Use Case: Sentiment Analysis of Nigerian Product Reviews using Bag-of-Words and Naive Bayes Models (20 Marks)

Objective

The goal of this assignment is to perform sentiment analysis on product reviews from Nigerian e-commerce platforms using Bag-of-Words (BoW) features and Naive Bayes models. You will gather your own dataset, preprocess the data, create BoW feature vectors, and then apply Naive Bayes classifiers to predict the sentiment of the reviews. Additionally, you will use Doccano for labeling your text data.

Background:

In the rapidly growing Nigerian e-commerce market, understanding customer sentiment is crucial for businesses to improve their products and services. By analyzing product reviews, businesses can gain valuable insights into customer satisfaction and identify areas for improvement. This assignment will help you develop the skills needed to process and analyze text data, which is highly relevant in the context of the Nigerian market.

Instructions:

1. **Data Collection (10%)**:
    - Choose a Nigerian e-commerce platform (e.g., Jumia, Konga) or any other source of product reviews relevant to the Nigerian context.
    - Collect a dataset with at least 1000 (Do up to 10,000 if you can 😊) product reviews.
2. **Data Labeling with Doccano (20%)**:
    - Install and set up Doccano for text annotation (Watch some tutorials on YouTube).
    - Label your collected dataset using Doccano for sentiment (e.g., positive, negative, neutral).
    - Export the labeled data from Doccano for further processing.
3. **Data Preprocessing (15%)**:
    - Clean the text data by removing special characters, stop words, and performing lowercasing (Refer to the recent lecture video and code on Sentiment Analysis).
    - Tokenize the text data.
    - Perform exploratory data analysis (EDA) to understand the distribution of classes and the characteristics of the text data.

4. **Feature Extraction using Bag-of-Words (15%)**:
    - Use the CountVectorizer from the scikit-learn library to convert the text data into a Bag-of-Words feature matrix.
    - Explore and document different preprocessing options within CountVectorizer (e.g., n-grams, max features, stop words).
5. **Model Building and Evaluation (30%)**:
    - Train Naive Bayes classifiers (e.g., MultinomialNB) on the BoW feature vectors.
    - Compare the performance of different models like SVM (Optional).
    - Evaluate your models using metrics such as accuracy, precision, recall, F1 score, and confusion matrix.

Report (10%):

- Write a comprehensive report summarizing your approach, experiments, results, and any challenges you faced.
- Include visualizations, tables, and charts to support your findings.
- Discuss the implications of your results and suggest potential improvements or future work.

Deliverables:

- **Python Code**: Well-documented Jupyter notebooks or Python scripts containing your code for data preprocessing, feature extraction, and model building.
- **Dataset**: The dataset you collected and labeled.
- **Report**: A detailed report in PDF format summarizing your approach, experiments, results, and conclusions.

Submission Instructions:

- Submit your code, dataset, and report in a single ZIP file.
- Ensure that your code is executable and includes instructions for setting up the environment and running the scripts.

Good Luck!!!