.Twitter.consumerSecret = Nuj7Yv0e6uILkDBdXCgjotY8YdEjPQcLLYZHIgDRbqsxXqV36Z
TwitterAgent.sources.Twitter.accessToken = 386241290-5uENDtASQ0jxvngUAx7Dx8r30QWdgMsG3lvglrkh
TwitterAgent.sources.Twitter.accessTokenSecret = o50qGf4e19tCQ5YdJoKrDtW7caf7LggIJHEZyC07EUGuK
TwitterAgent.sources.Twitter.keywords = Univ. de Guayaquil, universidad estatal de guayaquil,UG

TwitterAgent.sinks.HDFS.channel = MemChannel
TwitterAgent.sinks.HDFS.type = hdfs
TwitterAgent.sinks.HDFS.hdfs.path =hdfs://10.0.1.195/data/customer/tweets/landing/year=%Y/month=%m
TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream
TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text
TwitterAgent.sinks.HDFS.hdfs.batchSize = 1000
TwitterAgent.sinks.HDFS.hdfs.rollSize = 0
TwitterAgent.sinks.HDFS.hdfs.rollCount = 10000

TwitterAgent.channels.MemChannel.type = memory
TwitterAgent.channels.MemChannel.capacity = 10000
TwitterAgent.channels.MemChannel.transactionCapacity = 100
```

Figura 5. Contenido del archivo de configuración de FLUME flume.conf

Para almacenar la información extraída de la ejecución de FLUME se crea una base de datos. El resultado final del proceso de extracción de información se almacena en una tabla que representa la información de manera ordenada y clasificada, facilitando la comprensión de la misma. Esto facilitará el procesamiento de la información como en un manejador de base de y la visualización de los datos en la tabla si así se desea. Otra de las ventajas consiste en el desarrollo de una aplicación que se conecte con la base de datos y muestre los datos según un conjunto de criterios predefinidos.

Este paso es crucial para la toma de decisiones. El análisis de los tweets, clasificados en positivos y negativos, por el método de análisis de sentimiento organiza la información y la almacena en la base de datos HIVE. Para facilitar la toma de decisiones es necesario la utilización de herramientas especializadas. En la presente investigación se desea apoyar a la toma de decisiones en la universidad de Guayaquil mediante la organización de la información en gráficos estadísticos. De esta manera la visualización y la comprensión de la misma mejora de manera considerable.

Para la visualización de la información se utilizó Microsoft Excel del paquete de Microsoft Office. Se realizó una configuración previa en la PC servidor relacionado al origen de los datos. Esta configuración permitirá establecer la conexión entre la PC cliente y la base de datos. Posteriormente se inició Excel y se agregó la configuración del origen de los datos en la hoja de cálculo. Esto permite que sean importados los datos en una tabla dinámica, lo que facilita la obtención de los datos clasificados en una hoja de cálculo simple y a su vez realizar diferentes tipos de filtrado de datos sobre la tabla dinámica. El filtrado de los datos permite agregar o quitar la información que se desea visualizar según las necesidades del usuario para la inspección de la información final. En el caso de requerir una visualización global de la información se utilizará PowerMap para visualizar la tabla dinámica una vez filtrada y así visualizar la información de manera global como se puede observar en la figura 6:

Figura 6. Visualización de información usando PowerMap.

La utilización de estas herramientas proporciona gráficos estadísticos con la información clasificada. En estos gráficos se puede apreciar la información extraída de los tweets de diferentes maneras, ya sea desde diferentes tipos, formas. Con el uso de la herramienta PowerMap se podrán realizar diferentes gráficos estadísticos como gráficos de pasteles o diagramas de barras. De esta manera se podrá apoyar a la toma de decisiones pertinentes de requerirlo el caso u obtener un reporte del estado actual de un tópico en particular. Este tipo de análisis y representación tiene una marcada relevancia para conocer si la situación en la que se esté involucrado es la adecuada o se está cumpliendo con las expectativas deseadas.

4 RESULTADOS

En la presente investigación se ha realizado la extracción de información de un conjunto de tweets obtenidos en la Universidad de Guayaquil utilizando la técnica de análisis de sentimiento. Como resultado se obtuvo una base de datos con la información extraída, clasificada en positiva, negativa y neutrales para su posterior análisis. La información puede ser visualizada desde cualquier PC cliente utilizando las herramientas y la metodología propuesta. La facilidad proporcionada por la suite de Microsoft, específicamente Microsoft Excel permite la visualización de la información de manera sencilla mediante el uso de tablas dinámicas, gráficos de barras, gráficos estadísticos. Como alternativa a esta herramienta se propone PowerMap, mediante la cual se puede realizar la visualización de gráficos estadísticos lo que dará una mejor perspectiva visual de la información que se dispone, sectorizada globalmente, para apoyar a la toma de decisiones.

Otro de los resultados del proyecto es un conjunto de servicios que pueden ser implementados a partir de los datos obtenidos. Estos servicios pueden ser empleados en análisis posteriores durante un proceso de evaluación o análisis de sentimientos periódico de la Universidad de Guayaquil.

5 CONCLUSIONES Y TRABAJO FUTURO

Con el desarrollo de la presente investigación se identificó un conjunto de requisitos mínimos de hardware para la ejecución del proyecto y la utilización de sus servicios. La implementación del API ProEle facilitó la extracción de información de los tweets y su clasificación para ser almacenada en una base de datos. La extracción y clasificación de la información contenida en los tweets facilitó el análisis de la información. La representación de la información en gráficas y tablas sirvió de apoyo a la toma de decisiones y posibilitó el análisis de la problemática planteada en la Universidad de Guayaquil. Como trabajo futuro se propone expandir el proyecto a otras universidades del país. La captura de los requisitos específicos por universidad y las necesidades particulares mejorarían la usabilidad de la aplicación, con gran impacto social. Se prevé además la inclusión de varias técnicas de tomas de decisiones que facilitarán el trabajo de los expertos.

AGRADECIMIENTO

A la universidad de Guayaquil por el apoyo brindado durante la investigación y por facilitar los datos y el equipamiento necesario para llevar a cabo la investigación. Al equipo de trabajo que apoyó la implementación de ProEle y diseño la base de datos que fue utilizada en la investigación para almacenar los datos clasificados.

REFERENCIAS

[1] HEnriquez, Carlos. "Analisis de sentimiento." prueba 1.1 (2015).
[2] White, Tom. Hadoop: The definitive guide. " O'Reilly Media, Inc.", 2012.
[3] Das, Devaraj, et al. "Adding Security to Apache Hadoop." hortonworks report, http://www. Hortonworks. com (2011).
[4] Georgiou, Anastasia. Storing Data Flow Monitoring in Hadoop. No. CERN-STUDENTS-Note-2013-144. 2013.
[5] Foley, Matt. "High availability HDFS." 28th IEEE Conference on Massive Data Storage, MSST. Vol. 12. 2012.
[6] O'Malley, Owen. "Hadoop Benchmarking." (2012).
[7] Russom, Philip. "Integrating Hadoop into Business Intelligence and Data Warehousing." TDWI Best Practices Report (2013).
[8] Reddy, Y. "Access control for sensitive data in hadoop distributed file systems." Third International Conference on Advanced Communications and Computation, INFOCOMP. 2013.
[9] Murthy, Arun C., et al. Apache Hadoop YARN: Moving Beyond MapReduce and Batch Processing with Apache Hadoop 2. Pearson Education, 2013.
[10] Wadkar, Sameer, and Madhu Siddalingaiah. "Apache Ambari." Pro Apache Hadoop. Apress, 2014. 399-401.
[11] Aravinth, Mr SS, et al. "An Efficient HADOOP Frameworks SQOOP and Ambari for Big Data Processing." International Journal for Innovative Research in Science and Technology 1.10 (2015): 252-255.
[12] Padhy, Rabi Prasad, and Deepti Panigrahy. "A Gentle Introduction to Hadoop Platforms."
[13] Faghri, Faraz, et al. "Failure scenario as a service (FSaaS) for hadoop clusters." Proceedings of the Workshop on Secure and Dependable Middleware for Cloud Monitoring and Management. ACM, 2012.
[14] Arevalo, cabrera, et al. "uso de la plataforma pig sobre hadoop como alternativa a una rdbms para el analisis de datos masivos. Prueba de concepto utilizando registros de detalles de llamadas." (2010).
[15] Moncada Cerón, Jesús Salvador. "Big data en las empresas: una nueva era de la información." (2015).
[16] Hernández Dominguez, Antonio, and Adrian Hernández Yeja. "Acerca de la aplicación de MapReduce+ Hadoop en el tratamiento de Big Data." Revista Cubana de Ciencias Informáticas 9.3 (2015): 49-62.
[17] Dong, Fei. Extending starfish to support the growing hadoop ecosystem. Diss. Duke University, 2012.
[18] Wadkar, Sameer, and Madhu Siddalingaiah. "HCatalog and Hadoop in the Enterprise." Pro Apache Hadoop. Apress, 2014. 271-282.
[19] Arora, Nitika. "Hadoop: Components and Working." International Journal of Advanced Research in Computer Science 6.7 (2015).
[20] Analyzing Social Media and Customer Sentiment.http://hortonworks.com/hadoop-tutorial/how-to-refine-and-visualize-sentiment-data/.
[21] Analyse Tweets using Flume, Hadoop and Hive. http://www.thecloudavenue.com/2013/03/analyse-tweets-using-flume-hadoop-and.html.
[22] How to Install and Configure the Hortonorks ODBC driver on Windows 7. http://hortonworks.com/hadoop-tutorial/how-to-install-and-configure-the-hortonworks-odbc-driver-on-windows-7/.
[23] ELiRF-UPV en TASS-2013: Análisis de Sentimientos en Twitter. http://users.dsic.upv.es/~lhurtado/papers/pdfs/2013_pla13_tass.pdf, Ferran Pla y Llu´ıs-F. Hurtado.
[24] Análisis de sentimientos de tweets. http://www.cyt.uc.edu.py/jit-cita/2013/images/Trabajos/jitcita2013_NunhezJaraPezzino.pdf, Jorge José Jara Ruiz.