Basic Information

Project Title

Exploring Movies (tentative)

Name

David Lu

Email

dlu1@andrew.cmu.edu

Project Repository

https://github.com/dylu/movie-vis

Background and Motivation

I love watching movies, and I'll generally enjoy any movie I watch, in a variety of genres. However, while it would be nice to just "watch them all," there are simply too many movies and too little time. I would like to see if there's a relationship between certain types of metadata (such as movie length, director, actors, etc), and its overall rating.

Additionally, if it proves interesting, I would like to compare data between different specific movies – this way, I can see how a movie I may have particularly liked or disliked stands in the system compared to all the other movies available.

As a bonus, movies also tend to be a fairly well-organized set of data, making it less of a headache to deal with missing or "dirty" data.

Project Objectives

I want to see how movies measure up against one another, and see if there can be a meaningful prediction on how well a movie gets rated based on its metadata.

If there are interesting trends or general patterns among movie hits, we may be able to pick out the ones “most likely” to be rated well, and watch those. In addition, if there tends to be a type of movie an individual may like, we may be able to find similar ones based off a variety of factors.

Of course, it may just be interesting to see how certain favorite (or hated) movies an individual has measures up to the giant landscape of movies overall or perhaps if movies being made are trending in a certain direction over the last few years. (Maybe certain genres are getting more popular, or movies may be getting shorter overall!)

The benefits of this visualization:

- Allows someone to see how certain movies they like measure up to others
- Easily lets someone identify similar movies (movie recommendations, maybe?)
- May help someone predict if they might like a new movie based off its metadata
- Lets the user explore general trends in movies over time

Data

IMDb provides its information in a raw data format, available here:
http://www.imdb.com/interfaces/

MovieLens also provides its data across 20 million ratings and tags, available here:
https://grouplens.org/datasets/movielens/

I will likely be cross-referencing movies between both datasets; IMDb provides solid metadata information, however MovieLens provides user-specific data as well as individual ratings (and timestamps of said ratings). I believe if I want to explore not just overall trends but also how movies relate to one another (or perhaps building a similar movie list, with recommendations based off whether a user liked or disliked certain movies), then the MovieLens dataset would prove to be especially useful.

However, due to its vast database and basic overall metadata sources, I think IMDb will still be a good dataset to start with. Both datasets are free for non-commercial uses.

Data Processing

IMDb Dataset

Unfortunately it seems like the IMDb dataset will require a bit of cleaning before I start visualizing. I plan on extracting most of the basic metadata information, including the movie's Title, Actors/Actresses, Director, Length, Year, Genre, among others if I find them useful (or well-populated enough).

I'm most comfortable with Java, so if I end up requiring significant pre-processing / cleaning of the data, I will likely write a few Java classes to process this data directly from the raw text files downloaded from IMDb. In the case that this data isn't accurate compared to the website, I may write a few small smart-crawlers to handle it.

I plan on saving the result as a .csv file, and access it using Javascript – most of the minor processing will be handled using Javascript from my generated .csv files.

MovieLens Dataset

The MovieLens dataset seems much more organized and clean, requiring minimal pre-processing. However, due to how it's arranged, I may require a lot of custom objects or data-structures in my Javascript code to handle its multiple table cross-references. (As well as 'linking' it to the IMDb data)

Extra Data

If I have time, I may choose to further explore trailer data from scraping YouTube (or other sources if they arise), however this seems unlikely at the moment.

Visualization Design

My general idea is to design various modules, which I can then combine to create an overall visualization. For example, one module may display ratings over time, while another one may separate data-points into their respective genres.

Since these data entries should be obtained from a single source, it should be easy to link the data-points between different modules if applicable, providing an easier exploration experience for the user.

For the modules in question, please refer to the sketches document.

Must-Have Features

- Drilldowns to see where individual movies (or genres) fit in.
- Separate views for at least 3 different main metadata indicators.
- Interactivity between views (e.g. hovering over one changes how others respond)

Optional Features

- A way to search for how an individual movie fits in.
- Comparison page between two different individual movies.
- Individual movie page view, presenting information regarding that specific movie.

Project Schedule

There are a total of six weeks before the final presentation is due, with a project milestone due in three. I have my project schedule organized as follows:

Week 1:

Finish collecting and preprocessing all data needed.
Start brainstorming about necessary data-structures in frontend (Javascript).

Week 2:

Read in data to d3.
Basic visualization of one "module" completed.

Week 3:

Basic visualization of three separate "modules" completed.
Modules may not need to be linked yet.
Project Website up and running.

Week 4:

Modules linked together.
Explore different interactivity methods.

Week 5:

Start finalizing documentation and process book.
Iron out any bugs and last-minute feature additions.

Week 6:

Finish Project Screen-Cast.
Debugging.