Film Analytics

Data Dawgs

(Siddhant Sutar, Dalton Childers, Neil Matin, Abhimanyu Tomar)

PROGRESS REPORT

Achievements

- Populated an SQLite database of 1.1 million movies fetched from IMDb. A typical row in the database table features fields such as the movie title, plot summary, director, movie (IMDb) rating amongst others.
- Currently storing the data in an SQLite database, but have potentially looked at MongoDB and PostgreSQL because of their faster write operations.
- Using OMDb API to fetch the IMDb and Rotten Tomatoes movie data. However, the API does not support the retrieval of box office data.
- Originally planned on making the API calls by searching by movie title using the list of titles obtained from IMDb, but parsing the movie title is nearly impossible to achieve because of strange movie names. The current algorithm features brute forces through the list of all possible IMDb movie IDs.
- Wrote a web scraping script to retrieve the movie box office data using urllib2 (Python 2.7) and Regex. Looked at BeautifulSoup, a Python web scraping library, but it seems to be counter-productive for our purpose as well as inefficient.

Future Plans

- Analyze trends based on the data, for example:
  - production houses with higher box office success rate
  - comparison between commercially and critically acclaimed movies based on IMDb and Metascore ratings
  - box office success rate for a movie based on the release time
  - establish a relation between plot analysis (using Python’s Natural Language Processing (NLTK) library) and critical and commercial success

- Automate the process and make the database dynamic by updating it with new movies and change over time in user ratings
- Develop a web front-end