Viral Patel
Ketul Patel
Madeeha Khan

PROJECT REPORT-2

Project Title: Artist Recommendation System

Team Members: Madeeha Khan, Ketul Patel, Viral Patel

Date: 4/18/2018

Tasks Completed so far:

1) We have collected 1500 tweets for location Albany, New York using Twitter Rest API.
2) Merged generated dataset files and converted into a single data file(api-data.txt) file.
3) After that, we removed duplicate tweets using Python script removeDuplicatetweets.py, which will generate file having no duplicate tweets.
4) Then, just for our knowledge we created sentiment_analysis.py file to generate the positive and negative tweets in postive.txt and negative.txt file.
5) Next, to remove excess data, we did data processing on the tweets using preprocessor.py and generated the fruitful data file named preprocessed.txt.
6) We created one script file named scriptfromTweets toCSV.py can convert any text file to csv and output as output.csv.

(Checkpoint-1 Completed.)

7) We have labeled the tweets collected as positive and negative and stored them in 'labeled_tweets.txt'
8) Then, we found the top 10 features after removing the trivial words, spaces, alpha-numeric characters, etc.
9) After applying SVM classifier and 10-fold cross validation on the labeled dataset, we got 'predicted_tweets.txt', which has predicted class labels for the tweets which had not been labeled manually using script svm.py.
10) For finding association rules, we have used alternate.py script to get the tweets in the proper format.
11) Script_to_get_dataInList.py is used to get data from the dataset file as a list and stores it into Music_Words.txt
12) Python script association_rule.py produces association rules of the dataset and stores the association rules to the file named "result.txt"

Task Completed By Madeeha :

- Task no. (2) ,(6), (7) and (10)

Task Completed By Ketul :

- Task no. (1), (4),(9) and (11)

Task Completed By Viral :

- Task no. (3) ,(5),(8) and (12)

- Following is the link of google drive having collected data: https://drive.google.com/open?id=1z2OVLnKbeGAXrdhVHWYFuyopnK3YpBMi

Challenges We Have Encountered:

- While collecting tweets, due to tweeter API limit, we had to wait for 15 minutes to collect more tweets. To solve the issue, we wrote python script so that we didn’t have to worry about it, the script would automatically run after 15 minutes and collect more tweets.
- Finding Regex for removing a URL and extra characters from tweets took a long time and required extra efforts.