ICS 661 Assignment 2

Overview

This dataset contains movie reviews labeled with binary sentiment polarity (positive or negative) and is intended as a benchmark for sentiment classification tasks. This document outlines the structure of the dataset and the tasks associated with its usage in this project.

Dataset

The core dataset consists of 50,000 movie reviews, split equally into 25,000 reviews for training and 25,000 for testing. The label distribution is balanced, with 25,000 positive and 25,000 negative reviews.

To reduce correlation in ratings for the same movie, no more than 30 reviews are included for any single movie. Additionally, the training and test sets contain reviews from different sets of movies, ensuring that performance is not skewed by memorizing movie-specific terms linked to their sentiment labels.

- A **negative** review is defined as having a score of **4 or less** out of 10.
- A **positive** review is defined as having a score of **7 or more** out of 10.
- Reviews with neutral ratings (scores 5 or 6) are excluded from the dataset.

File Structure

The dataset is organized into two top-level directories: train/ and test/, which correspond to the training and test sets. Each of these directories contains two subdirectories:

- pos/: Positive reviews
- neg/: Negative reviews

Each review is stored as a text file, named according to the pattern [id]_[rating].txt, where:

- [id] is a unique identifier for the review
- [rating] is the original star rating of the review (on a scale of 1-10)

For example, the file test/pos/200_8.txt contains a positive review from the test set, with a unique id of 200 and a rating of 8/10. However, in this project, you only need to classify reviews as positive or negative, without predicting their specific ratings.

Project Task

The primary task is to classify the reviews as either positive or negative based on their text content. Text data can often be noisy, containing typos, punctuation, and irrelevant tokens, making it more challenging to handle compared to image data.

1. **Text Preprocessing**:
    - Clean the dataset by removing stop words, punctuation, and normalizing the text. You may refer to the following guide for text cleaning: Cleaning Text for Machine Learning
2. **Vocabulary Extraction**:
    - After cleaning the data, extract a vocabulary set from the provided dataset, which will serve as the input feature set for the model.
3. **Model Training**:
    - Train a Recurrent Neural Network (RNN) model using the provided training set to classify the reviews as positive or negative. You do not need to infer the exact ratings, only the binary sentiment (positive vs negative).
    - Your goal is to achieve at least **75% accuracy** on the test set using the RNN model.

Good luck!