TWITTER – USERS' CONTENT RECOMMENDATION SYSTEM USING TWEETS

A PROJECT REPORT

Submitted by

SHREENIDHI S 2017103620
PAVITHRA KARTHY 2017103059
SUPRADEEPA VELLA 2017103628

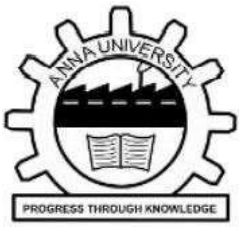
in partial fulfilment for the award of the degree

of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING



COLLEGE OF ENGINEERING, GUINDY

ANNA UNIVERSITY: CHENNAI 600 025

NOVEMBER 2020

TABLE OF CONTENTS

1. PROBLEM STATEMENT

The aim of this project is to build a recommendation system that can help an organization to target their customers based on their specific interests. The main objectives of our recommendation system include:

- Sentimental analysis of individual users
- Visualization of individual user interests
- Grouping of users with same interests
- Comparison of users with same interests

2. ABSTRACT

With the advent of the internet into our everyday lives, online social networks such as Facebook and Twitter have taken up a major role in networking, information deployment and entertainment. As of 2017, Twitter's outreach is over 317M monthly active users generating more than 320M tweets everyday, thus making it one of the fastest information deployment mediums of this era. In order to aid data distribution without causing a glut of information to the users, we develop a recommender system focusing on a vital aspect of most preferred interests of the individual users. The information collected from the most recent tweets of a user is used to find other users whose recent tweets contain similar information, and grouping of users with common interests are also done. By making use of the continuous and real time updating of data on social networks, we develop a method to ensure our training sets consist of relevant information for classification, thus preserving accuracy while reducing training set sizes for probabilistic learning models. We use two methodologies to detect tweets of common topics, namely a Lematization and a Cosine similarity Classifier and further compare their complexity and accuracy.
The information collected depicts the percentage of the most tweeted and less tweeted topics users are interested in and then grouping other users with common ideas and interests in common categories and the individual interests of the users are shown in the pie chart.

The aim of this project is to build a recommendation system that can help an organization to target their customers based on their specific interests. This can help the organization in forming effective marketing strategies to attract customers. Users interests are gathered from their twitter account in the form of tweets, favourites etc. For collecting data, we used tweepy API to extract tweets, favourites etc. Then we classified the users interests into relevant categories with an accuracy of about 85% by using NLP algorithms. We also performed sentiment analysis of tweets, retweets and were able to achieve an accuracy of 90%. Individual user interests are displayed using pie chart built using matplotlib whereas overall user interests are displayed using bar graph built using bokeh.

3. BASE PAPER

1) Short and Tweet: Experiments on Recommending Content from Information Streams

Authors: Jilin Chen

LINK:https://hci.stanford.edu/publications/2010/zerozero88/zerozero88-chi2010.pdf

2) Twitter-User Recommender System using Tweets: A Content-based Approach

Authors: Nidhi R.H, Annappa B

Link: https://ieeexplore.ieee.org/document/8272631

4. ABOUT CODE

FULL CODE:
https://github.com/snidhi99/TWITTER_SENTIMENT_ANALYSIS/blob/master/Recommendation_System_User_Interests(1).ipynb

The code is not the same as the research paper.

Research Paper idea

- In the research paper, the users are categorized based on their popularity and the information collected from the most recent tweets are used to find other users whose tweets contain similar information.
- Naïve bayes classification is done
- So, it is a user recommender system based on tweets

Our Project idea

- In our project, we are categorizing the users based on their specific interests, i.e, based on the content that they tweet and are visualizing each users' interest through a pie chart.
- Ultimately, the most tweeted ideas correspond to their category interests.
- This is what we would recommend, by eventually grouping them under interested categories.
- Additionally, sentimental analysis followed by lemmatization and cosine similarity, is performed

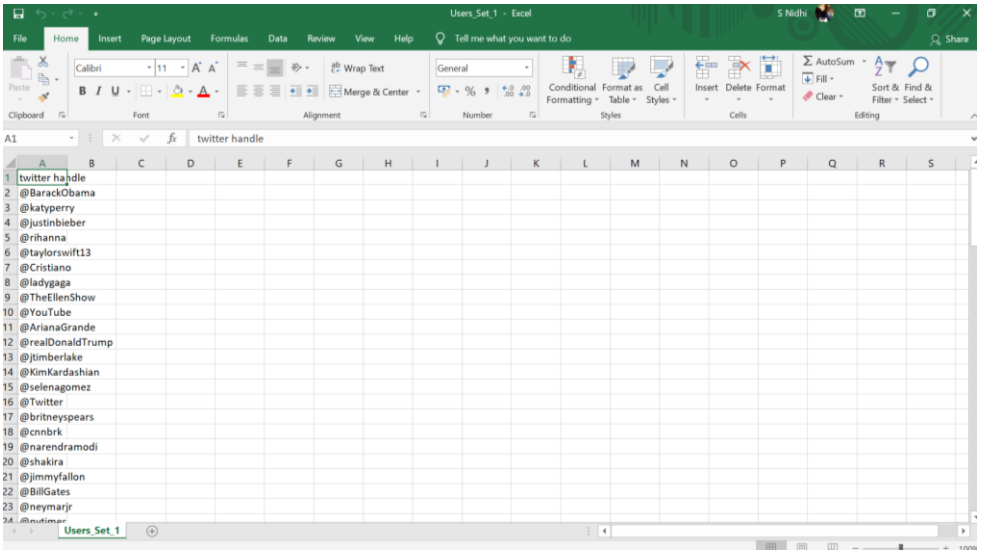
5. MODIFICATIONS DONE

- Sentimental Analysis – Classification of tweets, retweets and favorite tweets as positive and negative
- Getting Topwords and categories – identifying highest frequency words and putting them under their respective categories
- Bag of words - All the unwanted words and articles are removed, making it easier for us to identify the important keywords
- Lematization - process of grouping together the inflected forms of a word so they can be analysed as a single item, identified by the word's lemma, from the categories list

- Cosine Similarity - where we will have to convert the word to its vector form, and we will also have to convert the word in the categories list to its vector form

6. DATASET

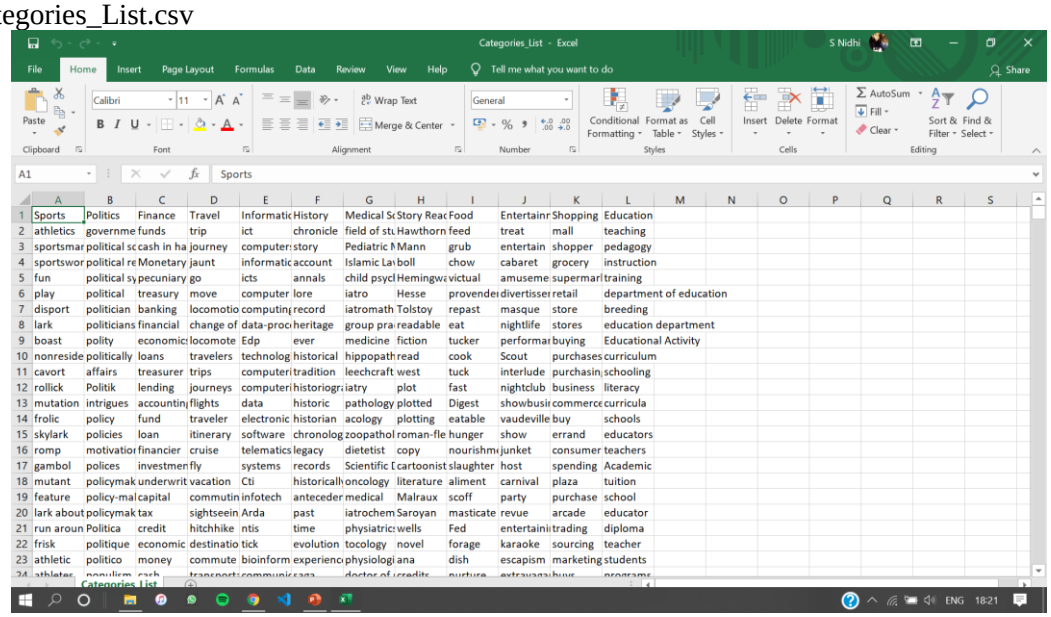
(1) User_Set_1.csv



DESCRIPTION:

- Twitter handles (User names) – Provides a list of all popular twitter users

(2) Categories_List.csv



DESCRIPTION:

- The various categories are the independent columns
- Each column has a list of all common verb forms
- Cosine similarity will be checked if the word matches with any of the words of the list

7. METHODOLOGY

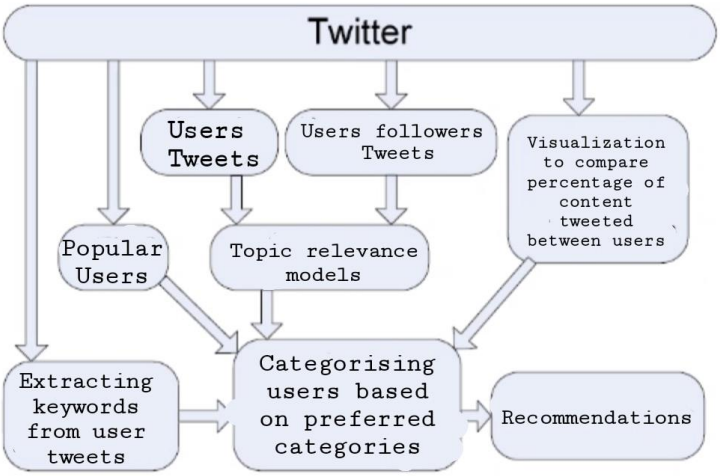


Figure 1. Conceptual Model of the Whole Recommender

(1) API AUTHENTICATION

- Generated data from twitter streams and then converted into a csv file
- Importing Data From Pre-Defined Datasets And Writing The Output Recommendation To A New Csv File

(2) FUNCTION TO EXTRACT TWEETS, RETWEETS AND FAVOURITE TWEETS

- The API.retweets() method of the API class in Tweepy module is used to return a list of retweets of a tweet.
- Using the retweets() method with count parameter to only fetch a certain number of retweets.

(3) FUNCTION TO PERFORM SENTIMENTAL ANALYSIS OF TWEETS

- The tweets are classified as positive/negative

(4) TO GET TOPWORDS

- Implementation of a function that gets the highest frequency words from the unigram, bigram and trigrams list

(5) FIND WHAT CATEGORY THE WORDS WILL BELONG TO

- Implementation of a function that helps to keep track of the number of words that belong to specific categories

(6) CLASSIFY WORDS BASED ON CATEGORIES/PERCENTAGE OF OCCURANCE

- categoryWisePercent function that classifies words belonging to each category and calculates categoryWise percentage

(7) PLOT USERS' RECOMMENDATIONS AS PIE CHARTS

- Visualization in the form of pie charts that shows the classified category interests of the users

(8) FUNCTION TO GET UI INTERFACE TO LOAD AND EXECUTE DATASET

- We are using the tkinter function which is a Python binding to the Tk GUI toolkit. **Tkinter** provides a powerful object-oriented interface to the Tk GUI toolkit. We will be able to load the input dataset and execute the command to get the desired results.

8. CONFIGURATION PARAMETERS

The stream tweets are extracted from the twitter API. For streaming the tweets using the twitter API, the keys are required as the tweets cannot be streamed withot the keys. Thus the keys are the primary configuration. In order to determine the polarity of the tweets, the module for the Vander polarity analyser is also important.

9. RESULTS

Sentimental Analysis of classification of tweets of users as positive/negative with the help of stopwords, wordcloud

(1)Positive Tweets (content)

```
In [ ]: wordcloud = WordCloud(stopwords = STOPWORDS, background_color = "white", max_words = 1000).generate(positive_sentiments)
        plt.figure(figsize = (12, 8))
        plt.imshow(wordcloud)
        plt.axis("off")
        plt.title("Positive tweets Wordcloud")

Out[ ]: Text(0.5, 1.0, 'Positive tweets Wordcloud')
```

(3)Negative Tweets (Content)

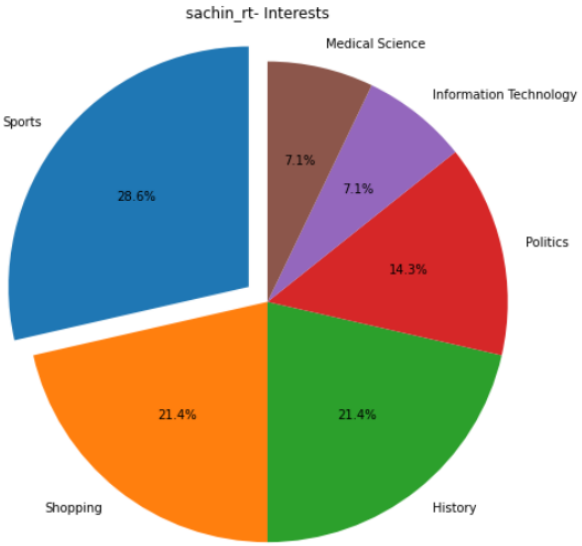
```
In [ ]: wordcloud = WordCloud(stopwords = STOPWORDS, background_color = "white", max_words = 1000).generate(negative_sentiments)
        plt.figure(figsize = (12, 8))
        plt.imshow(wordcloud)
        plt.axis("off")
        plt.title("Negative tweets Wordcloud")

Out[ ]: Text(0.5, 1.0, 'Negative tweets Wordcloud')
```

Sample pie charts of different users and their category interests are depicted:

(1)

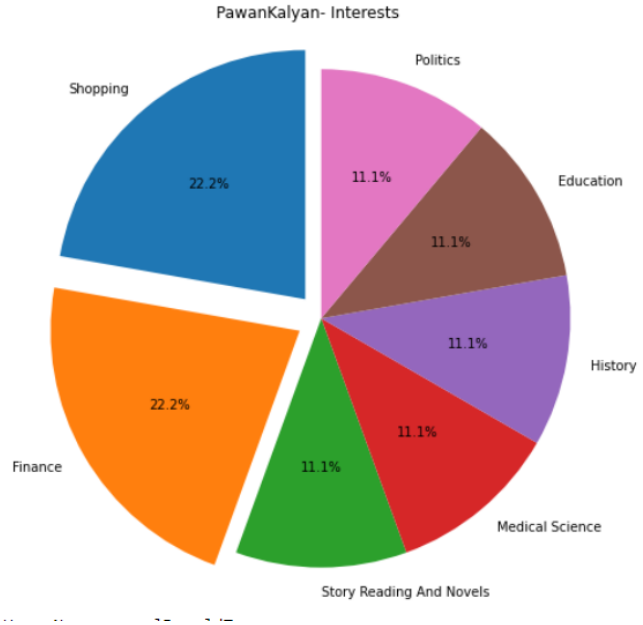
```
User Name : sachin_rt
                  Category  Words  Percentage
0                   Sports      4          28
1                 Shopping      3          21
2                  History      3          21
3                 Politics      2          14
4  Information Technology      1           7
5          Medical Science      1           7
```



(2)

User Name : PawanKalyan

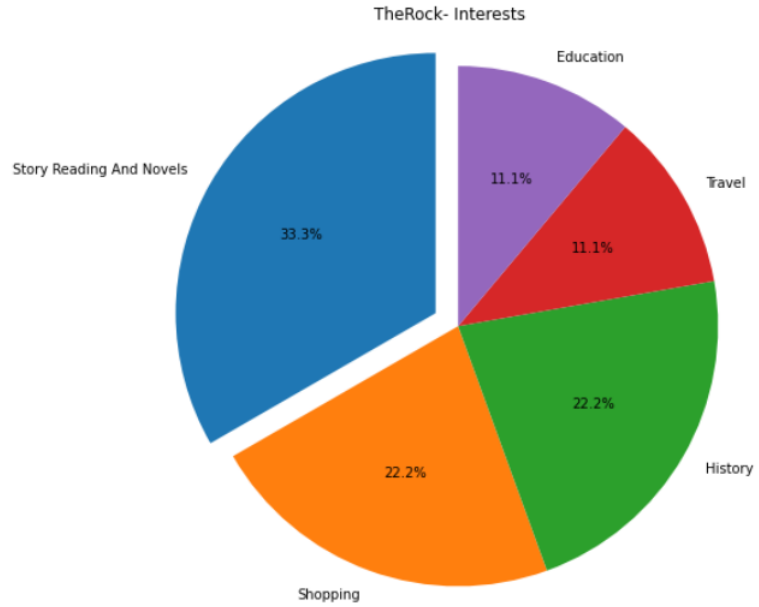
	Category	Words	Percentage
0	Shopping	2	22
1	Finance	2	22
2	Story Reading And Novels	1	11
3	Medical Science	1	11
4	History	1	11
5	Education	1	11
6	Politics	1	11



(3)

User Name : TheRock

	Category	Words	Percentage
0	Story Reading And Novels	3	33
1	Shopping	2	22
2	History	2	22
3	Travel	1	11
4	Education	1	11



Bar Chart to display individual users grouped into various categories. The number of users in each category and the total percentage of content used is tabulated.

The number of users' highest percentage of content tweeted, corresponding to each user's category interest will be recommended to the user.

	Categories	Number Of Users	Percentage	User Names
0	Sports	7	16	sachin_rt,virendersehwag,Cristiano,JustinTrudeau,realDonaldTrump,RaviShastriOfc,KP24
1	Shopping	6	13	sachin_rt,virendersehwag,Cristiano,JustinTrudeau,PawanKalyan,KP24
2	History	6	13	sachin_rt,virendersehwag,Cristiano,JustinTrudeau,realDonaldTrump,KP24
3	Travel	4	9	Cristiano,JustinTrudeau,realDonaldTrump,KP24
4	Story Reading And Novels	6	13	Cristiano,JustinTrudeau,lbjamesharden,realDonaldTrump,KP24,TheRock
5	Finance	3	6	JustinTrudeau,PawanKalyan,KP24
6	Food	2	4	JustinTrudeau,realDonaldTrump
7	Politics	4	9	JustinTrudeau,realDonaldTrump,KP24,PTTVOnlineNews
8	Entertainment	3	6	realDonaldTrump,KP24,PTTVOnlineNews
9	Education	1	2	realDonaldTrump
10	Medical Science	1	2	realDonaldTrump

10. EVALUATION

ALGORITHMS USED

- Content based filtering
- finding cosine correlation based similarities
- NLTK algorithms like stemming, lemitization