Anteproyecto Big Data
Clasificador de noticias

Daniel Enrique Pazmiño Ortega
Juan Felipe Mazo Sanchez
Universidad Autónoma de Manizales
Manizales, Colombia
daniele.pazminoo@autonoma.edu.co
juanf.mazos@autonoma.edu.co

Se pretende hacer una caracterización de noticias, en la que por medio del contenido de estas se logren agrupar por categorías (deporte, negocios, tecnología, política,entretenimiento) estableciendo una similitud.

Se usará una base de datos la cual fue encontrada en la página Kaggle (anexo 1) dicha base tiene un total de 2225 noticias que están repartidas en 1490 artículos para entrenar el modelo y 735 para testear el modelo (el dataset consta del número de artículo y del texto de la noticia). Cada noticia se separará por palabras, se hará un conteo de estas y se establecerá cuales son las palabras que más se repiten; cabe aclarar que también se les hará un tratamiento en el que no se tengan en cuenta las palabras que carecen de sentido cuando se escriben solas (stopwords). Adicionalmente se les aplicará la operación TF-IDF que es un cálculo estadístico adoptado usado para medir qué términos son más relevantes para un asunto, analizando la frecuencia con que aparecen en una página, en comparación con su frecuencia en un conjunto más grande de páginas.
Establecido el TF-IDF de cada texto es posible establecer relaciones entre las diferentes noticias y a partir de las palabras que más se pueden mostrar para dar una idea del tópico al que pertenecen.
En adición, por la forma que tiene el dataset se puede conocer la clasificación que se le dió originalmente y con esto se podrá hacer una comparación con el método implementado analizado su eficiencia y exactitud.
Para finalizar, se pretende generar gráficas que ilustren los temas más comunes y si es necesario implementar una interfaz.

ANEXOS

[1] "BBC news classification | kaggle". Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com/competitions/learn-ai-bbc/data?select=BBC+News+Train.csv (accedido el 27 de marzo de 2023).

REFERENCIAS

[1] C. Maklin. "TF IDF | TFIDF python example". Medium. https://towardsdatascience.com/natural-language-processing-feature-engineering-using-tf-idf-e8b9d00e7e76 (accedido el 26 de marzo de 2023).