CSE 4392 Special Topic: Natural Language Processing

Homework 4 - Spring 2024

Due Date: Feb 24, 2024, 11:59 p.m. Central Time

In this assignment, you will have the opportunity to dive deep into the inner workings of logistic regression and implement the algorithm from scratch without relying on any external libraries like PyTorch or scikit-learn. By completing this assignment, you will gain a solid understanding of the underlying principles of logistic regression and strengthen your skills in coding and machine learning.

Problem 1 - 100%

Your task is to build a text classification model to classify product descriptions from an E-commerce website into four categories: Electronics, Household, Books, and Clothing & Accessories. You will use the provided dataset in CSV format, which contains two columns: the first column represents the class name, and the second column represents the corresponding product description. You can find the dataset here: `https://drive.google.com/file/d/1YfG0iy0vpv7L0aqzDxvzHlW79HDdpZvg/view?usp=drive_link`

Step 1 Data Exploration

- Load the dataset from the provided CSV file into your program.
- Explore the dataset to gain familiarity with its structure and characteristics.
- Analyze the distribution of classes to understand the class imbalance, if any.

Step 2 Feature Extraction

- Perform necessary preprocessing steps on the text data. This may include removing special characters, converting text to lowercase, tokenization, and removing stop words.
- Extract **features of your choice** from the preprocessed text data.

Step 3 Model Training

- Implement a logistic regression classification algorithm to train a text classification model using the features you have extracted.
- Train and test your classification model using 10-fold cross-validation.

Step 4 Model Evaluation

- Evaluate the performance of your trained model by 10-fold cross-validation.
- Calculate evaluation metrics such as average accuracy, precision, recall, and F1-score to assess the model's performance.
- Analyze the confusion matrix to understand the model's performance across different classes.

Step 5 Documentation and Submission

- Show the results, details about the preprocessing techniques, feature extraction methods and the analysis in the pdf file and compress the code file and pdf file into a zip to submit.
- Discuss the model's performance, including strengths, weaknesses, and potential areas for improvement.
- Write code in Python 3.