Aliva Panigrahi (alivabp2)

Chirag Gupta (chiragg4)

Shreyah Prasad (sprasa20)

CS 410 Course Project Progress Report

For the past few weeks, we have been working on the following deliverables specified in our project proposal:

- Learning/using BeautifulSoup
- Scraping site contents
- Evaluating Letterboxd
- Sentiment analysis of the reviews

Current Progress & Challenges

We have successfully scraped the contents of Letterboxd that we need for our project: all movie reviews posted to the site this year, including those covering recent releases. We decided on pulling this specific set of data, as we found it would give the most relevant results for our system. To accomplish this, we used BeautifulSoup which we took some time to get familiar with. As a result, we have a large set of samples consisting of the movie, the corresponding numerical rating, as well as the text review upon which we plan to conduct sentiment analysis.

Upon starting to devise our sentiment analysis algorithm, we immediately spotted one challenge: the sarcasm within the text reviews. The comedic nature of Letterboxd content is why we chose this particular site, as the reviews are more personable. However, this also makes sentiment analysis a significantly more challenging problem, as sarcasm in a positive review can result in a review being categorized as negative. This is a problem that we are working to address, by balancing in other factors such as the numerical rating and potentially the number of likes that a review has.

Moving on to sentiment analysis, we have been using the NLTK Python library to process the text reviews. We found that NLTK has a movie review corpus which we are able to use in our implementation. We have been training and testing with a variety of models to gauge accuracy, including: linear support vector machine, random forest classifier, etc… Both these models have been tested with multiple feature vectors. So far the best test accuracy is around 87% with a linear support vector machine using a feature vector with length 8,000. All of these models have been trained using the NLTK movie review corpus. Additionally, we downloaded a dataset of iMDB reviews to try a different training method. We hope with these additional datasets, we will be able to improve our assessment of Letterboxd text reviews.

Upcoming Deliverables

Once we are able to improve the accuracy, we need to then develop a ranking method based on the results of our classifier. This is our next step, in addition to continuing to improve the model. We are continuing to find ways as mentioned above to better analyze reviews with sarcastic content. We also plan on testing out regression models to predict ratings rather than simply creating a basic positive/negative classifier. This way, we can get more information and a more accurate estimate of what a reviewer thinks of a movie. Another method we will try is stemming and lemmatization to gauge if this will help improve accuracy. Ultimately, if time permits, we will make the recommender system more personalized (ie: users can select what they are recommended based on positivity, negativity, rating, genre, etc…).