GARKL Project Proposal

Team Information

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

Free Topic Theme

We have chosen to pursue the idea of a Netflix movie recommender system that is based on IMDb reviews. Netflix has a notoriously awful recommender system, and the built in review doesn't have the completeness that IMDb reviews have. We want to build a tool with a website serving as a GUI so that a user can input details of what kind of movie they are looking for and we can use that data to query netflix movie titles and combine that with data we collect from the IMDb to give a better recommendation than the sole netflix recommender system can.

We are thinking we will need some kind of back-end probably done in python for the actual recommender portion of the project and then use the common web development languages for the front-end. We will need to perform sentiment analysis on the IMDb reviews (collected via their api) and possibly some other type of text retrieval algorithms, and combine that with the netflix query dataset we have to inform the user of the recommendation(s). We will evaluate the tool by performing relevance judgements ourselves and perhaps this can be a feature of the tool to allow the user to do as well.

Data sets we are considering using include Netflix's platform data set, which includes all information regarding the movies and television shows available on their website. Along with this, a data set with review information (google or IMDb) that includes more information in regards to the movies and tv shows in the netflix data set. This would allow for API calls and possible database usage to provide a recommendation of the movies and television options provided in Netflix.

To summarize, We will make a web application which will ask the user what they are looking for (mood, genre, previous relevance judgements) and the recommender system will run at the backend to look into movie reviews, metadata information and suggest top 5 movies that match best with the user choices. We will also ask the user to judge (thumbs up / down) whether the recommender output was relevant to their choice.

Programming Languages

We plan to use python as our recommender system, data crawling/scraping, and dataset preparation. Frontend will be built on javascript, HTML, CSS.

Workload Justification

1. Please justify that the workload of your topic is at least 20*N hours, N being the total number of students in your team. You may list the main tasks to be completed, and the estimated time cost for each task.

   Our work would be distributed around below main areas:
   - Data Preparation - *20 hours*
     - Research into whether we can query into existing data stores, or we have to create our own
     - Scrape data from various internet sources
     - Clean and process data from kaggle, google movies, IMDb movie reviews
     - Prepare the dataset for analysis
   - Server Management - *10 hours*
     - Deploy backend server
     - Deploy website
   - Tool Building - *50 hours*
     - Build datasets
     - Build programs
     - Integration with the frontend
   - Testing/Tuning - *20 hours*
     - User Interface testing
     - Dataset testing
     - Program logic testing