|  |  |  |  |
| --- | --- | --- | --- |
|  | | **FACULTAD DE INGENIERÍA**  **MAESTRÍA EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN** | |
| **TRABAJO DE GRADO – PROPUESTA DE PROYECTO – PZ-2011-1-XX** | | | |
| **TÍTULO DEL PROYECTO** | **Framework de pre-procesamiento de datos en minería de texto basado en tecnologías de Big Data** | | |
| **DATOS DEL ESTUDIANTE** | **Wilson Alzate Calderón** | **CORREO**  **ELECTRÓNICO** | [walzate@javeriana.edu.co](mailto:juanpp@javeriana.edu.co) |
| CC 1.032.400.001 | [wilson.alzate@gmail.com](mailto:juanito@gmail.com) |
| **DIRECTOR DE TRABAJO DE GRADO**  **ASESOR (opcional)** | Ing. Alexandra Pomares PhD | **MODALIDAD** | Profundización |
| [pomares@javeriana.edu.co](mailto:egon@javeriana.edu.co) | **ÁREA DE ÉNFASIS** | Sistemas Información e Ing. Software |
|  | **GRUPO Y LÍNEA DE INVESTIGACIÓN** | ISTAR – Educación |
|  | Sub-línea - Gestión de Información |

|  |  |
| --- | --- |
| **OBJETIVOS** | **Objetivo General**  Generar un *framework* que ofrezca funcionalidades pre-construidas para ejecutar tareas de pre-procesamiento en minería de texto basado en tecnologías de Big Data.  **Objetivos Específicos**   1. Seleccionar las técnicas, métodos y herramientas asociados a Big Data que sean relevantes para realizar tareas de pre-procesamiento en minería de texto. 2. Diseñar la arquitectura de un *framework* que ofrezca funcionalidades pre-construidas para ejecutar tareas de pre-procesamiento en minería de texto basado en las tecnologías de Big Data seleccionadas. 3. Construir un prototipo del *framework* que implemente la arquitectura diseñada. 4. Validar las funcionalidades provistas por el *framework* a través de su utilización en un caso de estudio aplicado al análisis de historias clínicas electrónicas del *Hospital Universitario San Ignacio.* |

|  |  |
| --- | --- |
| **PROBLEMA**  **DE**  **INVESTIGACIÓN**  **O**  **APLICACIÓN** | En la actualidad, el monto de información almacenada en el mundo se estima alrededor de 1.200 exabytes(1018 bytes)[1], los cuales son producidos por múltiples fuentes, por ejemplo, Google procesa más de 24 petabytes(1015 bytes) de datos por día y Facebook obtiene más de 10 millones de fotos cada hora[1]. En el mundo de los negocios el 80% de los datos existen en un formato no estructurado [4] y si este monto fuera solamente libros impresos, ellos cubrirían la superficie entera de Estados Unidos en 52 capas de grueso [1].  Teniendo en cuenta estas grandes cantidades de datos que se están generando actualmente, surge el paradigma de Big Data, que es la habilidad de la sociedad de aprovechar la información de nuevas formas para producir bienes y servicios que generen valor. Siendo además la manera de aplicar matemáticas a grandes cantidades de datos con el fin de inferir hallazgos [1]. Un problema de Big Data es aquel que cumple con las siguientes tres características [5]:   * Volumen (Escala de los datos): Big Data trabaja con grandes cantidades de datos, lo que en la actualidad es algo frecuente teniendo en cuenta que por ejemplo para 2012 cerca de 2.5 exabytes de datos fueron creados cada día y ese número se está duplicando cada 40 meses [5]. * Variedad (Diferentes formas de datos): Existen diferentes formas de datos, tanto estructurados como no estructurados, dentro de los que se encuentran: mensajes, actualizaciones, imágenes en redes sociales, lecturas de sensores, señales de GPS desde teléfonos celulares, compras en línea, entre otros [5]. * Velocidad (Análisis de flujo de datos): Para muchas aplicaciones, la velocidad de creación de datos es incluso más importante que el volumen. La Información en tiempo real, o en tiempos cercanos a ello, hace posible para una compañía ser mucho más ágil que sus competidores y poder tomar decisiones de negocio mucho más informadas y precisas [5].   El cuidado de la salud, en común con muchas otras industrias, está generando grandes cantidades de datos rutinarios, los cuales pueden ser minados e incluso combinados con tweets y blogs. Es un gran reto procesar, analizar y conservar dicha masa de datos. Darle sentido a esa gran cantidad de información ofrece oportunidades para el mejor tratamiento de una enfermedad, abordar temas de salud pública o para el funcionamiento eficiente de los proveedores de servicios de salud [6]. El análisis de datos contenidos en las historias clínicas electrónicas usando técnicas computacionales es un problema de Big Data[4] debido al gran volumen de información contenida, a la variedad entre datos estructurados y no estructurados, así como también, la velocidad con la que se requieren ciertos análisis.  Las técnicas de minería de texto son las más indicadas para el análisis del contenido narrativo de las historias clínicas electrónicas. Algunas de las tareas relacionadas a la minería de texto son el pre-procesamiento de colecciones de documentos (categorización de texto, extracción de información, extracción de términos), el almacenamiento de representaciones intermedias, técnicas para analizar dichas representaciones intermedias (como análisis de distribución, *clustering*, análisis de tendencias y reglas de asociación), así como la visualización de los resultados [11].    Figura 1 Arquitectura de alto nivel de un sistema de minería de texto[11]  Las operaciones de pre-procesamiento se centran en la identificación y extracción de características representativas para documentos en lenguaje natural y son responsables de la transformación de datos no estructurados almacenados en colecciones de documentos a un formato intermedio explícitamente estructurado [11]. Siendo tanto las tareas de pre-procesamiento como las del núcleo, las dos áreas más críticas para cualquier sistema de minería[11] por lo que se encuentra allí una gran oportunidad de optimización de procesos utilizando tecnologías de Big Data que permitan agilizar los tiempos de respuesta de los sistemas de minería.  Por otra parte, la atención de un paciente en una institución de prestación de servicios de salud puede generar en un día más de 100 registros nuevos asociados a la información de medicamentos, procedimientos y observaciones. Hecho que si se considera a largo plazo, en un hospital de primer nivel, puede generar millones de registros no estructurados (observaciones y procedimientos en texto narrativo) que se deben procesar. Dichas cantidades de información son un activo importante para el análisis de tendencias [6], pero con las tecnologías tradicionales, los procesos de extracción de conocimiento o de clasificación pueden tardar demasiado tiempo, lo que va en contraposición de lo que el sector necesita; que la velocidad de respuesta sea lo suficientemente rápida para lograr obtener la información correcta en el momento preciso [7]. Además de ello aunque existen en el mercado algunas tecnologías asociadas al paradigma de Big Data, como Hadoop [8], Cassandra, HBase, Hive, NoSQL [7] o Cloud Computing [9], con las cuales se pueden llegar a abordar las complejidades de Volumen, Variedad y Velocidad del análisis de historias clínicas electrónicas, su orquestación y gestión puede llegar a ser un tanto complejas incluso para un profesional con dominio técnico. Por otro lado, Se han realizado aproximaciones en el uso de tecnologías de Big Data para trabajar procesos *core* de minería de datos, tanto usando los tweets públicos [12], como noticias o patentes [13], pero no en análisis de historias clínicas electrónicas.  Teniendo en cuenta la problemática antes descrita se llega a la pregunta de ¿Cómo generar un framework que ofrezca funcionalidades pre-construidas para ejecutar tareas de pre-procesamiento en minería de texto basado en tecnologías de Big Data, en donde se logre adicionar un nivel de abstracción a la complejidad que implica tener que enlazar las diferentes tecnologías disponibles y así poder mejorar los tiempos de procesamiento en sistemas de minería de texto?, para lo cual en este proyecto se propone generar un modelo de aplicación de Big Data que soporte la fase de pre-procesamiento en los proyectos de minería de texto a través de un *framework*, el cuál será validado en un caso de estudio basado en análisis de historias clínicas electrónicas, específicamente en el análisis de notas de ingreso, notas de enfermería y plan de tratamiento. |

|  |  |
| --- | --- |
| **METODOLOGÍA** | |
| **DESCRIPCIÓN GENERAL** | La metodología marco que se seguirá en el presente proyecto se basa en el paradigma de la ciencia basada en el diseño, la cual permite abordar problemas de investigación en informática y ciencias de la computación [10], cuyo objetivo es contribuir con soluciones novedosas a problemas que aunque son relevantes, no han sido resueltos.  La ciencia basada en el diseño establece un proceso iterativo, de constante retroalimentación y mejoramiento que se compone de tres ciclos estrechamente relacionados:   1. Rigor: En este ciclo se desarrolla una revisión de los principios científicos, teorías y experiencias de investigaciones anteriores consignadas en bases de datos especializadas. Para el contexto del proyecto, se verificarán las teorías, técnicas, tecnologías y algoritmos relacionados con el pre-procesamiento en minería de texto, luego se revisarán las técnicas y tecnologías de Big Data que puedan ayudar a mejorar los tiempos de respuesta. El objetivo de esta fase es identificar cuáles funcionalidades debe proveer el *framework* para soportar la ejecución de tareas de pre-procesamiento de minería de texto. 2. Relevancia: Se continúa con un análisis del contexto del problema, en este caso, de las necesidades de la comunidad de minería de texto en cuanto a velocidades de pre-procesamiento, así como también de la organización y del entorno tecnológico del Hospital Universitario San Ignacio alrededor del análisis de historia clínica electrónica. Gracias a esta fase será posible identificar los insumos que permitirán validar el *framework*. Estos insumos incluyen las fuentes de datos de las que provienen los textos y las características de los textos. 3. Diseño central: Es la construcción y evaluación de teorías o artefactos, teniendo como materia prima tanto los resultados de las fases de rigor como los de relevancia, esto con el fin de generar una contribución a la base de conocimiento y también una aplicación que sea relevante en el contexto del problema.   A continuación se presenta un diagrama de la adaptación del paradigma de la ciencia del diseño aplicada al proyecto (Figura 1).  Metodología (1).png  Figura 2 Fases de la ciencia del diseño  Las fases de la ciencia del diseño para el proyecto son:  **Fase 1: Análisis de la base de conocimiento (Rigor):** Inicialmente se realiza una revisión de los métodos y herramientas, basados en teorías, artículos, libros y tecnologías tanto de pre-procesamiento en minería de texto como de Big Data con el fin de identificar las funcionalidades que debe proveer el *framework*.  **Fase 2: Análisis del entorno (Relevancia):** Se evalúan las necesidades de la comunidad de minería de texto en cuanto a pre-procesamiento de grandes cantidades de contenido narrativo. Se identifican además los insumos y el entorno tecnológico del caso de estudio (Hospital Universitario San Ignacio) con el cual se validará el framework.  **Fase 3: Diseño, construcción y validación del framework (Diseño central):** Esta fase se centrará en el diseño de la arquitectura y la construcción de un prototipo del *framework* que ofrezca funcionalidades generales y pre-construidas para ejecutar tareas de pre-procesamiento en minería de texto basado en tecnologías de Big Data, tomando los resultados tanto de las fases de rigor como de relevancia, para luego validar el prototipo en un caso de estudio aplicado al Hospital Universitario San Ignacio.  Vale la pena indicar que ciertas actividades de las Fase 1 y 2 NO son secuenciales y se convierten en un proceso iterativo en el cual se refinan unas actividades de la Fase 1 con otras de la Fase 2. |
| **FASE 1**  **(Rigor)**  **Análisis de la base de conocimiento** | El análisis de la base de conocimiento es la que permite la obtención de los fundamentos teóricos y metodológicos generales mediante los cuales se guiará el proyecto, brindando las diferentes teorías, técnicas, tecnologías y conceptos que se usarán en la fase de diseño central. En el marco de este proyecto la base de conocimiento se compondrá de fundamentos de minería de texto (específicamente en técnicas de pre-procesamiento), así como también técnicas y tecnologías de Big Data. Para el diseño de la arquitectura se tomará de la base de conocimiento aquellos meta-modelos arquitectónicos que puedan ser relevantes.  La construcción de la base del conocimiento está asociada con la problemática existente y a la inexistencia de un método que permita la integración de técnicas de pre-procesamiento de minería de texto con tecnologías de Big Data que permitan mejorar los tiempos de respuesta de los sistemas minería de texto. Esta fase se relaciona con el primer objetivo.  Para el desarrollo de esta fase se realizarán las siguientes actividades:   1. Identificar, clasificar y analizar la bibliografía relacionada con técnicas, modelos, métodos, herramientas y algoritmos relacionados con pre-procesamiento de minería de texto y Big Data. 2. Especificar y definir las técnicas y algoritmos de pre-procesamiento en minería de texto que se van a aplicar para el análisis de texto narrativo. 3. Especificar y definir las técnicas y tecnologías de Big Data que se usarán para mejorar los tiempos de respuesta en el pre-procesamiento de texto narrativo. 4. Realizar una prueba de concepto de las técnicas y herramientas usadas en pre-procesamiento de minería de texto teniendo en cuenta texto narrativo. 5. Realizar una prueba de concepto de las técnicas y tecnologías de Big Data que puedan ser usadas para el mejoramiento de los tiempos de respuesta en las tareas de pre-procesamiento en minería de texto. 6. Escritura de un informe con el reporte de pruebas de concepto de las técnicas, algoritmos y herramientas usadas. |
| **FASE 2**  **(Relevancia)**  **Análisis del entorno** | En el desarrollo de esta investigación en la fase de análisis del entorno se identifican las necesidades de la comunidad de minería de texto en cuanto a las velocidades de pre-procesamiento y se reconocen los insumos y el entorno tecnológico para la posterior validación del framework.  Esta fase se relaciona directamente con el cumplimiento del segundo objetivo específico, el cual se logra a partir de la realización de las siguientes actividades:   1. Especificar los requerimientos funcionales y no funcionales usando los conceptos obtenidos de la fase de rigor y realizando entrevistas estructuradas a miembros de la comunidad de minería de texto a los que se pueda tener acceso directo e interacción a través de foros virtuales de las herramientas de minería. 2. Validación y documentación de los requerimientos teniendo en cuenta las necesidades generales de la fase de rigor y las particulares de los *stakeholders*. 3. Identificar la infraestructura (hardware y software) de sistemas de información con los que se va a integrar o van a ser fuente de datos en el caso de estudio. |
| **FASE 3**  **(Diseño central)**  **Diseño, construcción y validación del framework** | En esta fase se desarrollará el *framework* con base en los resultados de las dos fases anteriores, enmarcados por la aplicabilidad y uso de los fundamentos y metodologías con las que se cuentan. Para la construcción del *framework* se aplicará la metodología Personal Xtreme Programming (PXP)[14], contando con un proceso iterativo en donde inicialmente los requerimientos obtenidos se traducirán en historias de usuario, las cuales serán priorizadas, luego se escribirán las pruebas unitarias y se escribirá el código fuente, se realizará el refactor necesario y se escribirán pruebas unitarias, manteniendo tres líneas base de código (Desarrollo, Refactor y Producción).  Esta fase tiene estrecha relación con el cumplimiento del tercer Objetivo específico. La validación de los procesos de estructuración y extracción a través del *framework* se realiza mediante el cumplimiento del cuarto objetivo específico. Para la validación del *framework* se analizan las tecnologías, aplicaciones y el sistema SAHI (Sistema de información médico que contiene las historias clínicas electrónicas del Hospital Universitario San Ignacio). La extracción, análisis y preparación de los datos se realizará a partir del texto narrativo de las historias clínicas electrónicas que se encuentran en el sistema de información SAHI.  Para el desarrollo de esta fase se realizarán las siguientes actividades:   1. Describir y caracterizar el diseño del framework. 2. Desarrollar el *framework* haciendo uso de la metodología PXP. 3. Definir y ejecutar el protocolo de pruebas. 4. Validar la funcionalidad del *framework* desarrollado a través de su aplicación en el caso de estudio seleccionado. 5. Documentar los resultados finales del proyecto. |

|  |  |
| --- | --- |
| **RESULTADOS ESPERADOS** | |
| **ASIGNATURA MISyC PROYECTO 1** | Caracterización y análisis de las técnicas, modelos, métodos, tecnologías, herramientas y algoritmos relacionados con pre-procesamiento de minería de texto y Big Data. (Entregable fase 1 en cumplimiento del objetivo 1). |
| Documento con los requerimientos funcionales y no funcionales recopilados (Entregable fase 2 en cumplimiento del objetivo 1). |
| Diseño preliminar del *framework* (Entregable fase 3 en cumplimiento del objetivo 2). |
| **ASIGNATURA MISyC PROYECTO 2** | Diseño final del *framework* (Entregable fase 3 en cumplimiento del objetivo 2). |
| Desarrollo del *framework*. Además de los programas fuentes y ejecutables, se entregará la documentación técnica y el manual de instalación y uso (Entregable fase 3 en cumplimiento del objetivo 3). |
| Validación de la funcionalidad del *framework* con el caso de estudio previamente establecido (Entregable fase 3 en cumplimiento del objetivo 4). |
| Escritura y sometimiento de un artículo de investigación que incluye la validación y resultados del framework. Este artículo será presentado para publicación en una revista indexada (Entregable fase 3 en cumplimiento del objetivo 4). |
| Memorias del proyecto de grado. |

|  |  |
| --- | --- |
| **PROSPECTIVA DE INNOVACIÓN** | |
| **POTENCIAL DE INNOVACIÓN** | El potencial de innovación del presente trabajo de grado se encuentra centrado principalmente en dos frentes:   * El uso novedoso de tecnologías de Big Data para el mejoramiento de los tiempos de respuesta de las aplicaciones de minería de texto, específicamente en las tareas de pre-procesamiento, al suministrar una serie de funcionalidades pre-construidas que puedan ser usadas por otros sistemas. * Incrementar el nivel de abstracción para disminuir la complejidad que implica tener que enlazar las diferentes tecnologías disponibles tanto de minería de texto como de Big Data.   La literatura analizada permitió establecer que existen trabajos relacionados a este tipo de problemática; sin embargo, no existen estudios sobre el uso de tecnologías de minería de texto conjugadas con Big Data para el análisis de texto narrativo. Por tal razón se aprecia una gran oportunidad para abrir campos y oportunidades de desarrollo e investigación en la comunidad académica y científica. |
| **PROPIEDAD INTELECTUAL** | Este trabajo de grado se realizará dentro del marco de investigación del grupo de investigación ISTAR de la Pontificia Universidad Javeriana de Bogotá, se contará con el apoyo del Hospital Universitario San Ignacio que proveerá toda la información concerniente a las historias clínicas electrónicas. El resultado conceptual, modelos, código fuente y documentación del proyecto podrán ser empleados por otros proyectos asociados al departamento de Ingeniería de Sistemas.  El resultado del proyecto aportará al marco global del desarrollo del grupo ISTAR, servirá como base para futuros trabajos y proyectos del grupo, y estará a disposición del grupo, sin ninguna restricción de uso, ya sea éste de carácter académico o empresarial. |

|  |
| --- |
| **BIBLIOGRAFÍA** |
| [1]V. Mayer-Schönberger and K. Cukier, Big data: a revolution that will transform how we live, work, and think. Boston: Houghton Mifflin Harcourt, 2013.  [2]N. V. Chawla and D. A. Davis, “Bringing Big Data to Personalized Healthcare: A Patient-Centered Framework,” Journal of General Internal Medicine, vol. 28, no. S3, pp. 660–665, Jun. 2013.  [3]T. O’Reilly, M. Loukides, J. Steele, and C. Hill, How Data Science Is Transforming Health Care, 1st ed. O’Reilly Media, 2012.  [4]T. B. Murdoch, “The Inevitable Application of Big Data to Health Care,” JAMA, vol. 309, no. 13, p. 1351, Apr. 2013.  [5] Andrew McAfee and Erik Brynjolfsson. Big data: The management revolution. Harvard business review, 90(10):p60{68.  [6] Harshana Liyanage, Siaw-Teng Liaw, and Simon de Lusignan. Accelerating the development of an information ecosystem in health care, by stimulating the growth of safe intermediate processing of health information (IPHI). Informatics in primary care, 20(2):81{86, 2012. PMID:23710772.  [7] IBM. (2012, June 18). ¿Qué es Big Data? CT316. Retrieved February 28, 2014, from http://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/  [8] McAfee, A., & Brynjolfsson, E. (n.d.). Big Data: The Management Revolution. Harvard Business Review, 90(10), p60–68.  [9] Purkayastha, S., & Braa, J. (2013). Big Data Analytics for developing countries – Using the Cloud for Operational BI in Health. The Electronic Journal of Information Systems in Developing Countries, 59(0). Retrieved from https://www.ejisdc.org/ojs2/index.php/ejisdc/article/view/1220  [10] Hevner, A., March, S., Park, J., Ram, S. MIS Quarterly Design Science in Information Systems Research, pp. 75-105, 2004.  [11] Feldman, R., & Sanger, J. (2013). The Text Mining Handbook (1 edition.). Cambridge University Press.  [12] Das, T. K., & Mohan Kumar, P. (n.d.). Big data analytics: A framework for unstructured data analysis. School of Information Technology and Engineering, VIT University.  [13] Tablan, V., Roberts, I., Cunningham, H., & Bontcheva, K. (2012). GATECloud.net: a platform for large-scale, open-source text processing on the cloud. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 371(1983), 20120071–20120071. doi:10.1098/rsta.2012.0071  [14] Agarwal, R., & Umphress, D. (2008). Extreme programming for a single person team (p. 82). ACM Press. doi:10.1145/1593105.1593127 |