UNIVERSIDAD DE GUAYAQUIL

INGENIERIA INDUSTRIAL

INGENIERÍA EN SISTEMAS DE INFORMACIÓN

TAREA 1 2PARCIAL

TÍTULO

MINERÍA DE TEXTO

ALUMNO:

MADELINE MUÑOZ VILLEGAS

MATERIA:

CIENCIA DE DATOS

DOCENTE:

ING. HECTOR HUMBERTO DULCEY

CURSO:

8VO SEMESTRE NOCTURNO

AÑO LECTIVO:

CICLO I

2023-2024

DESARROLLO

La minería de texto es una disciplina que se encarga de extraer información valiosa y conocimiento a partir de grandes cantidades de datos de texto no estructurados. Una de las tareas fundamentales en la minería de texto es la representación adecuada de los documentos de texto para que puedan ser procesados y analizados de manera eficiente. Existen diversos modelos y técnicas de representación de texto que juegan un papel clave en este proceso. A continuación, se presenta una investigación y resumen sobre algunos de los modelos más comunes utilizados en la minería de texto.

1. **Bolsa de Palabras (Bag of Words, BoW)**

El modelo de bolsa de palabras es uno de los enfoques más simples y populares para representar texto en minería de texto. En este modelo, cada documento se representa como un conjunto desordenado de palabras sin tener en cuenta la estructura gramatical ni el orden de las palabras. Se crea un vector de características, donde cada dimensión representa una palabra única y se cuenta la frecuencia de aparición de cada palabra en el documento. A pesar de su simplicidad, la bolsa de palabras es ampliamente utilizada debido a su facilidad de implementación y aplicabilidad en diversas tareas.

2. **Word Embeddings (Incrustaciones de Palabras)**

Los word embeddings son representaciones vectoriales densas de palabras que capturan relaciones semánticas y similitudes entre palabras. Estos modelos de representación utilizan algoritmos de aprendizaje profundo, como Word2Vec, GloVe o FastText, para asignar cada palabra a un espacio vectorial de alta dimensión. La ventaja de los word embeddings es que pueden capturar el contexto y el significado de las palabras en función de su contexto en el texto.

3. **TF-IDF (Term Frequency-Inverse Document Frequency)**

TF-IDF es una técnica utilizada para ponderar las palabras en el modelo de bolsa de palabras y resaltar la importancia relativa de cada palabra en un documento en relación con el conjunto de documentos. La frecuencia del término (TF) mide cuántas veces aparece una palabra en un documento específico, mientras que la frecuencia inversa del documento (IDF) mide cuánto es común o rara una palabra en el conjunto de documentos. La multiplicación de TF y IDF produce el peso del término en el documento, lo que permite enfocarse en palabras más relevantes para cada documento.

4. **Modelos de Lenguaje Pre-entrenados**

Los modelos de lenguaje pre-entrenados, como BERT (Bidirectional Encoder Representations from Transformers) y GPT (Generative Pre-trained Transformer), han revolucionado la minería de texto. Estos modelos utilizan arquitecturas basadas en transformers y se entrenan en grandes cantidades de datos textuales para capturar relaciones semánticas y contextuales más complejas. Al utilizar modelos de lenguaje pre-entrenados, es posible obtener representaciones de texto de alta calidad y mejorar el rendimiento en diversas tareas de procesamiento de lenguaje natural.

5. **Análisis de temas y Modelos Latentes**

Estos modelos buscan identificar patrones subyacentes en el texto y agrupar documentos relacionados en temas o categorías. Algunos ejemplos son LDA (Latent Dirichlet Allocation) y LSA (Latent Semantic Analysis).

En conclusión, la representación adecuada de texto es esencial para el éxito de las tareas de minería de texto. Los modelos mencionados, como la bolsa de palabras, TF-IDF, word embeddings y los modelos de lenguaje pre-entrenados, son herramientas fundamentales para transformar el texto en formatos numéricos que puedan ser procesados y analizados por algoritmos de aprendizaje automático y técnicas de minería

de datos. Cada modelo tiene sus ventajas y desventajas, y la elección depende del problema específico y de los datos disponibles. La continua investigación en este campo sigue enriqueciendo y mejorando las técnicas de representación de texto en la minería de texto.