Dylan Tan

Data Science with User Ratings and Reviews

In 2009, Netflix awarded $1,000,000 to BellKor's Pragmatic Chaos team for creating a recommendation system that beated Netflix's own recommender by 10.06%. Popularity of recommendation systems has exploded since then, and can be found in high-trafficked sites such as Facebook, Youtube, and Reddit. The impact of recommendation systems on commercial ventures has been huge. For instance, Amazon has been quoted to have increased product sales from recommendations by 35%[1]. Thus, our challenge is to gather data of user review and ratings and create a recommendation system that can predict user preferences on unrated items.

Milestones:

1. Project Selection + Data Gathering:

Form teams of 2 or 3 and select a data source from the provided list.

- ❑ **Anime Reviews from AniList** - http://anilist-api.readthedocs.io/en/latest/anime.html
- ❑ **Amazon Food Reviews** - https://www.kaggle.com/snap/amazon-fine-food-reviews
- ❑ **IMDb Movie Reviews** - http://www.imdb.com/
- ❑ **MetaCritic Movie, Music, Video Game Reviews** - http://www.metacritic.com/
- ❑ **Other Sources of Reviews**

1. *http://venturebeat.com/2006/12/10/aggregate-knowledge-raises-5m-from-kleiner-on-a-roll/*

2. Literature Review:

Here's a list of relevant articles and papers:

1. High Level Overview of Content Based Recommendations - http://www.analyticsvidhya.com/blog/2015/08/beginners-guide-learn-content-based-recommender-systems/
2. General Overview of Recommendation Systems - http://infolab.stanford.edu/~ullman/mmds/ch9.pdf
3. Python Collaborative Filtering Example - http://www.salemmarafi.com/code/collaborative-filtering-with-python/
4. Netflix Prize - http://netflixprize.com/assets/GrandPrize2009_BPC_BellKor.pdf
5. Linear Regression with Recommenders - http://www.cse.psu.edu/~xxg113/fskd11.pdf

3. Data Exploration and Cleaning:

Perform the following exploration steps:

- ❖ Decide on a suitable database to store the data, and on a computing resource to process the data (AWS, Microsoft Azure, personal computer).
- ❖ Perform Feature Extraction/Selection (where can you get other features?)
- ❖ Remove/Clean records with spurious entries (e.g. null values, unmatched titles, etc)
- ❖ Visualize popularity of different genres.
- ❖ Check for any correlations between features.

4. Proposal:

Propose methodologies and ideas to be implemented, tested and interpreted for your final project. What are the key questions that your project will answer?

For example:

- ★ What are the features that makes movies popular?
- ★ Can we predict sentiments of reviews?
- ★ Can we predict user preferences?

5. Implement Baselines + Recommender:

- Decide on the *performance metric* to evaluate prediction.
- Implement the following technique(s):
  - *Simple averaging*: Predict user ratings of items based on user and item preferences
  - *Content-Based Recommendations:* Model user + item attributes and interactions with utility matrix
  - *Collaborative Filtering*: Recommend items based on similarity measures between users and/or items.
  - Other models such as linear regression