CS579 Project 2 weekly sample report

Bhanuja Arekatla(A20449753)
Navya Medarametla(A20442648)

WEEK-2 Date 11/05/2020

1. Your progress this week

   The 2 tables that we considered in our analysis are review and restaurant tables. In the restaurant table, important columns to be considered are restaurantID, name and categories. Whereas, in the review table, the important columns to be considered are reviewerID, restaurantID and rating.

   We are working with restaurant and review data but we named the restaurant as item to more generic to all use-cases, similarly with the other columns too.

```
user_col = 'reviewerID'
item_col = 'restaurantID'
value_col = 'rating'
name_col = 'name'
categories_col = 'categories'
```

   OS module in python provides functions for interacting with the Operating System.We have used os.path module which is a sub-module of OS module used for common pathname manipulation.os.path.join() method to join one or more path components.

```
rating_path = os.path.join('C:/Users/navii/OneDrive/Desktop/Fall 2020 SEM-3/OSNA/yelpResData', 'review.csv')
```

```
item_info_path = os.path.join('C:/Users/navii/OneDrive/Desktop/Fall 2020 SEM-3/OSNA/yelpResData', 'restaurant.csv')
```

But, the length of df_raw which is the dataframe for rating_path is 788471
And for item_info_path is 242652.

   As the number of entries in the tables is too large , we are only considering the entries with rating greater than or equal to 4

```
df_rating = df_raw[df_raw[value_col] >= 4.0].copy
```

   Sparse matrices come up in observations that record the occurrence or count of an activity. Compressed Sparse Row(CSR) algorithm is one of the types of Sparse matrices provided by Scipy.

```
<31647x175434 sparse matrix of type '<class 'numpy.float32'>'
        with 506402 stored elements in Compressed Sparse Row format>
```

2. Your plan for next week.
   Given this dataframe we will use the reviewerId, restaurantId and rating to perform the random train/test split (we can split based on the time if preferred) and feed the training set into a collaborative filtering based algorithm to train the model, so we can generate item recommendations for users.

3. The problems you encountered and how you solve them

We tried merging df_rating and df_item and both dataframes have column name "rating", Since same column existed in both dataframes, it's automatically renamed as rating_x and rating_y, so when we tried to print(df_rating[value_col]) we got a keyError:rating because the value_col equals rating but not rating_x.
So we resolved that with Suffixes parameter with the right outer join on merge (df_rating and df_item)so as shown below:

```
df_rating = df_rating.merge(df_item,on=item_col,suffixes=('','_y'))
```

4. Teamwork contribution

   NavyaMedarametla: worked on retrieving values from database, data preprocessing (reduced len(df) from 788471 entries to 242652 entries)and calculated the sparse matrix using CSR .

   Bhanuja Arekatla : Resolved the problem encountered with same column name in both tables and performed merge operation on 2 data frames and performed cosine similarity .