Análisis de ventajas y desventajas del proceso de minería de datos

La minería de datos es el conjunto de técnicas y herramientas que se emplean para la extracción de información relevante en grandes conjuntos de datos. Algunas de las ventajas que ofrece son:

Información de confianza. Una de las grandes ventajas de la minería de datos es que la información que se extrae con ella es totalmente fiable. Por eso, por ejemplo, se emplea en la investigación de mercados para ver que tipos de productos les interesan a los clientes.

Mejoras y ajustes en los procesos empresariales. La minería de datos tiene como ventaja la ayuda que proporciona para realizar ajustes operativos en las empresas. Esto sobresale, ante todo, en todo lo que se refiere a la mejora de los procesos logísticos.

- *Una mejor toma de decisiones*. Las decisiones basadas en datos siempre van a ser mejores. La minería de datos da información objetiva y fiable, por lo que las empresas y los analistas pueden tomar decisiones mucho mejores para el futuro y el negocio de la compañía.
- *Analiza grandes cantidades de datos rápidamente*. Gracias a la minería de datos se puede procesar una mayor cantidad de información en menos tiempo.
- *Predicciones*. Gracias a los datos extraídos se pueden realizar predicciones de comportamiento basadas en patrones. También es útil, en este sentido, para la creación de algoritmos para aprendizaje automático y el diseño de aplicaciones y programas específicos de IA

Ahora que ya conocemos las ventajas de la minería de datos, vamos a ver cuáles son sus desventajas. Y es que, aunque tenga muchas aplicaciones y potencial, no por ello es infalible o no tiene inconvenientes. Estas son algunas de las principales desventajas de la minería de datos:

- *Herramientas complejas*. La mayoría de las herrameintas que se emplean para minería de datos son complejas y requieren que las manejen profesionales formados y especializados. Es decir, se requiere capacitación y, en ocasiones, certificaciones específicas para poder manejarlas. Esto hace que los profesionales sean escasos y muy demandados.

- *No es infalible*. Aunque se trata de un conjunto de técnicas fiable, la minería de datos no es infalible y no siempre proporciona información totalmente precisa. Por ejemplo, en la creación de algoritmos de machine learning para la predicción (como los que se usan para recomendaciones en Netflix o Spotify) se puede dar el caso (y se da) de que las predicciones no son totalmente precisas.

- *Privacidad*. Uno de los inconvenientes de la información, sobre todo en el ámbito de la empresa privada, es el tratamiento de datos personales. Existen muchas personas preocupadas porque las empresas puedan compartir entre ellas información privada sobre ellos, aunque solo sea para ofrecer un servicio determinado.
- *Bases de datos*. Para extraer información de manera más precisa y eficaz se requieren grandes bases de datos, espacio de almacenamiento y capacidad de procesamiento para tratarla.
- *Costos*. El punto anterior nos lleva a los costes de la minería de datos, que, si no se trabaja con las herramientas adecuadas, puede ser muy elevado.

Pasos a seguir para hacer un proceso de minería de datos:

Adquisición de Texto

Primero debemos obtener datos sin procesar de alguna parte. La Adquisición de Texto es el primer paso y el más importante antes de la minería de texto. Muchas personas escribirían sus propias arañas(crawlers) usando python u otros idiomas para raspar datos en sitios web. Las bibliotecas como BeautifulSoup4, request o Tweepy se han utilizado ampliamente. Pero para aquellos que no tienen una habilidad de programación de alto nivel o no entienden la estructura web tan bien, la programación parece ser el mayor obstáculo para sus proyectos. En este

caso, otra opción es utilizar algunas herramientas de web scraping necesarias para la codificación

Procesamiento de texto

Por lo general, los humanos procesan textos en nuestro cerebro al leerlos línea por línea para comprenderlos y concluirlos. Durante la minería de texto, la computadora borraría automáticamente de cierta información inútil y cuantificaría los textos útiles transformándolos en números.

1) Procesamiento lingüístico de textos: Para el proyecto de minería de texto, la computadora no podía entender la semántica de las palabras, por lo que solo podía reconocer palabras basadas en la estructura. Por lo tanto, todo el pasaje de textos se dividiría en unidades de texto específicas, como una oración o, más frecuentemente, una palabra. Tokenization, Lemmazation o Stemming son las formas más comunes de separar todo el texto. Después de dividir el texto en palabras, podemos clasificarlas de acuerdo con sus partes del discurso. Como sabemos, habría una cadena sin sentido en textos como “a”, “the” o algunos signos de puntuación. Estos textos se llaman stopword. Una última cosa que se debe hacer al procesar texto es eliminar todas las stopwords y conservar solo los datos significativos.

2) Procesamiento matemático de textos: Después de separar los textos y eliminar todas las stopwords, podríamos comenzar a hacer un procesamiento matemático, que es cuantificar los textos transformándolos en números basados en diferentes parámetros. El parámetro más común es la frecuencia de palabras (Countvectorizer).

Minería de texto

Después de procesar todos los datos, podríamos comenzar nuestros proyectos de minería de texto. Estos son algunos de los ejemplos más comunes de minería de texto:

Nube de palabras: Cree una nube de palabras según la frecuencia de las palabras. Todas las palabras aparecerían dentro de una nube. Las palabras de alta frecuencia aparecerían más grandes que las palabras de baja frecuencia.

Análisis sentimental: El Análisis Sentimental es un proceso que podría ayudarnos a identificar el sentimiento a partir de opiniones basadas en las palabras. Una biblioteca de Python llamada TextBlob podría ayudarnos a analizarlos y generar un puntaje de fuerza de sentimiento positivo o negativo. (por ejemplo, monitoreo de producto o marca)

Modelado de temas: El modelado de temas podría ayudarnos a identificar el tema de un texto. Latent Dirichlet Allocation (LDA) es un ejemplo de modelado de temas que podría clasificar el texto de un documento a un tema en particular. Crea un tema por modelo de documento y palabras por modelo de tema, modelado como distribuciones de Dirichlet. (por ejemplo, etiquetado para reseñas/ noticias/artículos).

Referencias:

Fuente: https://www.octoparse.es/blog/mineria-de-texto-con-octoparse

Fuente: https://www.tokioschool.com/noticias/ventajas-desventajas-mineria-datos/