GARKL Project Progress Report

Team Information

Team Member	NetID
Captain: Abi Venkat	abinand2
Karan Gulati	kgulati2
Gabin Ntankeu	ntankeu2
Leon Li	angl2
Ratul Saha	ratuls2

Completed Tasks

We identified that movie reviews from IMDB, cast and plot details from Kaggle, movie listing from Netflix will give us a good set of data to work on. We have started conducting data preparation. We have looked at a couple of data sets and also have used the IMDb api to query certain results. We have a list of about a thousand movies that have an ID, title, genre(s), and IMDb ids.

We finalized a list of features to build, e.g., search algorithm on reviews, genre, plot, cast etc. to show the movie listing. Running a sentiment analysis on reviews to identify what movies are suitable for the user based on their search.

We have also created a mock up of what our web-based front end could look like. We are thinking of building a Jupyter Notebook based interface which does not depend on too many technical layers compared to a functional web application.

Overall we are still in the research / data gathering phase of the project and have done minimal coding. We have an idea of how we want our data sets to interact, we need data sets of ratings and movie information from IMDb and movie information from Netflix. We can join the IMDb sets and order them by rating, then join this table on the movie titles with the Netflix data set. This way we will be able to query through this data set based on user inputs within our search engine.

The input of the program would be a genre and search text (a query) by the user. The expected output of the program would be a list of Netflix movies with their reviews from IMDB, metadata details, along with results from the sentiment analysis. The result would be sorted and filtered on positive or negative reviews.

Pending Tasks

We still need to figure out the flow of our application in explicit terms such as what the input and outputs are for certain parts of the project (e.g., what input to use for this part of the program and where will that output go as an input for another part of the program). We have a general idea of how it would look at a high level, but it would be beneficial for us to write down an exact workflow of our application at a medium level scope (e.g., the user input is passed into a specific function to figure out what specific parameters we are looking for).

We also need to figure out the way we will clean and join our data, as using two different data sets (IMDb, Netflix) will mean that they are probably not stored the same way. We may need to create our own ID's or clean the data such that we can join and search it accordingly.

We still need to finalize what data sets we are using. This includes curating our own data set(s) that are ready to use in our project. We still have a bulk of the coding to do and this includes mostly (1) implementing a method for retrieving a bag of words relevant to the user's query and (2) using sentiment analysis as a filter to only look at movie titles that have a majority of positive reviews.

In terms of infrastructure as well we are pretty sure we will use something locally to demo the program instead of having it hosted somewhere. We also talked about using google colab.

Challenges

So far we have not faced any direct challenges. One thing we were thinking about changing was the mood input into our program. Most sentiment analysis results will be a binary classifier telling us positive or negative. We originally thought about using this for determining mood of a text, but since mood is a non-binary classification this may be tough to do.