UNIVERSIDAD EAFIT
MAESTRÍA EN CIENCIA DE DATOS Y ANALÍTICA
ST1800 ALMACENAMIENTO Y RECUPERACIÓN DE INFORMACIÓN, 2024-1

Trabajo 2 – analítica de texto

Fecha de entrega: hasta el 14 de abril de 2024

Descripción del trabajo 2

Durante las sesiones de: Sistemas de IR, Sistemas de recomendación y Analítica de texto (NLP y text mining - TM), se pudo ver diferentes técnicas de preparación de texto (tokenización, normalización, remoción de stopwords, stemming y lematización; luego la representación de documentos con diferentes técnicas como BoW-TFIDF y embeddings; además de selección de características (features) como bit-vector, tf, tf-idf y embeddings. Una vez representado el texto a nivel vectorial, se tienen diferentes aplicaciones como los motores de búsqueda como Apache Solr, OpenSearch, Sistemas de Recomendación y diferentes aplicaciones de la minería de textos. Se puede ejecutar una variedad de modelos que nos permitirán extraer información y conocimiento. Se revisaron los sistemas de recomendación en general, modelos de clasificación, principalmente basados en Naive Bayes, si bien hay muchos otros, y detección de tópicos principalmente con LDA. Este trabajo les permitirá aplicar los conocimientos vistos en la clase e investigar, en las siguientes partes del trabajo.

Descripción del problema

Conocer el sentimiento o tendencia de opinión sobre el cambio climático, es un tema muy relevante desde la aparición de las redes sociales, en las que la opinión pública es un factor importante para divulgar o socializar propuestas o movimientos en torno al fenómeno del Cambio Climático. Redes Sociales como X (antes twitter), han logrado crear espacios de opinión en torno a temas tan importantes como este. Este dataset de trabajo (climateTwitterData.csv) recopila más de 70 mil mensajes, los cuales ya han sido parcialmente analizados con técnicas de Aprendizaje Estadístico o Automático, por ejemplo, ya se ha hecho Análisis de Sentimientos.

Los retos de este trabajo2 será aplicar diferentes técnicas de preparación de texto (datos), ejecutar diferentes modelos de aprendizaje estadístico tanto supervisado (análisis de sentimientos y clasificación de texto), así como técnicas no supervisadas como LDA en el cual podamos encontrar insigths por grupo de mensajes. También se pretende crear un sistema de Recuperación de Información, basado en los nuevos modelos LLM, en los cuales como primera etapa se plantee un sistema de consultas sobre este data set.

A continuación, se detallan los diferentes requerimientos:

Desarrollo

Parte 1: Aplicar las diferentes técnicas de preparación de datos, los cuales incluyen un proceso de tokenización, optimización del BoW (con reducción de dimensionalidad), representación de características y de documentos. El objetivo es obtener el BoW óptimo (reducido) para pasar a la fase de representación de característicos y de documentos. (se tiene como columna de entrada 'text'). Realizar la preparación de texto tanto en 1) librerías python como nltk, spacy, gensim o una combinación ellas, y 2) en SparkML o SparkNLP utilizando pyspark.

(nota: en el caso de tokenización, ensaye: `from nltk.tokenize import TweetTokenizer`, un tokenizador especial para datos de twitter, compárelo contra el tokenizador estándar/convencional)

Parte 2: Realizar preparación y representación de documentos basado en LLMs, revisar y escoger algún modelo pre-entrenado de embeddings para texto (ej: openai, hugging face, cohere, etc). Una vez creada la representación vectorial de los textos en embeddings, realizar la indexación y almacenamiento en una base de datos vectorial, de preferencia una de las más sencillas como chromadb (https://www.trychroma.com/) u otra más robusta como opensearch (https://opensearch.org/) o Pinecone (https://www.pinecone.io/)

Parte 3: Realizar detección de tópicos en este dataset. A parte de averiguar cómo hallamos el k óptimo, la intuición del número de tópicos puede ser encontrada en el número de etiquetas diferentes en la columna: 'search_hashtags'.

Parte 4: Realizar Análisis de sentimientos y completar la columna 'sentiment1' o 'sentiment2' del dataset. Se suministran 30 mil tuits ya con Análisis de Sentimiento, y los profesores cuentan con el resto de los tuits analizados para validar la capacidad de predicción del modelo empleado por uds. Además, realizar un modelo de clasificación de tuits que permita estimar el grupo al cual pertenece, puede utilizar 'search_hashtags' como labels o los tópicos detectados con LDA.

Parte 5: Realizar un sistema de recuperación basado en la base de datos vectorial escogida en la Parte 2. Realizar diferentes consultas, que permia a un usuario recuperar un listado de tuits ranqueados.

Como un bonus extra, puede enfrentar uno de los siguientes retos basado en la vector storage:

1) un modelo de clasificación de texto, entrenando un modelo basado en una etiqueta (label) extraído de 1) la columna 'search_hashtags' en el dataset.
2) un modelo de clusterización.
3) búsqueda o recuperación semántica de documentos.
4) opcionalmente, aplicar un modelo moderno de QA (Question & Answering) o generación de prompt y consultas en NLP hacia chatgpt.

Utilice los siguientes sitios para apoyar la realizar de esta parte5:

https://python.langchain.com
https://cohere.com/
https://openai.com/

Se va a emplear el siguiente conjunto de datos:

climateTwitterData.csv.zip

Requerimientos:

Cada uno de los grupos, deberá centrarse en realizar diferentes técnicas y métodos de Preparación de texto (tokenización / optimización, representación de características y representación de documentos; usar técnicas o modelos no supervisados como LDA y clustering; y supervisados como clasificación y otros elementos solicitados en la parte 4.

Se tendrán en cuenta 4 aspectos:

1. Preparación de datos, características y representación en TF-IDF (visión clásica)
2. Preparación de datos, generación de embeddings, indexación y almacenamiento en una bd vectorial (visión moderna)
3. Detección de tópicos con LDA
4. Aplicaciones avanzadas con embeddings y LLM.
5. Implementación de un sistema de Recuperación de Información (IR)

Podrá usar cualquier librería de python de NLP (nltk, spacy, gensim, pyspark, etc)

Criterios de evaluación

En síntesis, el alcance de evaluación que emplearemos en este trabajo son:

Parte 1: 20% - preparación de texto, características, representación en TF-IDF

Parte 2: 20% - preparación de texto, características, representación en embeddings, indexación y almacenamiento en vector storage.

Parte 3: 20% - LDA

Parte 4: 20% - Clasificación de tuits.

Parte 4: 20% - aplicaciones sobre embeddings y LLMs

Regla de ética y transparencia para todas las alternativas:

Si encuentran soluciones públicas o de reúso de código de alguna de las partes requeridas en este trabajo debe EXPLÍCITAMENTE DECLARAR:

- Declarar explícitamente: De que referencias en kaggle, médium, datacamp, toward data science, o de otro sitio, ud empleo parte del código y la solución para realizar su propio trabajo.
- Declarar explícitamente: cual fue el aporte específico que el grupo realizó en el trabajo.

Entregables: POR BUZÓN DE ENTREGA

1. Un documento en PDF que describa los integrantes, un resumen de lo que realizaron (incluyendo lo que deben declarar explícitamente) y cómo usar el código que hicieron.
2. Una carpeta .ZIP con todo lo demás que han usado (excepto los datos), esto incluye notebooks, scripts y documentación adicional que quieran anexar.