ERG2050: Introduction to Data Analytics

Final Project

April 15, 2021

Due: 11:59 pm, May 5, 2021

Topic: Sentiment Analysis

Descriptions: This dataset contains movie reviews along with their associated binary sentiment polarity labels.

The core dataset contains 50,000 reviews split evenly into 25k train and 25k test sets. The overall distribution of labels is balanced (25k pos and 25k neg).

In the entire collection, no more than 30 reviews are allowed for any given movie because reviews for the same movie tend to have correlated ratings. Further, the train and test sets contain a disjoint set of movies, so no significant performance is obtained by memorizing movie-unique terms and their associated with observed labels. In the labeled train/test sets, a negative review has a score <= 4 out of 10, and a positive review has a score >= 7 out of 10. Thus reviews with more neutral ratings are not included in the train/test sets.

There are two top-level directories [train/, test/] corresponding to the training and test sets. Each contains [pos/, neg/] directories for the reviews with binary labels positive and negative. Within these directories, reviews are stored in text files named following the convention [[id]_[rating].txt] where [id] is a unique id and [rating] is the star rating for that review on a 1-10 scale. For example, the file [test/pos/200_8.txt] is the text for a positive-labeled test set example with unique id 200 and star rating 8/10 from IMDb.

We also include the IMDb URLs for each review in a separate [urls_[pos, neg].txt] file. A review with unique id 200 will have its URL on line 200 of this file. Due the ever-changing IMDb, we are unable to link directly to the review, but only to the movie's review page.

You are required to perform the binary sentiment analysis on the above dataset.

Grading standard:

- Work distribution (The distribution should be included in the reports and slides.)
- Project Report (at leaset including the motivation/insight, approach, experiments and conclusion) (The report should be submitted on time.)
- Project Presentation (25 mins) (The slides should be submitted on time.)
- Code (The code should be submitted on time.)

NOTE: The accuracy is NOT the only criteira.