Project Progress Report

Names: Elizabeth Lin (el18@illinois.edu), Nirali Rai (narai2@illinois.edu)

GitHub link: https://github.com/el18illinois/CourseProject-CS410-MovieBee

The progress we have made for our project is that we have collected the raw data which includes initial data of various movie scripts including Thor, Jaws, and other classic Hollywood movies.

The data collected was in the form of movie scripts/screen plays. The data is formatted as a text file that contains the script. Each script is in its own text files. The movies selected were from a variety of different categories, but also movies that are very commonly known to a diverse population of people.

We went through research by looking at various dataset on open source including Kaggle and Google libraries.

Next, we look into strategies to parse the data. Currently we are trying to tokenize each script into a bag of words to test and evaluate the sentiment. As we are brainstorming our approach we realize the overall sentiment of a specific movie is hard to access so we are using split from each movie script. That would mean we need to split the movie script into different parts to assess individual sentiment of different parts and then the overall conclusion of each script based on the scores of each part.

In this process, we have dived into Model Design. We plan to utilize bag of words models as a baseline and analyze sentiment using scoring methodology. We will first get rid of common words from our vocabulary and associate each movie script vocabulary (containing only relevant words) to a scoring based on different moods (happy, sad, scary, mysterious, etc.). We are deciding which moods to label each movie, which means the list will be updated in our next progress report and throughout our analysis.

Some challenges we faced in this process was collecting essential scripts from open source because it was hard to find accessible data that was not corrupted and formatted as easy to parse. It is also challenging to craft the design of labeling movies because there are many genres associated with movies and we do not want to associate genres as the mood/sentiment. Our approach is to predict the sentiment based on given script analysis then use the genre as a form of verification of our data. However many moods can match a certain genre. This is why we decide to split movies into portions so we can evaluate each part of the movie script separately in terms of the mood.

Next, we need to start implementing our model which includes writing Python code for Bag of Words model and Tokenization of scripts. I foresee a challenge will be to parse the data and organize effectively for model scoring.

Checklist:

1) ~~Initial Research: 4 hours~~
    a) ~~Kaggle~~
    b) ~~Reading descriptions of datasets and format~~
2) ~~Dataset Collection: 4 hours~~
    a) ~~Collect sample datasets to test with~~
    b) ~~Evaluate and compare datasets~~
3) ~~Data Parsing: 6 hours~~
    a) ~~Parse datasets dependent on program needs~~
4) Model Design: 10 hours
    a) Create structure/algorithm
5) Model Implementation: 10 hours
    a) Code up selected model(s)
    b) Verify model functionality against given dataset
6) Testing & Evaluation: 3 hours
    a) Data Validation Tests
    b) Computing Accuracy with Precision and Recall
7) Documentation: 3 hours
    a) Document how model runs and design reasoning

Total: 2 * (N=2) = 40 hours estimated work