RESUMEN ANALÍTICO EN EDUCACIÓN - RAE -	UNIVERSIDAD CATÓLICA de Colombia Vigilada Mineducación RIUCaC

FACULTAD DE INGENIERÍA
PROGRAMA DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
PREGRADO EN INGENIERÍA DE SISTEMAS
BOGOTÁ D.C.

AÑO DE ELABORACIÓN: 2017

TÍTULO: Desarrollo y aplicación de una herramienta de extracción y almacenamiento de datos de twitter a un contexto social de violencia política.

AUTOR (ES): Barriga Mariño, José Camilo

DIRECTOR(ES)/ASESOR(ES): Rincón Yáñez, Diego Alberto

MODALIDAD: Auxiliar de investigación.

PÁGINAS: 106 **TABLAS:** 3 **CUADROS:** 0 **FIGURAS:** 32 **ANEXOS:** 1

CONTENIDO:

INTRODUCCIÓN
1. GENERALIDADES
2. OBJETIVOS DEL PROYECTO
3. MARCO DE REFERENCIA
4. MARCO CONCEPTUAL
5. METODOLOGÍA
6. DESARROLLO DEL PROYECTO
7. RESULTADOS

CONCLUSIONES
RECOMENDACIONES
BIBLIOGRAFÍA
ANEXOS
GLOSARIO

RESUMEN ANALÍTICO EN EDUCACIÓN - RAE -	UNIVERSIDAD CATÓLICA de Colombia Vigilada Mineducación RIUCaC

DESCRIPCIÓN: Este proyecto se orientó a la construcción de una herramienta web para la extracción y almacenamiento de datos de la red social twitter, la cual permitirá a futuro con apoyo de un software externo o integrado, establecer un análisis estadístico de estos datos, enfocado en la necesidad del usuario y también alentar la construcción de nuevas solución.
.

METODOLOGÍA:

Para su realización se integraron bases teóricas enfocadas en conceptos relacionados con BigData, Almacenamiento y minería de datos, haciendo énfasis en los procesos de extracción, almacenamiento y curación de datos. Así mismo, se realizó un proceso de desarrollo de software en el cual se aplicó la metodología PXP, la cual es una adaptación de la metodología de programación extrema enfocada al desarrollo llevado a cabo por un solo programador.
Las principales herramientas implicadas en el desarrollo de la herramienta se listan a continuación:

- **Eclipse Mars 2.**
- **MySQL 5 7**
- **MongoDB 3.0**
- **Liferay portal 6.2 ce ga5**
- **API de Twitter**
- **Twitter4j**

PALABRAS CLAVE:

ALMACENAMIENTO DE DATOS, BIGDATA, DESARROLLO DE SOFTWARE, EXTRACCIÓN, MINERÍA DE DATOS

CONCLUSIONES:

Luego de la ejecución del proyecto y tomando como base los resultados obtenidos referentes a los procesos de extracción y almacenamiento de datos de la red social twitter, tanto en la fase de pruebas del proyecto, como en el resultado visual de los datos y teniendo en cuenta la calidad de estos, la cual es necesaria para

poder establecer un almacenamiento sencillo y eficaz y así posteriormente con el apoyo de desarrollos o integraciones futuras, poder generar un análisis estadístico de varianza, con posibilidad de enfoque al a análisis de sentimientos y con un carácter predictivo. Se concluye de esta forma, que los objetivos planteados para el desarrollo del proyecto y enfocados en la descripción de la problemática a tratar fueron cumplidos, abarcando en sí cada aspecto planteado para el desarrollo de una forma adecuada que permitió con base al alcance y limitaciones del proyecto, brindar la solución esperada por los interesados tras este ciclo de implementación.

El desarrollo del proyecto permite definir a la extracción y almacenamiento de datos de redes sociales como una práctica o conjunto de procesos que implica la comprensión de una base teórica referente a arquitectura de software, programación y almacenamiento de datos, ya que sin esta base conceptual el entorno de desarrollo de una herramienta que permita establecer estas funcionalidades es invisible a los ojos del desarrollador.
Los tipos de datos recolectados en una extracción a un medio definido pueden variar, y esta estructura variable entre ellos implica el uso de diferentes técnicas de manejo de datos y sistemas de almacenamiento. El correcto tratamiento de estos datos puede definir en su totalidad los resultados obtenidos al culminar el proceso de desarrollo. En este caso modelos relacionales y no relacionales fueron integrados en el mismo sistema, con el fin de alcanzar el nivel de calidad y cumplir con los objetivos del proyecto, encontrando diferentes ventajas al definir correctamente la estructura general del desarrollo.
Definitivamente la extracción de datos de redes sociales, aplicando una metodología de desarrollo ágil como lo es el concepto de Personal Extreme Programming, seleccionado para la construcción de la herramienta y siguiendo un ciclo de investigación, planificación, diseño, desarrollo y pruebas, provee al implementador de una gran cantidad de información, que en una diferente fuente sería muy difícil de recuperar. El uso de herramientas como las APIs, en este caso generadas por comunidades de desarrolladores de las diferentes redes sociales, permite implementar de forma gratuita desarrollos de minería de datos, minería web e inclusive desarrollos que apuntan al concepto BigData en cuanto a la recolección masiva de datos para su posterior análisis y visualización. Lo que permite que el concepto de datos abiertos se reproduzca y las soluciones a diferentes problemáticas se den de una manera mucho más sencilla.
La formación de un desarrollador frente a un proyecto con un concepto tan amplio, el cual implica el manejo de una gran variedad de herramientas y conocimientos de la ingeniería de sistemas y computación, crea una evolución en las habilidades de análisis y desarrollo de forma personal, y genera un tipo comprensión del flujo

de información a escala global. Cambiando así el aspecto general del cómo se ven y relacionan las cosas en el entorno en el cual vivimos.

FUENTES:

(INTECO), I. N. de T. de la C. (2009). Estudio sobre la privacidad de los datos personales y la seguridad de la información en las redes sociales online. Retrieved from http://www.uv.es/limprot/boletin9/inteco.pdf

Agarwal, R., & Umphress, D. (2008). Extreme programming for a single person team. Proceedings of the 46th Annual Southeast Regional Conference on XX - ACM-SE 46, (March), 82. https://doi.org/10.1145/1593105.1593127

Bustamante, D., & Rodríguez, J. C. (2014). Metodología Actual Metodologia XP. Retrieved from http://blogs.unellez.edu.ve/dsilva/files/2014/07/Metodologia-XP.pdf

Clases, D. De, Objetos, D. De, Estados, D. De, Secuencias, D. De, Actividades, D. De, Colaboraciones, D. De, & Componentes, D. De. (2001). Diagramas del UML, 1–23.

Eisenberg, T., Bigrigg, M. W., Kathleen, P., Kunkel, F., Chieffallo, D., & Diesner, J. (2010). AutoMap : Office.

Facchín, J. (2016). Las Redes Sociales más importantes del Mundo “Lista 2016.” Retrieved April 20, 2017, from http://josefacchin.com/2013/03/15/las-redes-sociales-mas-populares-del-planeta/

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

García García, P., & Rodríguez, C. A. (2017). Minería de Datos aplicada a las Redes Sociales. Retrieved from http://www.it.uc3m.es/jvillena/irc/practicas/08-09/08.pdf

Gartner Inc. (2007). “Dirty datos” es un problema de negocios, no un problema de TI, según Gartner. Retrieved April 3, 2017, from http://www.gartner.com/newsroom/id/501733

Gartner Inc. (2013). What Is Big Data? - Gartner IT Glossary - Big Data. Retrieved May 7, 2017, from http://www.gartner.com/it-glossary/big-data%5Cnhttp://www.gartner.com/it-glossary/big-data/

git. (2015). Una breve historia de Git. Retrieved May 23, 2017, from https://git-scm.com/book/es/v1/Empezando-Una-breve-historia-de-Git

Glez-Peña, D., Lourenzo, A., López-Fernández, H., Reboiro-Jato, M., & Fdez-Riverola, F. (2013). Web scraping technologies in an API world. Briefings in Bioinformatics, 15. https://doi.org/10.1093/bib/bbt026

Hootsuite. (2017). Hootsuite (Home Page). Retrieved March 16, 2017, from https://hootsuite.com/es/

IA, D. C. de la C. e. (2014). Pruebasunitarias. Retrieved May 2, 2017, from http://www.jtech.ua.es/j2ee/publico/lja-2012-13/sesion04-apuntes.html

IBM Bluemix. (2015). Insights for Twitter. Retrieved May 28, 2017, from https://console.ng.bluemix.net/catalog/ibm-insights-for-twitter/

IETF. (1999). RFC 2616 - Hypertext Transfer Protocol -- HTTP_1. Retrieved May 15, 2017, from https://tools.ietf.org/html/rfc2616

Issi, G. (2003). Metodologías Ágiles en el Desarrollo de Software. Retrieved from http://issi.dsic.upv.es/archives/f-1069167248521/actas.pdf

Liferay. (2017). Liferay Portal Feature Overview | Liferay. Retrieved April 4, 2017, from https://web.liferay.com/es/products/liferay-portal/features/portal

Logicalis. (2015). Redes sociales como fuentes de datos_ el caso de Twitter. Retrieved May 27, 2017, from https://www.marketingdirecto.com/digital-general/social-media-marketing/breve-historia-de-las-redes-sociales

Lotfy, A. E., Saleh, A. I., El-Ghareeb, H. A., & Ali, H. A. (2016). A middle layer solution to support ACID properties for NoSQL databases. Journal of King Saud University - Computer and Information Sciences, 28, 133–145. https://doi.org/10.1016/j.jksuci.2015.05.003

Marketing Directo. (2011). Breve historia de las redes sociales - Marketing Directo. Retrieved May 27, 2017, from http://www.marketingdirecto.com/actualidad/social-media-marketing/breve-historia-de-las-redes-sociales/

Marset, R. N. (2007). REST vs Web Services. Retrieved from http://users.dsic.upv.es/~rnavarro/NewWeb/docs/RestVsWebServices.pdf

Matsuo, Y., Mori, J., Hamasaki, M., Nishimura, T., Takeda, H., Hasida, K., & Ishizuka, M. (2007). POLYPHONET: An advanced social network extraction system from the Web. Web Semantics. https://doi.org/10.1016/j.websem.2007.09.002

NTP-ISO/IEC 12207. (2006). isoiec12207[7]. Lima, Perú. Retrieved from http://www.senasa.gob.pe/senasa/wp-content/uploads/2014/11/Certificacion-citricos-a-mexico_26_mayo_2105_2.pdf

Olston, C., & Najork, M. (2010). Web Crawling. Foundations and Trends R in Information Retrieval, 4(3), 175–246. https://doi.org/10.1561/1500000017

Rojas, D., & Platzi. (2016). ¿Cuál es la diferencia entre Big Data y Business Intelligence? Retrieved April 20, 2016, from https://platzi.com/blog/diferencia-big-data-business-intelligence/

Rubira, J. (2011). Twitter4j, integración de tu aplicación Java con Twitter. Retrieved March 8, 2017, from https://www.genbetadev.com/frameworks/twitter4j-integracion-de-tu-aplicacion-java-con-twitter

SAIMA Solutions. (2013). Business Intelligence y Big Data, ¿independencia o cooperación? | Blog de Saima Solutions. Retrieved from http://www.saimasolutions.com/blog/business-intelligence-big-data/

Schroeck, Michael; Shockley, Rebecca; Smart, J. (2012). Analytics: el uso de big data en el mundo real. IBM. Informe Ejecutivo, 22. https://doi.org/10.1007/978-1-84996-226-1

Srivastava, J., Desikan, P., & Kumar, V. (2006). Web Mining — Concepts, Applications, and Research Directions. Retrieved from http://dmr.cs.umn.edu/Papers/P2004_4.pdf

Techopedia. (2016). What is Data Retrieval? - Definition from Techopedia. Retrieved March 15, 2017, from https://www.techopedia.com/definition/26464/data-security

Vásquez Vélez, M. (2012). EL HABEAS DATA EN LAS REDES SOCIALES. Retrieved from http://bdigital.ces.edu.co:8080/repositorio/bitstream/10946/1281/2/Habeas data.pdf

Web, S. (2011). Guía Breve de Servicios Web. Retrieved March 16, 2017, from http://www.w3c.es/Divulgacion/GuiasBreves/ServiciosWeb

Weigend, A. (2014). Sabías que es un Data Scientist? Retrieved May 5, 2017, from http://sabiasqueestadistica.blogspot.com.co/2014/03/sabias-que-es-un-data-scientist_3.html

Wood, D. (2010). Linking enterprise data. Linking Enterprise Data, 1–291. https://doi.org/10.1007/978-1-4419-7665-9

Wood, R., Zheludev, I., & Treleaven, P. (2012). Mining Social Data with UCL's SocialSTORM Platform. London. Retrieved from http://weblidi.info.unlp.edu.ar/worldcomp2012-mirror/p2012/DMI9011.pdf

LISTA DE ANEXOS:

- Anexo A: Especificación de requerimientos de software (ERS)