**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**

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

**Models**

- **Baseline**: Global mean + movie average + user average
- **CF**: Item-Item Collaborative Filtering
- **CF+Baseline**: Remove Baseline Effect CF on Residuals
- **CB-Features**: Movie Actors, Years, Genres (Regression)
- **CB-NLP**: Movie Description (Regression)
- **CB-Features+NLP**: Mix the Feature Vector
- **Hybrid**: CF+Baseline and CB-Features + NLP (Regression)

**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

**Future Work**

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