Problem 2 : Data Engineering Challenge

Task1 (Data Structuring):

The BBC_articles folder contained .txt files with the category and article ID present in the name and the text content present in the file. The files were read and its content was transferred to a pandas Data frame. The data frame was then saved in the "bbc_articles.csv" file.

Task 2 (Data Preprocessing for Model Training):

The data frame in "bbc_articles.csv" was read for preprocessing. For preprocessing, NLTK was used for tokenization and Word2Vec from genism.models was used for vectorized representation of the tokens in the form of word embeddings.

Tokenization

Stopwords and punctuations were removed. All alphabets were converted to lowercase.

Word2Vec

Word2Vec is a popular approach for preprocessing as it encodes the words in such a way that the cosine similarity between two similar words is maximized while that between dissimilar words is minimized.

The model assigning Word2Vec vectors to tokens was trained on the provided text corpus with 5 epochs with a vector size of dimensions 2 for demonstration purposes. In practice however, such a model would be trained on several epochs with a vector size of 100+.

After applying Word2Vec, the resultant dataset was saved with the required file name "vectored_dataset.csv".

Dependencies

The modules os, csv, pandas, NLTK, string must be required. BBC_articles folder must be in the same directory as the .ipynb file.