CS4375

Project Proposal

Project Title: Predicting User Film Ratings Based on Text Reviews

Data Sets to Use (Planned):

- http://ai.stanford.edu/~amaas/data/sentiment/
- https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset?select=IMDb+ratings.csv

Project Idea:

Our project will focus on the relationship between word choice in user text reviews of movies and the resulting user score of the film. Movies and media have been a mainstay of revenue generation across the world, but opinions often vary between individuals due to differing perceptions and values. Due to the complexity of emotion and thought, it can be difficult to boil down one's opinion to a 1-10 scale. However, there would be correlations between an individual's own choice of words and ultimately the value they assign the film.

Our goal is to generate a probable estimate of an individual's rating for any particular film based on the content of their written user review. By analyzing the user's word choice and comparing it to a predefined dictionary we can associate with positive or negative impressions, we can attempt to quantify and predict what score the individual will give to the movie. Realistically, it may be necessary to group together positive and negative scores (e.g. 7-10 being positive, 1-4 being negative) and see if there is a correlation for the middling reviews of 5-6. Furthermore, if inconsistencies emerge between the correlation of score and word choice, we can build a dictionary from the training data.

Software and Techniques:

- Text Classification Software - This software would analyze the user text reviews and determine the overall tone of the review based on the connotations of individual words and phrases. The ratio of positive words to negative will be a major factor in generating the final predicted score.
- Naive Bayes - This will serve as a baseline for our data analysis with the training set.
- Multiclass Classification - When analyzing the ratio of positive and negative keywords, it may be possible to categorize the results if some show stronger correlations than others.

Papers to read (relevant articles)

- https://www.aclweb.org/anthology/P11-1015.pdf

Group Team Members

James Baker: jlb170003
Natasha Trayers: nnt180002
Hansen Li: hxl180039

CS4375

Milestone Checklist

- Data Collection/Compilation
- Data Analysis
- Model Building
- Report
- Completion due May 9th