Arda Bedoyan

Progress Report

CS 410: Text Information Systems

Genre Analysis of Movie Scripts

Progress Made:

Most of my progress has been made in the form of research. I originally had hoped to be able to find a topic analysis algorithm or toolkit that would be able to identify and define different genres based on a list of predetermined genres. However, after doing some research and exploring the Stanford Topic Modeling Toolbox (TMT), the MeTA toolkit, and the Gensim toolkit, I realized that it won't be as simple as I had originally planned. I cannot enter a text file of a movie script and have the toolkits output the genres. Rather, what I noticed to be the case was that many of the toolkits would output a chosen number of topics with a list of the top words per topic, but they did not label or define the topics as genres. If I chose to go with this method, then further analysis would need to be done to determine which topic matches with which genre. Instead, I am thinking of using a feature available within the Gensim toolkit that offers the ability to find the similarity between documents. Therefore, I can start by labeling a chosen number of documents (scripts) by their top genre. Then for each new document I add to the collection, I will run the Gensim function to determine the similarity of the inputted document with the rest of the collection. The output will be the top three documents that the input was similar to, and this will be used to determine the top three genres for the inputted document. As more documents are added and the algorithm is run again and again, then the results may change to be more accurate.

I currently have fifteen movie scripts that I will start my collection with. Also, I have decided on 11 genres (i.e., topics), action, comedy, drama, romance, horror, science fiction, thriller, fantasy, war, sports, and western.

Remaining Tasks:

Remaining tasks include finalizing the algorithm with the Gensim similarity function and evaluating that the program will output genres for a given script. Additionally, I had originally planned to create a front-end web application where users could go and enter a movie title and see the top three genres for that movie. However, I think that I will take the advice of one of my reviewers and focus on getting the algorithm right before moving on to the web development. Therefore, I am pushing the front-end aspect of the project to bonus work that I will try to get to if time allows. I want to focus on ensuring that the program actually works and that the similarity feature provided by the Gensim toolkit will be able to provide accurate results.

Challenges and Issues:

The main challenge I faced was determining the best way to classify genres per movie scripts. I think that many common topic modeling functions did not provide the versatility I needed to be able to determine the movie genres without further human involvement. Therefore, I have settled on using a similarity measure to analyze how similar a script is to other scripts in the collection and then determine the genres based off the genres of the similar scripts. Another challenge I face is being able to get to the front-end development to have a project that is easily accessible from the web. However, with this now being bonus work, I am hopeful that I can deliver a working program that determines the genre of a movie script.