Collaborative Filtering

Asal Rahbari – 99222044

For applying collaborative filtering method on the beauty products dataset, there are a few steps that we have to do. For this dataset I'm going to have a user-based collaborative filtering, meaning that I'll find the users that are similar to each other based on their likings and taste (users that had have similar rating patterns for same set of products). So, whenever a user asks for recommendation based on their similar users we kind of predict how they would've rated a product , if they had have the chance to.

1) Loading Dataset:
   There's nothing new about this step, we just try to understand the features of our dataset like its size, columns, number of unique values of each column, and also the conceptual characteristics of each features like timestamp.
2) Null Values:
   Another essential step that we're totally familiar with is checking if there are any null values in our dataset or not. In this data it seems that there are no null values and therefore there's no need to handle them or fill them in.
3) Data processing:
   In this step we should make any required changes on our data, to convert them into a form that can be effectively and easily used for our collaborative filtering.
   The "Timestamp" column is not in the form of date time, instead it's in the unix format. So, the first thing to do is to convert this column's values to the data time format, which we can understand it.

By looking at our dataset’s size, we clearly realize that its large size may cause problem in the next steps. In this condition, we can get an appropriate sample of our data, and use the sample for collaborative filtering.

Some points about sampling that we should definitely consider while getting sample of our data are:

a) Adequate sample size: Our sample shouldn’t be neither too small nor too large compared to the size of our dataset. The reason is that if it’s too small then the output that we get from that may not be as accurate as we need to be, hence the result cannot be generalized to the whole dataset

b) Adequate sampling method: There exists various types of sampling, but they can’t be always appropriate for any dataset. In other words, we have to choose among all of the available method based on the features of our dataset. For instance, our dataset contains a column, in which the timestamp of each rating (row of the dataset) is stored. Its seems that the best way to get a sample of our dataset here is to make use of the timestamp column, and do the sampling based on that.

   Since we converted timestamp into understandable data time format, we can easily so the sampling based on that, and get the records related to each year as a sample.

   In our dataset years from 1998 to 2014 are included, so we will have 16 distinct samples that we can apply the collaborative filtering on each of them.

4) User Similarity:

   For finding the users similar to each other we use the cosine method here.

Cosine similarity measures the cosine of the angle of two vectors and by that it somehow measures the similarity between two vectors.

After finding the similar users, we use them to predict the ratings for the products that the user had not rated and they stand a chance for that specific user to like them. In this step we choose that how many of the similar users must be taken into account in the process (I set top 5 similar users).

The final step is to print top 10 recommendation found in the previous part and give them as an output to user.

The End