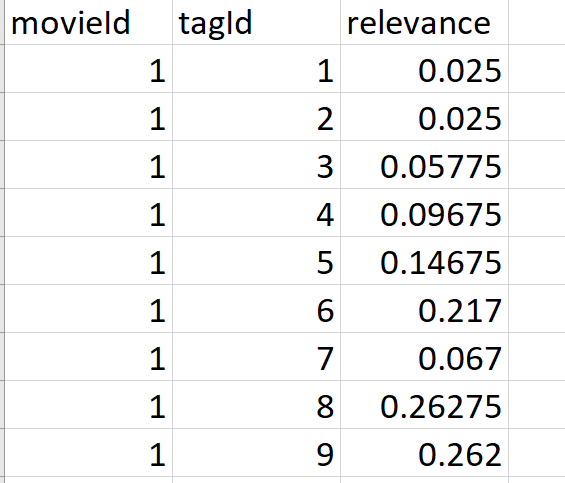
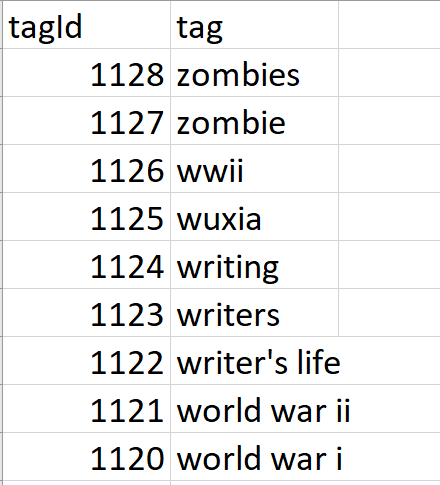
User generation information:

genome-scores.csv: movie id, tag id, relevance, indicating that for each movie-tag pair, the strength of this pair. Tags are user-generated metadata about movies. Each tag is typically a single word or short phrase. The meaning, value, and purpose of a particular tag is determined by each user. This data document should be seen together with genome-tags.csv.



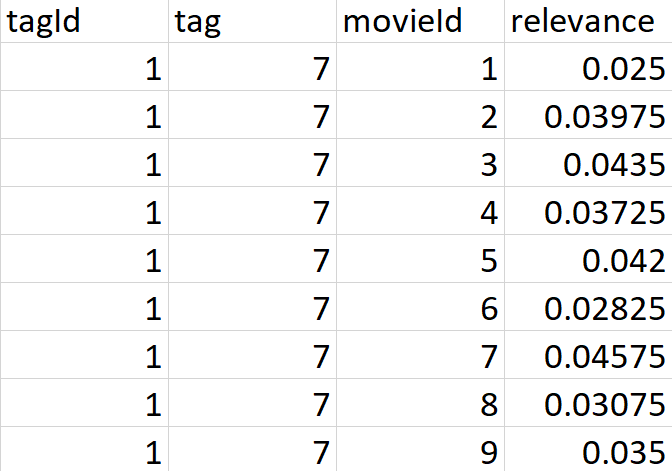
And genome-tags.csv: tagid, tag, indicating that each tagid corresponding to a specific tag. These tags are customers’ perspective which are personal words and not several specific types like movie genres. There are 1128 types of tags they use to describe the movies.



tags\_shuffled\_rehashed.csv: userId, movieId and tag, indicating that each user assigns a movie a tag. This movie id corresponds to the genome-scores and genome-tags which indicate the quality of the tags which a user identified. And each tag could be directly related to a user and a movie with score of it. The original order maybe ramdon compared to other csv files.



Because of the limitation of merge, I cannot combine all three csv files into one and I combined genome-scores and genome-tags into **movie\_tag\_Combin.csv**. The situation is every tags could assign to multiple movies, like tag1(7) who assigns to 10380 movies. Besides, one movie has multiple tags.



Official information:

movies.csv: movieid, title, genres, indicating each movie's name and corresponding genres.

Test:

Test-rating: The user id and movie id are connected, and we need to generate the user will like it or not to add a column after them.

test\_ratings.csv: test dataset. Please follow the format of the sample submission to submit your answer to this test set.

train\_ratings.csv: training dataset. It consists of three fields: userId, movieId, and rating (binary: 0 indicating dislike and 1 indicating like).

val\_ratings.csv: validation dataset with the same fields as the training dataset.

I combined train\_rating and movies.csv together into **rating\_movies.csv**.

