Sentiment Analysis Using Classical ML

The objectives of this assignment are as follow:

- Load and use available online dataset
- Build classical Machine Learning models using several feature engineering techniques
- Plot and interpret the result

Assignment Tasks:

Task 1 – Load the movie review dataset of IMDB

- Download the dataset csv file from the following website: IMDB Movie Reviews Dataset
- Load the file using pandas

Deliverable: Print the head of the dataset

Task 2 – Split dataset into training and testing

- Separate features (reviews) and targets (sentiments)
- Split the dataset into training (60%) and testing (40%) sets

Deliverable: Print the number of training and testing examples

Hint: use train_test_split from sklearn.model_selection

Task 3 – Build Model Using Count Vectorizer + Logistic Regression

- Create pipeline: CountVectorizer + Logistic Regression
  - Hint: use make_pipeline from sklearn.pipeline
- Train the model
- Evaluate it

Deliverable: Show Model Accuracy and Classification Report

Task 4 – Build Model Using TF-IDF Vectorizer + Logistic Regression

- Create pipeline: CountVectorizer + TfidfTransformer + Logistic Regression

- Train the model
- Evaluate it

Deliverable: Show Model Accuracy and Classification Report

Task 5 – Build Model Using Pretrained Word Embedding Model + Logistic Regression

- Load a pretrained model (word2vec, fasttext, or glove) from gensim
- Create pipeline: EmbeddingVectorizer + Logistic Regression
    - In embedding vectorizer, you may try different approaches and choose the best:
        - Average the embedding vectors of all existing words in the review
        - Concatenate the embedding vectors of maximum 200 words in the review, use padding with zeros if the review is less than 200 words
- Train the model
- Evaluate it

Deliverable: Show Model Accuracy and Classification Report

Hint: Check how to create the embedding vectorizer that can be added in the pipeline using sklearn

Task 6 – Build Model Using Your trained Word Embedding Model + Logistic Regression

- Train a word2vec model using the same dataset
- Repeat the same steps in Task 5

Deliverable: Show Model Accuracy and Classification Report

Task 7 – Plot and Interpret the Results

- Plot models accuracy in one bar chart graph
- Interpret the result

Submission Guidelines

- Submit a Jupyter notebook named: **SA_ML_Assignment_<YourName>.ipynb**
- Ensure each task is performed in a different cell
- Ensure prints and plot are shown in your notebook when submitting the file