Humanidades, Tecnología y Ciencia, del Instituto Politécnico Nacional
Certificado de Reserva de Derechos al Uso Exclusivo del Título: 04-2010-0326124-14000-203 ISSN: 2007-1957

IDENTIFICACIÓN MEDIANTE MINERÍA DE TEXTOS DE LA COMPRENSIÓN DEL SIGNIFICADO DE LA CIENCIA DE DATOS QUE TIENEN LOS ESTUDIANTES

Fabiola Ocampo Botello
Instituto Politécnico Nacional. Escuela Superior de Cómputo
focampob@ipn.mx

Enrique Ramírez Gutiérrez
Instituto Politécnico Nacional. Escuela Superior de Cómputo
eramirezg1504@alumno.ipn.mx

Yanina De Luna Ocampo
Instituto Politécnico Nacional. Escuela Superior de Cómputo
ydelunao1600@alumno.ipn.mx

Resumen

Mediante técnicas de minería de textos se identificaron los aspectos clave de la percepción del significado de la Ciencia de datos de diez estudiantes de nuevo ingreso del plan de estudios de Licenciatura en Ciencia de datos de la Escuela Superior de Cómputo (ESCOM) del Instituto Politécnico Nacional (IPN). Aplicando la normalización del corpus del conocimiento mediate la frecuencia relativa del término (tf) se descubrieron cuatro categorías: 1) Lo que entendieron del significado de la ciencia de datos, 2) Los sectores potenciales de su utilidad actual o proyección futura, 3) La utilidad: valor, decisiones, desarrollo, uso, productos y servicios y 4) Los retos que enfrenta; resultados coincidentes con los materiales bibliográficos utilizados en las actividades académicas realizadas por los aprendices.

Palabras clave: Ciencia de datos, minería de datos, minería de textos, frecuencia de término.

Humanidades, Tecnología y Ciencia, del Instituto Politécnico Nacional
Certificado de Reserva de Derechos al Uso Exclusivo del Título: 04-2010-0326124-14000-203 ISSN: 2007-1957

La minería de textos es un área de estudio que se fundamenta en las tareas de análisis de la minería de datos, la cual permite descubrir conocimiento oculto en los datos mediante la generalización, patrones de comportamiento o relación entre variables que no son visibles a simple vista.

Debido a las diversas actividades que desarrollan los estudiantes en su formación académica, en este artículo se presenta el análisis de la comprensión que tienen los estudiantes de nuevo ingreso sobre lo que es la ciencia de datos, lo anterior a partir de diversas actividades académicas como: organizadores gráficos, lectura de textos, exposiciones, todas estas guiadas por la profesora encargada de impartir el curso.

Los resultados encontrados revelaron los términos claves propios de la ciencia de datos, los cuales fueron agrupados en cuatro categorías, coincidentes con el marco teórico estudiado.

Marco teórico

Joyanes Aguilar (2019) define la ciencia de datos como: "una disciplina que se encarga de la extracción de conocimiento a partir de datos y que se encuentra en plena expansión" (p. 418). El mismo autor, señala que la ciencia de datos es el proceso de extraer conocimiento oculto a partir de cantidades masivas de datos estructurados y no estructurados mediante métodos de estadística, aprendizaje automático, minería de datos y analítica predictiva.

Diversos autores expresan que las áreas de conocimiento que incorpora la Ciencia de datos son: el Big Data para procesar datos, la Minería de datos para analizar e identificar relaciones ocultas, patrones y tendencias, y la visualización de datos para explicar y socializar mejor la información obtenida (Moreno Salinas, 2017), mientras que Joyanes Aguilar (2019) establece que es una ciencia interdisciplinaria que está cambiando el modo en que las organizaciones resuelven problemas y ganan ventaja competitiva, y que se concentra en las tres grandes disciplinas: Ciencias de la computación, matemáticas/estadística y dominio del conocimiento.

La minería de datos, como se ha expresado, es un área fundamental en la ciencia de datos, debido a que incorpora diversas tareas para el análisis de datos, ésta ha sido definida como el proceso de extraer información y conocimiento implícitos, potencialmente útiles y que son desconocidos por las personas, los cuales se encuentran en datos masivos, incompletos, difusos y aleatorios (Sahu, Shrma & Gondhalakar, 2011) e incorpora diversas tareas, las cuales tienen sus propios requisitos y la información que se obtiene difiere mucho de la obtenida por otra tarea, ejemplos de estas son: la clasificación, la regresión, el agrupamiento, las correlaciones, las reglas de asociación (Hernández, Ramírez y Ferri, 2004).

Considerando el estudio que se realizó en este proyecto, cuyo objetivo fue analizar la comprensión que tuvieron los alumnos sobre lo qué es la ciencia de datos, lo anterior, tomando en cuenta las lecturas realizadas, la creación de organizadores gráficos y exposiciones por equipo hechas por los alumnos y guiadas por la profesora, se utilizó la minería de textos debido a la forma en que se solicitó expresaran por escrito la comprensión del tema, lo cual se realizó mediante la siguiente pregunta abierta: Explique: ¿Qué es la ciencia de datos?

La extracción de la información de los datos contenida en los textos se realiza

Humanidades, Tecnología y Ciencia, del Instituto Politécnico Nacional
Certificado de Reserva de Derechos al Uso Exclusivo del Título: 04-2010-0326124-14000-203 ISSN: 2007-1957

mediante el área de estudio conocida como minería de textos.

Se utilizó la minería de textos debido a que según Joyanes Aguilar (2019) la minería de textos tiene el mismo propósito de la minería de datos, pero la entrada es una colección de archivos no estructurados o semi estructurados, tales como documentos Word, archivos PDF, resúmenes de textos, archivos XML, por mencionar. La minería de textos empieza con la recolección y almacenamiento de fuentes de datos, los procesa y analiza para extraer información relevante y conocimiento de los datos basados en textos a través de técnicas y herramientas de minería de datos.

Según Sarkar (2016) la minería de textos es una disciplina que tiene como objetivo principal extraer la información significativa de textos, lo cual conlleva de forma implícita el análisis de estos. El análisis de texto se refiere al proceso mediante el cual se extrae información relevante y procesable a partir de datos textuales, el cual es un campo interdisciplinario que combina técnicas de procesamiento del lenguaje natural, inteligencia artificial y minería de datos.

Método

La forma de trabajo que se realizó con los alumnos consistió en una intervención didáctica, en la cual se realizaron diversas actividades como: lectura de artículos de lo que es la ciencia de datos, las áreas de investigación que incorpora, las aplicaciones y fines que persigue en ellas, con lo anterior, los discentes realizaron organizadores gráficos y exposiciones en equipo guiadas por la profesora del grupo. Al finalizar la intervención, se solicitó a los alumnos que respondieran un cuestionario de preguntas abiertas, en donde una de ellas fue: Explique: ¿Qué es la ciencia de datos?

En este estudio se aplicó una técnica no probabilística para la elección de los participantes llamada dependencia de sujetos disponible (Babbie, 1988). La participación fue voluntaria y respondieron diez estudiantes.

Para el tratamiento y análisis de datos se utilizó la Plataforma Analítica de datos KNIME. El flujo de trabajo desarrollado se presenta en la figura número 1.

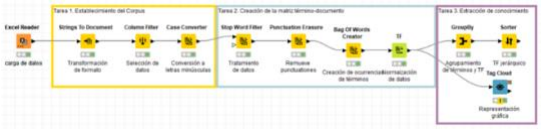

Figura 1. Flujo de trabajo desarrollado en la Plataforma Analítica KNIME.

El método para el análisis de textos que se aplicó en el desarrollo de este proyecto fue el propuesto por Sharda, Denle & Turban (2014), el cual se divide en tres tareas consecutivas, con sus entradas de datos y respectivas salidas, las cuales se describe a continuación.

Tarea 1. Establecimiento del corpus

El objetivo principal de la actividad de la primera tarea es recopilar todos los documentos relacionados con el contexto (dominio de interés) que se va a estudiar, se transforman y organizan de tal manera que todos tienen la misma forma de representación (p. 308).

En este estudio se concentraron todas las respuestas de los estudiantes en un solo archivo de tipo Documento para ser procesado por la herramienta de análisis y se eligieron los textos de la columna que contenía la pregunta a analizar.

Humanidades, Tecnología y Ciencia, del Instituto Politécnico Nacional
Certificado de Reserva de Derechos al Uso Exclusivo del Título: 04-2010-0326124-14000-203 ISSN: 2007-1957

Tarea 2: Creación de la Matriz Término-Documento

En esta tarea, los documentos digitalizados y organizados (el corpus) se utilizan para crear la matriz término-documento (*Term–Document Matrix, TDM*). En el TDM, las filas representan los documentos y las columnas representan los términos. Las relaciones entre los términos y los documentos se caracterizan por índices, que puede ser el número de ocurrencias del término en los respectivos documentos. El objetivo en esta tarea es convertir la lista de documentos organizados (el corpus) en un TDM, en donde se excluyen artículos, verbos auxiliares y términos utilizados en casi todos los documentos del corpus. Para tener un TDM más consistente para un análisis posterior, estos índices brutos deben normalizarse, con la intención de mostrar los recuentos de frecuencia reales (p. 309-310).

El tratamiento del corpus que se hizo en esta etapa consistió en la conversión a letras minúsculas para uniformizar los datos, se eliminaron las palabras vacías, es decir aquellas que no aportan ningún significado al análisis, por ejemplo, los conectores, los pronombres, etc. Un aspecto importante en esta tarea es la normalización de los datos, es decir, el cálculo de la frecuencia relativa del término (tf), este valor se calcula dividiendo la frecuencia absoluta de un término según un documento por el número de todos los términos de ese documento.

Tarea 3: Extracción del conocimiento

Una vez generado un TDM bien estructurado, se extraen patrones novedosos en el contexto del problema específico que se está abordando. Las categorías principales de los métodos de extracción de conocimiento son: la clasificación, el agrupamiento, la asociación y el análisis de tendencias (p. 312).

Considerando que el objetivo de este proyecto fue identificar las palabras claves para analizar la comprensión que tuvieron los alumnos de lo que es la ciencia de datos. En esta etapa se realizó un agrupamiento de términos para analizar aquellos con mayor frecuencia. Lo cual se realizó mediante una consulta de lenguaje de consulta estructurado.

Resultados

En la figura número 2, se presenta la expresión gráfica de la nube de etiquetas de los términos con mayor frecuencia escritos por los participantes, en la cual los términos con mayor frecuencia normalizada aparecen con letras de mayor tamaño.

Figura 2. Nube de etiquetas de los datos identificados.

Los resultados obtenidos respecto a las palabras clave que los estudiantes comprenden como ciencia de datos, se realizó considerando las acciones desarrolladas en la tarea número 2 del procesamiento de la minería de textos, referente al cálculo de la frecuencia relativa del término (tf), resultado de la división de la frecuencia absoluta de un término según un documento por el número de todos los términos de ese documento, se expresan en los siguientes párrafos:

Certificado de Reserva de Derechos al Uso Exclusivo del Título: 04-2010-0326124-14000-203 ISSN: 2007-1957

1) Lo que entendieron del significado de la ciencia de datos: como ciencia, datos, futuro, poder, cantidad, ayuda
2) Los sectores potenciales de su utilidad actual o proyección futura: marketing, empresas, mercados, sector, tecnología, sociedad, compañías, vida
3) La utilidad: valor, decisiones, desarrollo, uso, productos y servicios
4) Los retos que enfrenta: calidad, tratamiento, filtros, almacenamiento, diversificación, fuentes, problemas.

Considerando lo anterior, la comprensión general de los estudiantes respecto a lo qué es la ciencia de datos abarca los aspectos descritos en las definiciones anteriormente expuestas. Los discentes expresaron de forma escrita lo que comprendieron y mediante la minería de textos se logró identificar el uso predominante de las palabras clave, aquellas que obtuvieron un TF mayor, estas fueron: datos, cómputo, servicios, cantidad, uso y futuro. Por tal, se organizaron los términos en las categorías anteriormente descritas.

Conclusiones

Las diversas actividades académicas que desarrollan los estudiantes en el aprendizaje de un tema de estudio estimulan sus variadas formas de expresión tanto verbales como visuales, activas y reflexivas. Por ejemplo, a través de un organizador gráfico, establece, jerarquiza y articula los diversos conceptos, principios y relaciones que identifica en su proceso de comprensión. Otro recurso, es la expresión verbal escrita que los estudiantes realizan a través de respuestas a preguntas abiertas y que pueden ser analizadas mediante la minería de textos, la cual extrae conocimiento oculto inmerso en cantidades masivos de datos estructurados y no estructurados y con ello detectar los puntos clave que han entendido en el texto en cuestión.

Resulta de utilidad aplicar este tipo de métodos de análisis en la evaluación de las actividades que desarrollan los alumnos en sus quehaceres educativos para con ello estructurar y reestructurar el diseño de las secuencias didácticas para lograr mejores aprendizajes y la diversificación de tales actividades que favorezcan sus diversos estilos de aprendizaje.

Referencias

Babbie, R. E. (1988). *Métodos de investigación por encuesta.* Fondo de Cultura Económica. México.

Hernández O. J., Ramírez, Q. M. J. y Ferri, R. C. (2004). *Introducción a la Minería de datos*. Pearson.

Joyanes Aguilar, L. (2019). *Inteligencia de negocios y analítica de datos*. Una visión global de Business inteligence & Analytics. Alfaomega.

Moreno Salinas, J. G. (2017). *Científico de datos: codificando el valor oculto e intangible de los datos*. Revista Digital Universitaria. Vol. 18, Núm. 7, septiembre-octubre 2017.

Sharda, R., Denle, D., & Turban, E. (2014). *Business Intelligence and Analytics Systems for Decision Support*. TENTH EDITION. Pearson.

Sarkar, D. (2016). *Text Analytics with Python*. En Apress eBooks. https://doi.org/10.1007/978-1-4842-2388-8

Sahu, H., Shrma, S. & Gondhalakar, S. (2011). *A Brief Overview on Data Mining Survey*. International Journal of Computer Technology and Electronics Engineering (IJCTEE). Vol.1.