Stanford

A Movie Recommender System from Tweets Data

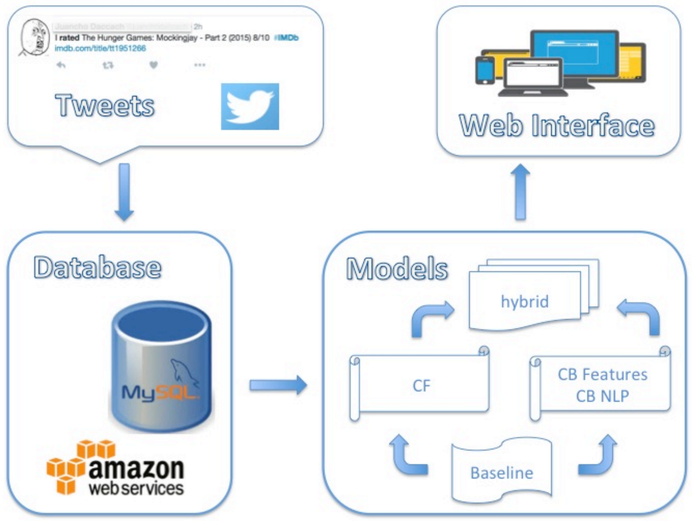
Mengyi Gao
Statistics

Xiang Zhang
Statistics

Introduction

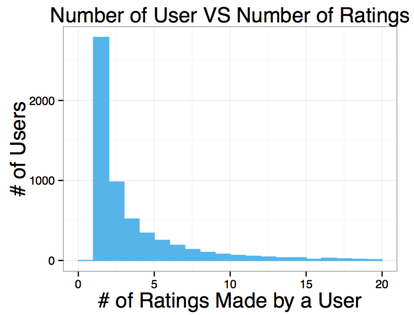
- Collected user ratings from Tweets
- Extracted movie information from IMDb website
- Built a collaborative filtering model, a content-based model, and a hybrid model incorporating the two
- Evaluated the models by test RMSE
- Created a web interface

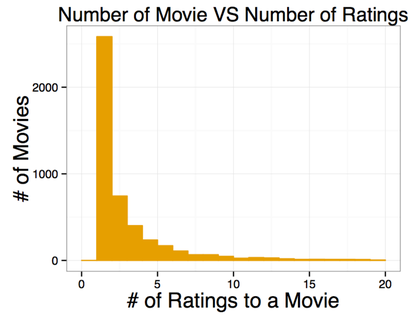
Pipeline



Data Preprocessing

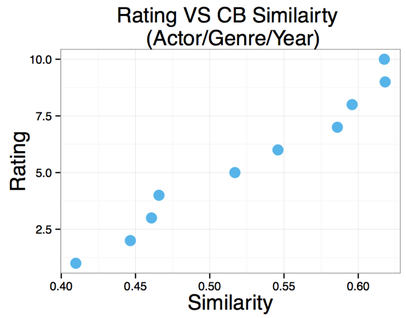
- Created a mysql database to store the data
- Automated updating database periodically through Twitter API and OMDb API
- Removed users who rate less than 5 movies
- Omitted movies with less than 5 ratings
- Split the data into 80% training set and 20% test set

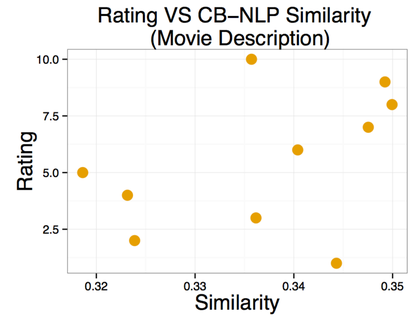


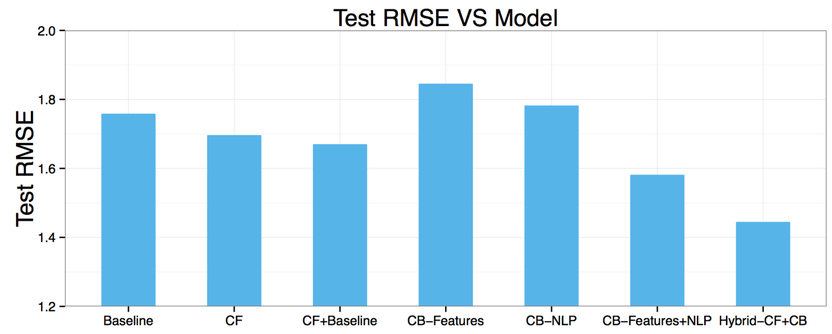


Models

- Defined baseline model as global mean + user bias + movie bias
- Built an item-item collaborative filtering model
- Constructed content-based models through 1. movie features (actor/genre/year) 2. NLP on movie descriptions
- Combined the above into a hybrid model

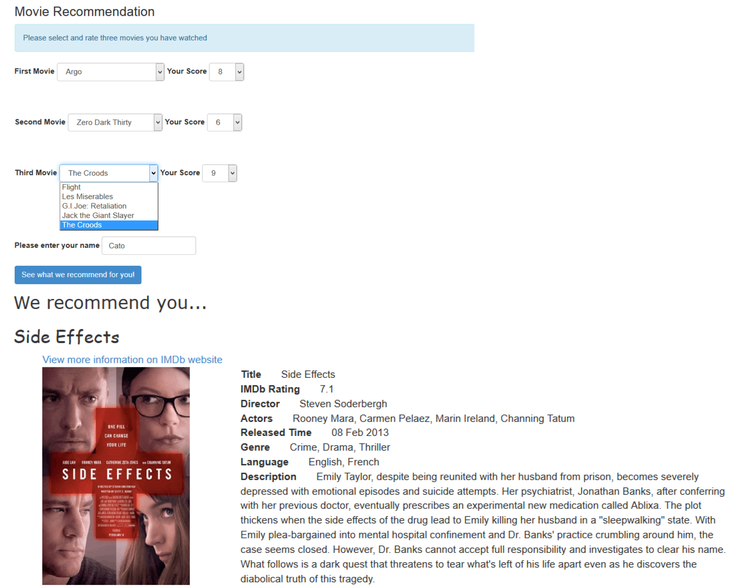






Model	Description	RMSE
Baseline	Global mean + movie average + user average	1.757
CF	Item-Item Collaborator Filtering	1.695
CF+Baseline	Remove Baseline Effect CF on Residuals	1.668
CB-Features	Movie Actors, Years, Genres (Regression)	1.844
CB-NLP	Movie Description (Regression)	1.781
CB-Features+NLP	Mix the Feature Vector	1.580
Hybrid	CF+Baseline and CB-Features +NLP (Regression)	1.443

Web Interface



Future Work

- Run diagnostics for bias vs. variance
- Increase model efficiency
- Explore more NLP methods
- Add more relevant features
- Improve web user interface