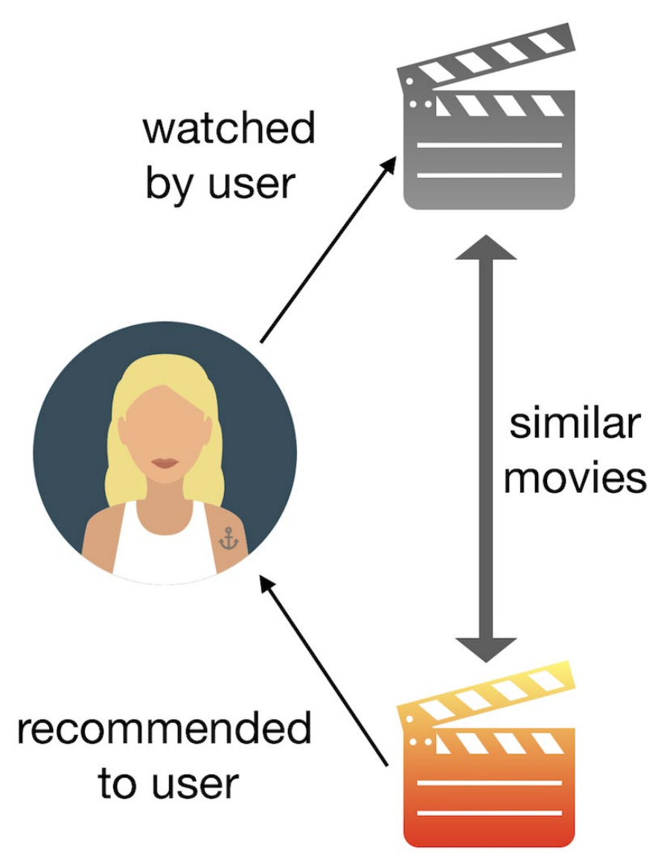
Content-based Movie Recommender

Ying Cai & Chong Meng

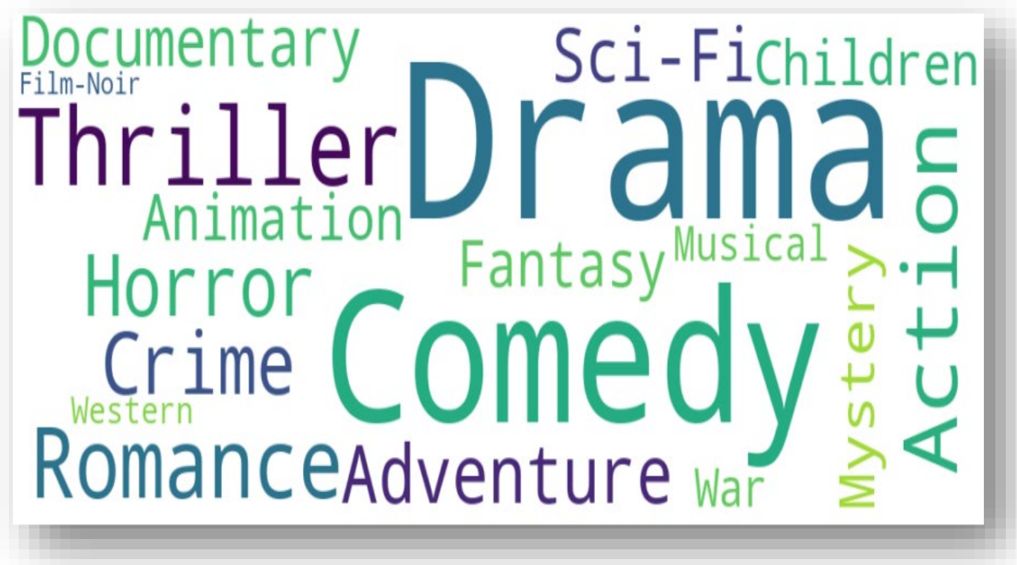
Introduction



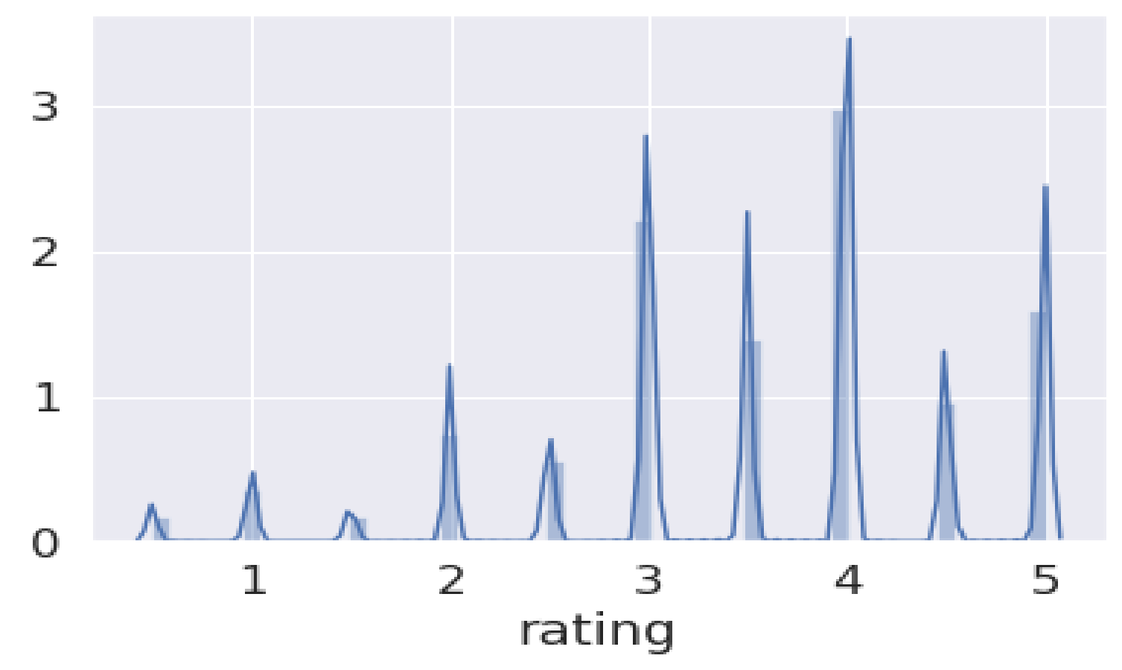
- ❑ The main goal of our project is to build a movie recommendation system based on movies' content.
- ❑ Our recommender works for following two kinds of recommendation.
  a) Recommend movies to user based on their personalized preference.
  b) Uncover potential audiences for a new movie advertisement.
- ❑ The dataset we are working with is MovieLens. It contain users' watching lists, ratings and movies' information.

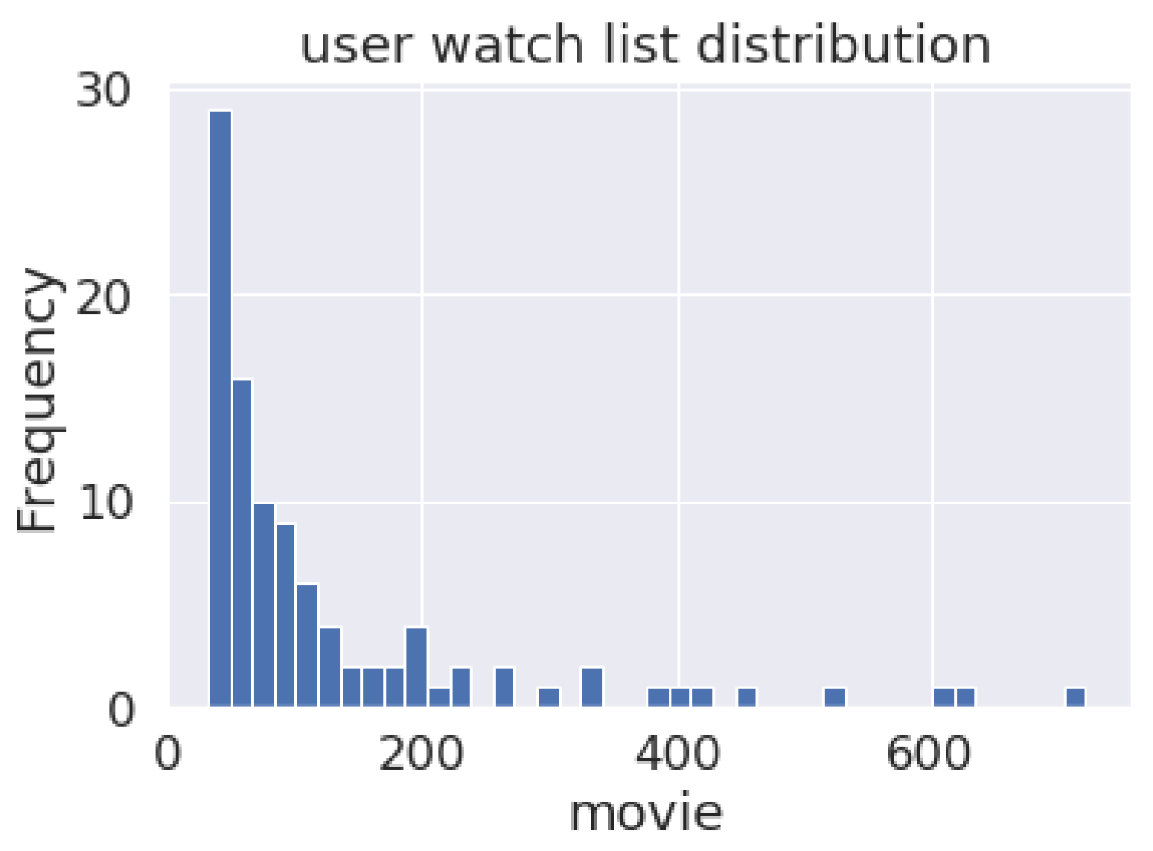
Data Preprocessing and Movie Features

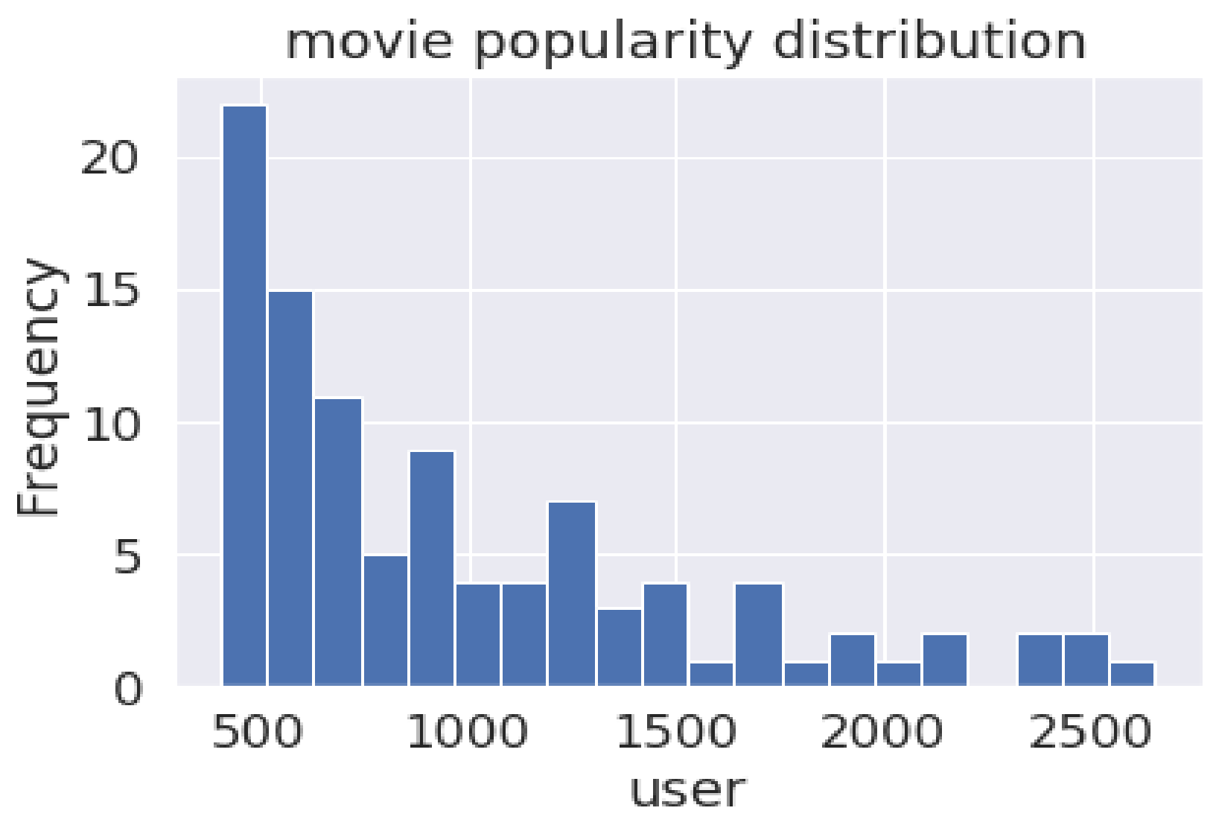
The original dataset contains about 27,000,000 ratings applied to 58,000 movies by 280,000 users. We refine the raw data by eliminating extreme cases (movies and users with tiny amount of records). After this step ,the dataset contains 16,655 movies, 153,524 users and 26,000,000 ratings.



word cloud of genres







Algorithms and Measurements

- ❑ "**Cosine similarity** is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them."*

$$\sigma_{ij} = \cos\theta = \frac{\sum_k A_{ik}A_{kj}}{\sqrt{\sum_k A_{ik}^2}\sqrt{\sum_k A_{kj}^2}} = \frac{n_{ij}}{\sqrt{d(i)d(j)}} \in [0,1]$$

- ❑ **KNN, the k-nearest neighbors algorithm.** It is "a non-parametric method used for classification and regression. The "input consists of the k-closest training examples in the feature space."*
- ❑ **User Preference and Movie Popularity.**
- ❑ **Precision and Recall.**

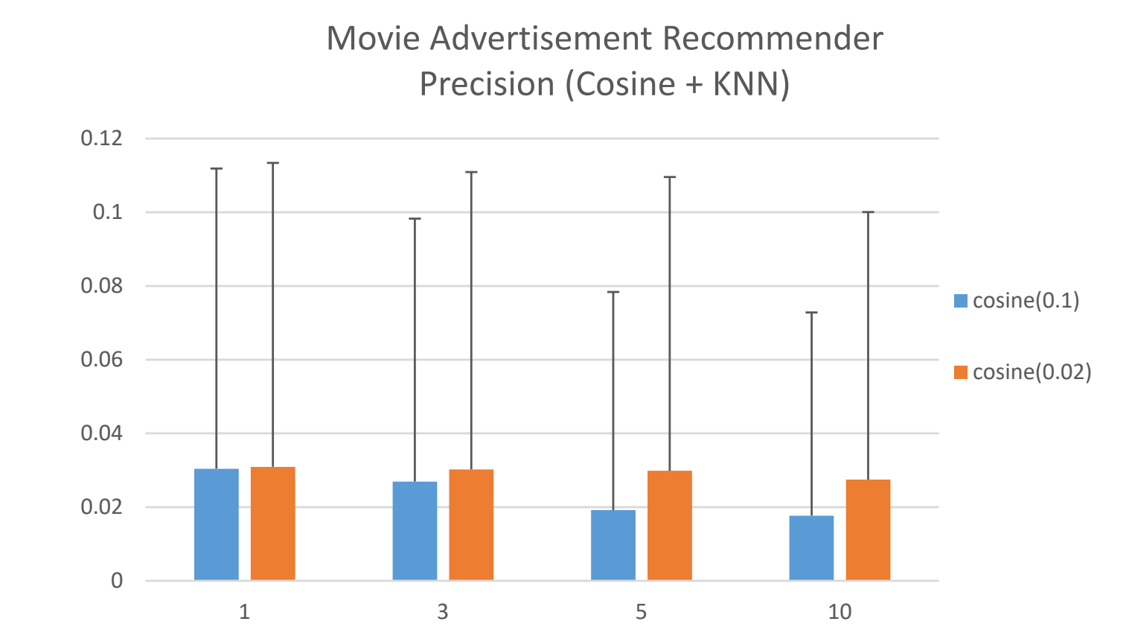
Precision is the fraction of relevant instances among the retrieved instances.

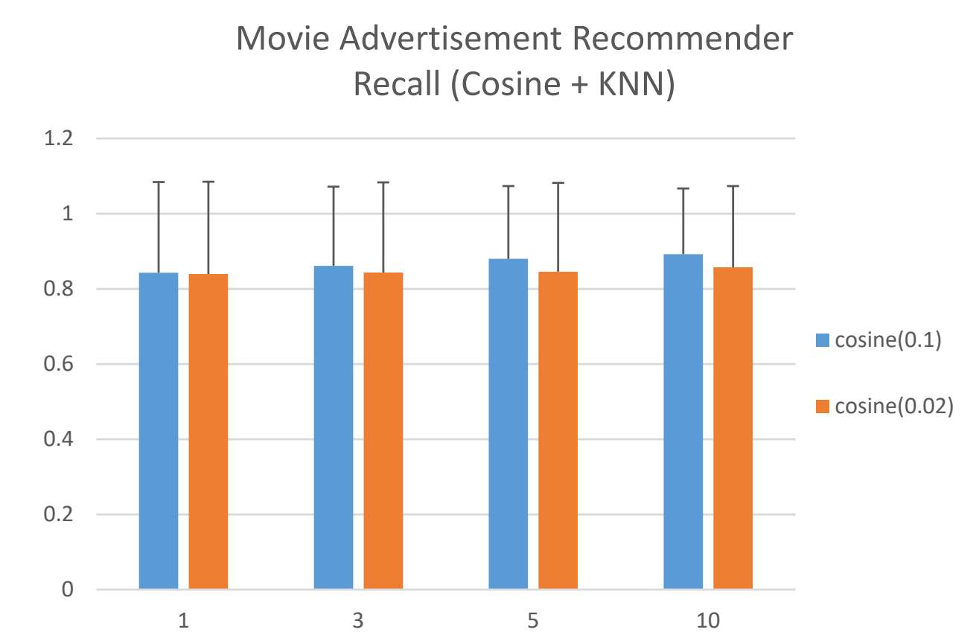
Recall is the fraction of relevant instances that have been retrieved over the total amount of relevant instances.

* From Wikipedia

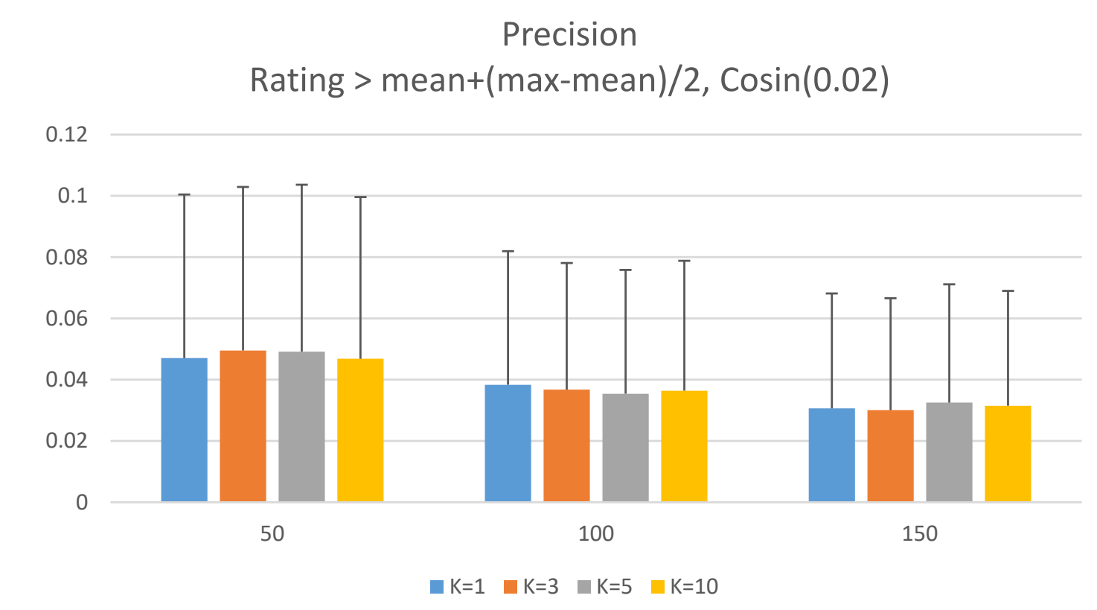
Parameter tuning and Results

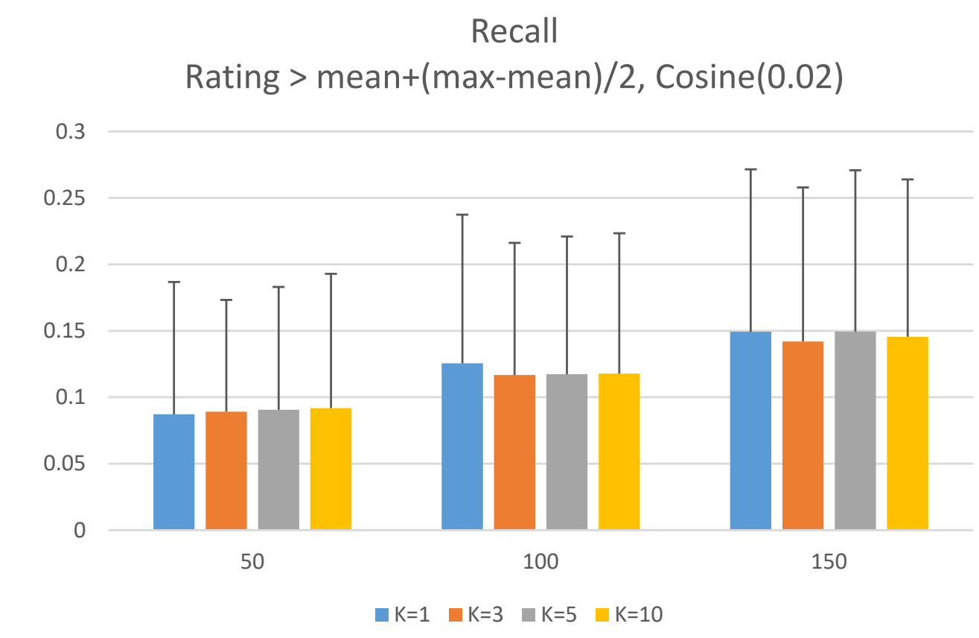
- ❑ ***Part A***. Uncover potential watchers. Choose k nearest neighbors, based on movie-movie Cosine similarity.

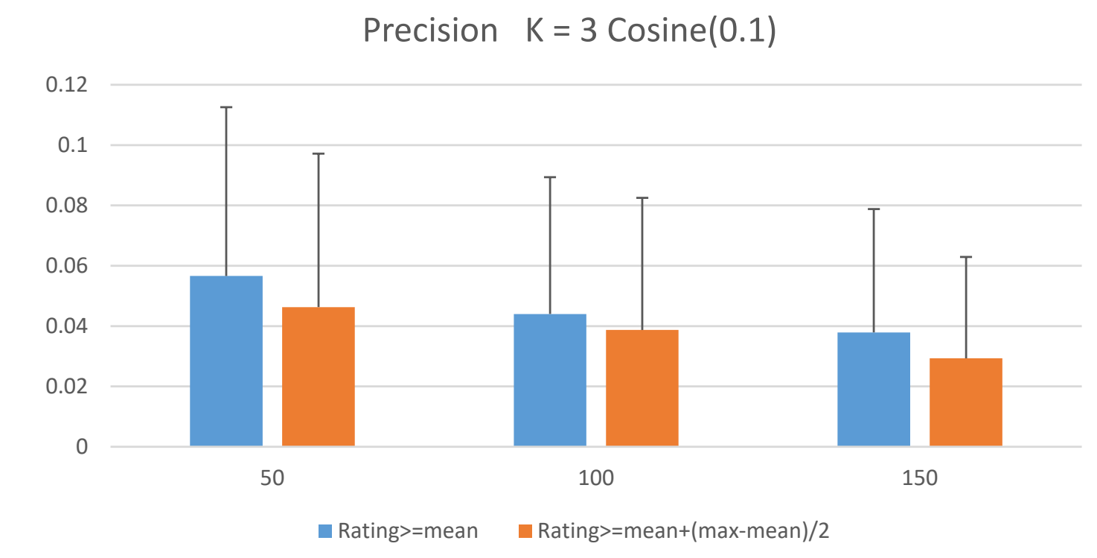


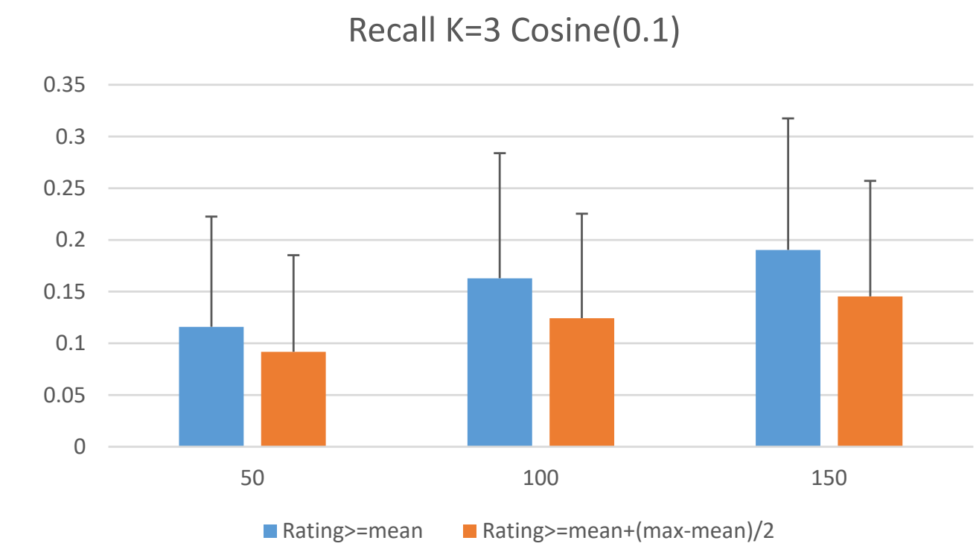


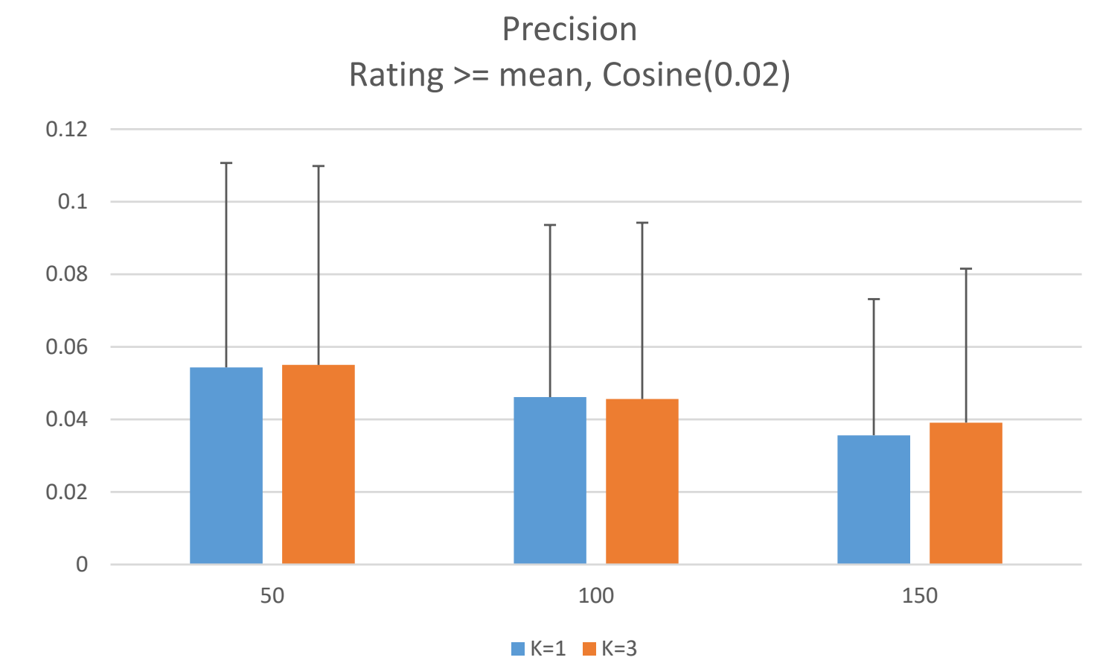
- ❑ ***Part B***. Recommend movies to the user based on the watching history. Choose the movies with higher ratings from the list, find similar unwatched movies by KNN, rank the list by movies' popularity, and then generate the recommendation list.

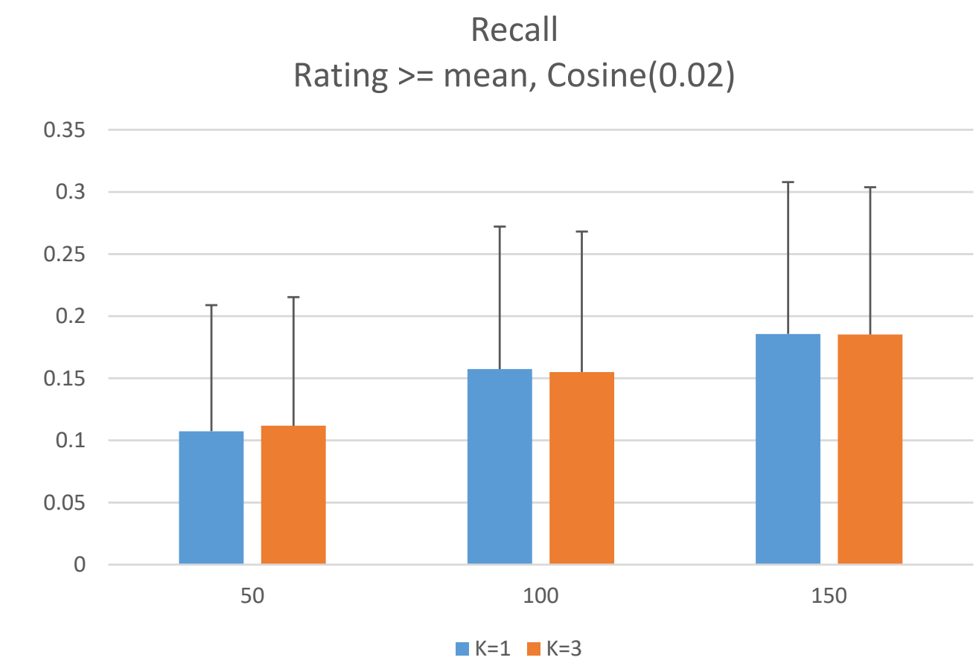






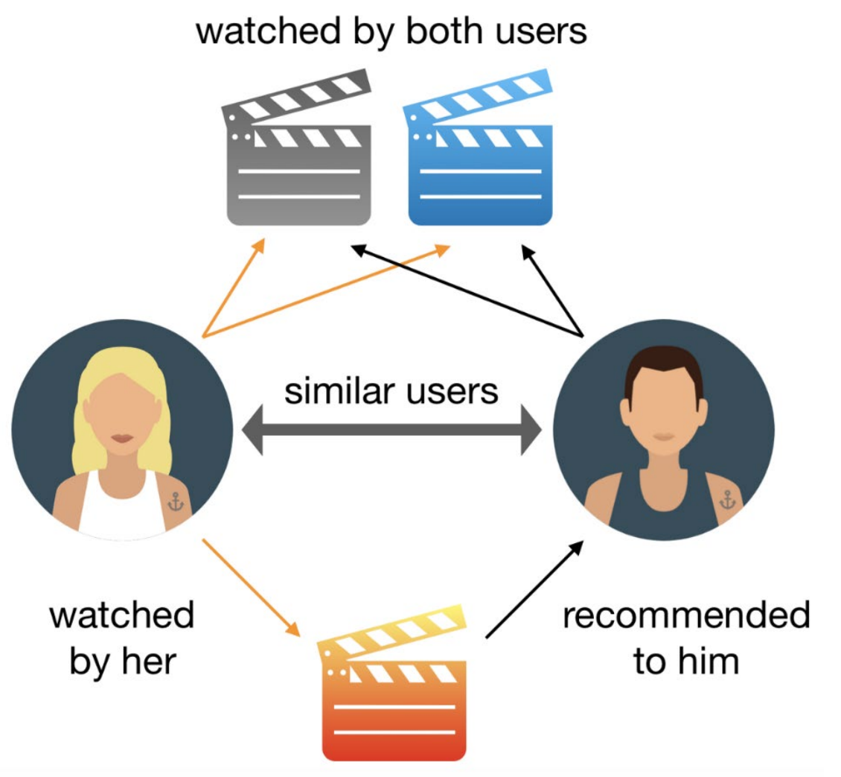






Discussion and Future Work

- ❑ Popularity is important for this recommendation system.
- ❑ The major purpose of Experiment A is finding potential audiences, thus the poor precision is not so unacceptable. Advertiser may prefer a higher recall. We now only have movie genre feature, more features are needed to further distinguish similarity and improve our model.
- ❑ Some users have very short watching lists and some movies have very few watchers. Both distributions have long tails. This can partially explain our big standard errors.
- ❑ We tried the collaborative filtering recommender system which is based on user-user similarity. The computational cost is so huge that we are still working to find an efficient way to generate a list.


watched by both users
similar users
watched
by her
recommended
to him