Project: Naive Bayes Classifier for Sentiment Analysis

Due: 06/13/2024 11:59 PM

Project Description:

In this project, students will implement a Naive Bayes Classifier (NBC) for sentiment analysis on a dataset containing reviews and their respective star ratings. The datasets, "train.csv" and "test.csv", will be provided. A review with a 5-star rating will be considered positive, while all other ratings will be considered negative. **Do not use** any publicly available code–your code will be checked against public implementations or AI-generated codes. Other packages like pandas are also not allowed to use. Built-in packpages like csv or os are good to use.

The project consists of three tasks:

Task 1: Feature Selection (10 points)

- Students will preprocess "train.csv" and select the top 1000 words (by frequency) as word features for their model. All other words will be ignored.
- Please print out the top 20-50 words from the selected features.

** Preprocessing Guideline:*

a. Convert all text to lowercase

b. Remove special characters.

c. Tokenize the text into words.

d. Remove stop words.

e. Select 1000 most frequently appeared words for the final features

Task 2: Model Training and Evaluation (15 points)

- Using "train.csv" and "test.csv", which they will use to train and evaluate their Naive Bayes Classifier with Laplace Smoothing
    - Laplace Smoothing: Implement Laplace smoothing in the parameter estimation. For an attribute Xi with k values, Laplace correction adds 1 to the numerator and k to the denominator of the maximum likelihood estimate.

- Evaluation measure: Accuracy

- Please describe your observations and provide an analysis of their model's performance.

Task 3: Learning Curve Analysis (5 points)

- Students will plot a learning curve by varying the amount of training data used [10%, 30%, 50%, 70%, 100%]. The testing set will remain unchanged.
- For this plotting task only, students may use external plotting packages like the Matplotlib.
- Students will describe their observations and provide an analysis of the learning curve.

Deliverables:

1. Python code implementation of the Naive Bayes Classifier.
2. README file for executing your code.
3. PDF report