**Building a movie recommender system**

Cosine similarity is a method to measure the difference between two non-zero vectors of an inner product space.

For building a movie recommender system based on users’ ratings to films and the genres of these films, the data set contained > 100.000 movie ratings, produced by nearly 600 users and included 9.000 movie titles.

Preprocessing

After uploading the data I checked for duplicates in movies’ Ids and movies’ titles. The set is almost neat. I only found 5 duplicated movie ids values and removed the ones that had the lowest count of user ratings. I dropped the unnecessary “timestamp” column and records with Nan values. Only the first watched 200 films by first 200 user ids were selected for building the system.

Building model

Item-Based Recommender: First built a function to calculate the similarity of items (In this project, each item is represented by a movie) based on ratings only. For example, in order to calculate the similarity between Iron Man (2008) and Iron Man 2 (2010) I created two vectors with all the reviews for each movie and then calculated the vector’s cosine similarity. The vector contains several 0 values to fill in null values when a move was not rated by a user.

 Item-Based and Genre Recommender: In the previous model, two films could get similar ratings but if we look into each film’s genre we find that they are slightly different. So, I added a new layer to the recommender. I generated another function to add a new column containing the genre cosine similarity. Then the final recommendation is done based on the mean of similarity values, the genre and the rating.

Results

The first 10 recommended films similar to “Toy Story” displaying the cosine similarity, genre similarity and their mean:



And the recommendations for “Waiting to exhale”:

