TWITTER DATA ANALYSIS

A Mini Project Report Submitted by

NEHA U	NIDHI RAI
(4NM18CS102)	(4NM18CS103)

UNDER THE GUIDANCE OF

Ms. MINU P ABRAHAM

Assistant Professor

Department of Computer Science and Engineering

in partial fulfillment of the requirements for the award of the Degree of

Bachelor of Engineering in Computer Science & Engineering

from

Visveshvaraya Technological University, Belgaum

N NITTE
EDUCATION TRUST

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

N.M.A.M. INSTITUTE OF TECHNOLOGY

(An Autonomous Institution under VTU, Belgaum) (AICTE approved, NBA Accredited, ISO 9001:2008 Certified) NITTE -574 110, Udupi District, KARNATAKA.

May 2021

NITTE
EDUCATION TRUST

N.M.A.M. INSTITUTE OF TECHNOLOGY
(An Autonomous Institution affiliated to Visvesvaraya Technological University, Belagavi)
Nitte – 574 110, Karnataka, India
(ISO 9001:2015 Certified), Accredited with 'A' Grade by NAAC
☎: 08258 - 281039 – 281263, Fax: 08258 – 281265

Department of Computer Science and Engineering

B.E. CSE Program Accredited by NBA, New Delhi from 1-7-2018 to 30-6-2021

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

"Twitter Data Analysis"

is a bonafide work carried out by

NEHA U(4NM18CS102)   NIDHI RAI(4NM18CS103)

in partial fulfillment of the requirements for the award of Bachelor of
Engineering Degree in Computer Science and Engineering prescribed
by Visvesvaraya Technological University,
Belgaum during the year 2019-2020.

It is certified that all corrections/suggestions indicated for Internal Assessment
have been incorporated in the report.

The Mini project report has been approved as it satisfies the academic
requirements in respect of the project work prescribed for the Bachelor of
Engineering Degree.

Signature of Guide   Signature of HOD

ACKNOWLEDGEMENT

We believe that our project will be complete only after we thank the people who have contributed to make this project successful.

First and foremost, our sincere thanks to our beloved principal, **Dr. Niranjan N. Chiplunkar** for giving us an opportunity to carry out our project work at our college and providing us with all the needed facilities.

We express our deep sense of gratitude and indebtedness to our guide **Ms. Minu P Abraham**, Assistant Professor, Department of Computer Science and Engineering, for her inspiring guidance, constant encouragement, support and suggestions for improvement during the course of our project.

We sincerely thank **Dr. K.R. Udaya Kumar Reddy**, Head of Department of Computer Science and Engineering, Nitte Mahalinga Adyantaya Memorial Institute of Technology, Nitte.

We also thank all those who have supported us throughout the entire duration of our project.

Finally, we thank the staff members of the Department of Computer Science and Engineering and all our friends for their honest opinions and suggestions throughout the course of our project.

<div align="right">

NEHA U(4NM18CS102)
NIDHI RAI(4NM18CS103)

</div>

TABLE OF CONTENTS

ABSTRACT

In this generation of digitalization, internet and social media has become an essential part of our society. On a daily basis social media is being used people to gain knowledge, get the news updates and present their opinions to others.

Twitter being one the apps has a large amount of tweets being exchanged on daily basis. This includes all kinds of information like economic, industrial, current affairs or governmental affairs by grouping the tweets into different categories as per customer demands. Since each category of information contains huge data it becomes a complex problem to store and process that huge data. We will be analyzing the huge amount of tweets in the project.

<u>INTRODUCTION</u>

Twitter has over a billion users and everyday people generate billions of tweets over 100 hours per minute and this number is ever increasing. To analyse and understand the activity occurring on such a massive scale, a relational SQL database is not enough. Such kind of data is well suited to a massively parallel and distributed system.

    The main objective of this project is to focus on how data generated from Twitter can be mined and utilized by different companies to make targeted, real time and informed decisions about their product that can increase their market share or to find out the views of people on a specific topic of interest.

    There are multiple applications of this project. Companies can use this project to understand how effective and penetrative their marketing programs are through sentiment analysis. In addition to it, companies can also evaluate the popular hash tags which are trending nowadays. Applications for Twitter data can be endless.

    This project can also help in analysing new emerging trends and knowing about people's changing behaviour with time. Also people in different countries have different preferences. By analysing the tweets/hash tags/sentiment etc., companies can understand what are the likes /dislikes of people around the world and work on their preferences accordingly.

DESIGN AND IMPLEMENTATION

The twitter dataset which is required to do analysis is downloaded and stored in the local system. Next step is to load this data into Hadoop Distributed File System (HDFS).

What is Big Data ?

Big data is exactly what it sounds like—a lot of data. Alone, a single point of data can't give you much insight. But terabytes of data, combined together with complex mathematical models and boisterous computing power, can create insights human beings aren't capable of producing. The value that big data Analytics provides to a business is intangible and surpassing human capabilities each and every day.

The first step to big data analytics is gathering the data itself. This is known as "data mining." Data can come from anywhere. Most businesses deal with gigabytes of user, product, and location data. In this tutorial, we'll be exploring how we can use data mining techniques to gather Twitter data, which can be more useful than you might think.

For example, let's say you run Facebook, and want to use Messenger data to provide insights on how you can advertise to your audience better. Messenger has 1.2 billion monthly active users. In this case, the big data are conversations between users. If you were to individually read the conversations of each user, you would be able to get a good sense of what they like, and be able to recommend products to them accordingly.

Using a machine learning technique known as Natural Language Processing (NLP), you can do this on a large scale with the entire process automated and left up to machines.

This is just one of the countless examples of how machine learning and big data analytics can add value to your company.

What is Twitter Data ?

Twitter is a gold mine of data. Unlike other social platforms, almost every user's tweets are completely public and pullable. This is a huge plus if you're trying to get a large amount of data to run analytics on. Twitter data is also pretty specific. Twitter's API allows you to do complex queries like pulling every tweet about a certain topic within the last twenty minutes, or pull a certain user's non-retweeted tweets.

A simple application of this could be analyzing how your company is received in the general public. You could collect the last 2,000 tweets that mention your company (or any term you like), and run a sentiment analysis algorithm over it. We can also target users that specifically live in a certain location, which is known as spatial data. Another application of this could be to map the areas on the globe where your company has been mentioned the most.

As you can see, Twitter data can be a large door into the insights of the general public, and how they receive a topic. That, combined with the openness and the generous rate limiting of Twitter's API, can produce powerful results.

Tools Overview

We'll be using Python 2.7 for these examples. Ideally, you should have an IDE to write this code in. I will be using google Collab .

To connect to Twitter's API, we will be using a Python library called tweepy.

CODE AND RESULTS

STEP 1:

In the first step of the project, a Twitter Developers Account is created. We have to have a Twitter Account Logged In. Once we created the account a project named TDA_BDA was created.
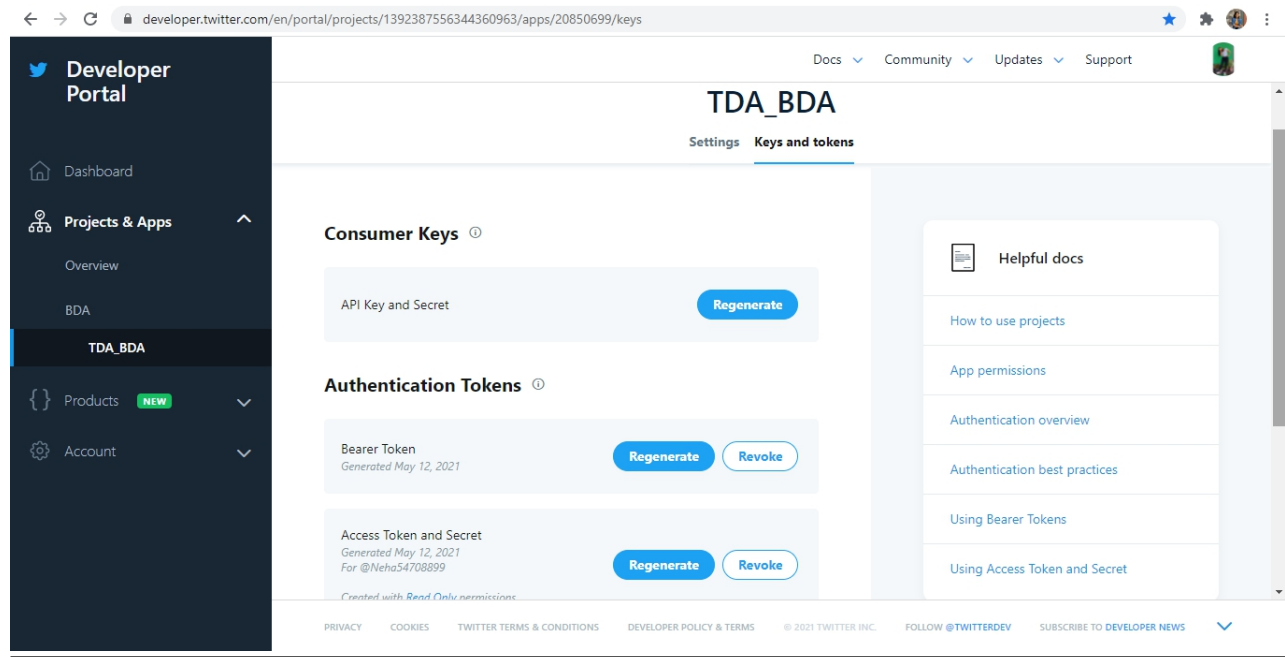
STEP 2: Save your App's key and tokens and keep them

Once you've been approved for developer access and have created a Project and App, you will be able to find or generate the following credentials within your developer App:

API Key - This is essentially a username, and allows you to make a request on behalf of your App.

API Key Secret - This is a password, and allows you to make a request on behalf of your App.

Access Token - This token represents the Twitter account that owns the App, and allows you to make a request on behalf of that Twitter account.

Access Token Secret - This token also represents the Twitter account that owns the App, and allows you to make a request on behalf of that Twitter account.

Bearer Token - This token represents your App and enables you to authenticate requests that require OAuth 2.0 Bearer Token authentication

Here the necessary packages that is needed to get the data in a structured way is imported and downloaded.

All the necessary authentication to the twitter account is done here. By mentioning the necessary keys. The next step is creating an OAuthHandler instance. Into this we pass our consumer key and secret. You can throw these into a database, file, or where ever you store your data.

To re-build an OAuthHandler from this stored access token you would do the above mentioned code.
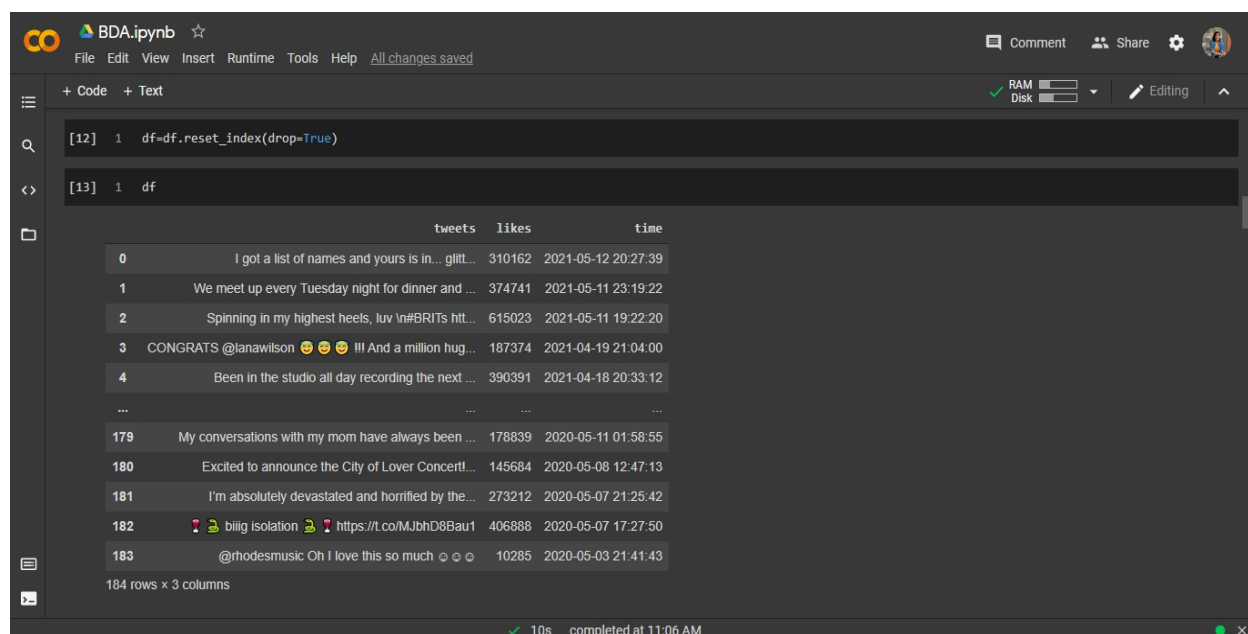
This code will give you the tweets likes and time of the twitter user that you have mentioned in the id. The Result will be as shown below:

The Below Code will give you out the data of the most liked tweets:

Now we have to see line by line and split the words. The 16th Code splits all the sentences up which makes it easier to work. So, now we have to standardize normalize the data. First, we have to remove all the punctuation using the regular expression using the 17th Code.

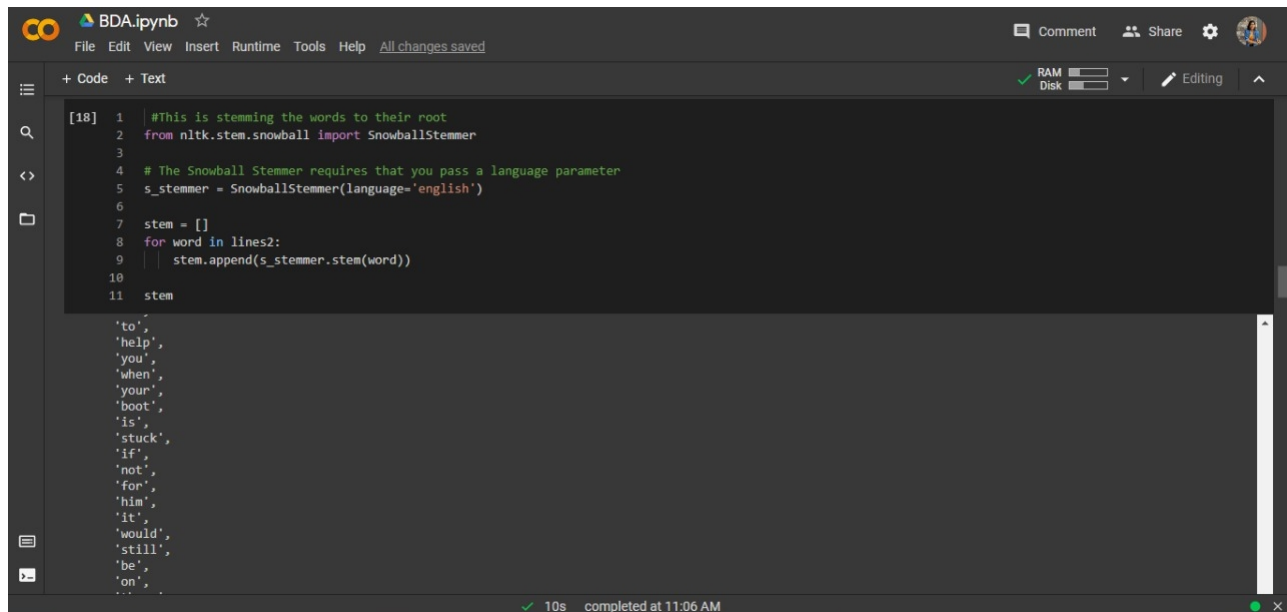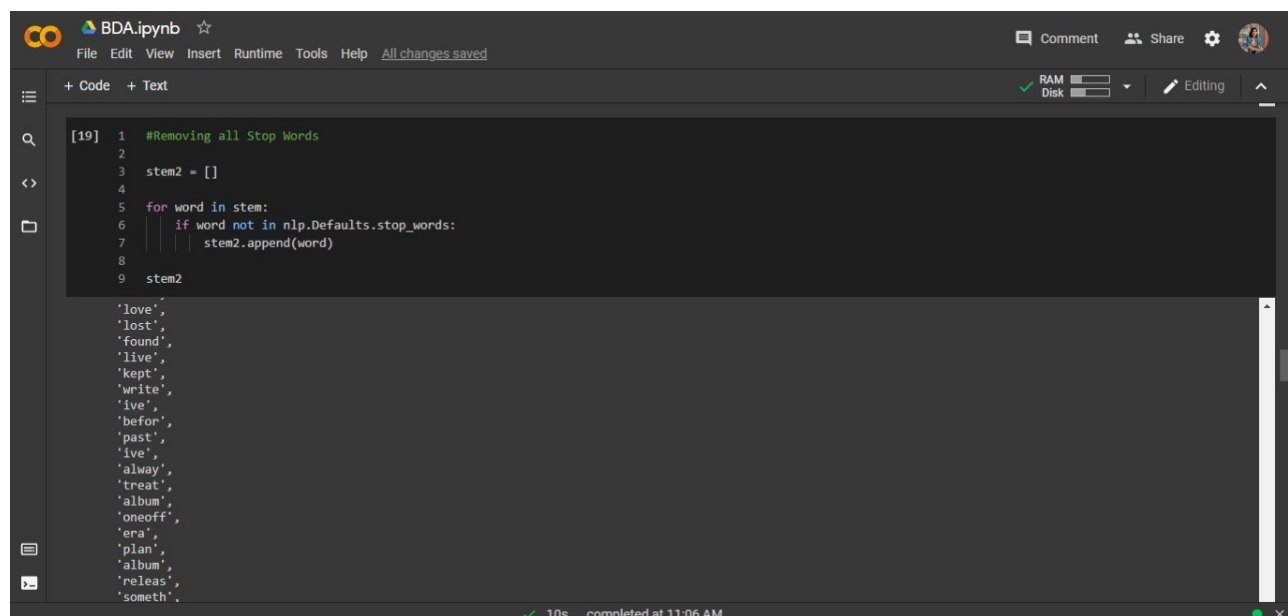

Then, steming - taking different variations from one word and breaking them in terms of stem or the root word.
For example - If run had different variations like running and runner we have to break it into root word run.

```
[18]  1   #This is stemming the words to their root
      2   from nltk.stem.snowball import SnowballStemmer
      3
      4   # The Snowball Stemmer requires that you pass a language parameter
      5   s_stemmer = SnowballStemmer(language='english')
      6
      7   stem = []
      8   for word in lines2:
      9       stem.append(s_stemmer.stem(word))
     10
     11   stem
```

```
'to',
'help',
'you',
'when',
'your',
'boot',
'is',
'stuck',
'if',
'not',
'for',
'him',
'it',
'would',
'still',
'be',
'on',
```

✓ 10s completed at 11:06 AM

Final step is to remove all the stop words. Stop words are really simple words like is, the, of, a since these words are repeated thousand times.


```
[19]  1   #Removing all Stop Words
      2
      3   stem2 = []
      4
      5   for word in stem:
      6       if word not in nlp.Defaults.stop_words:
      7           stem2.append(word)
      8
      9   stem2
```

```
'love',
'lost',
'found',
'live',
'kept',
'write',
'ive',
'befor',
'past',
'ive',
'alway',
'treat',
'album',
'oneoff',
'era',
'plan',
'album',
'releas',
'someth',
```

✓ 10s completed at 11:06 AM

The 21st Code gives out the Top 10 Words used by the user mentioned in the id.
The Graph will be generated using the Matplotlib (a low level graph plotting library in python that serves as a visualization utility.)
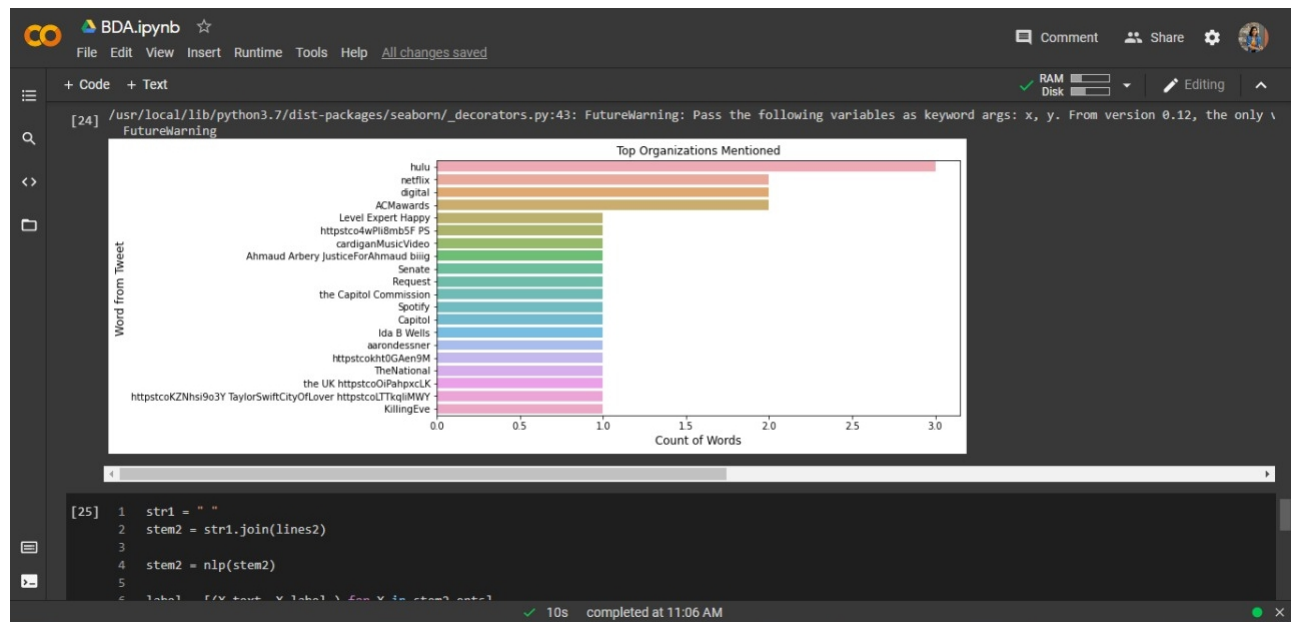
13
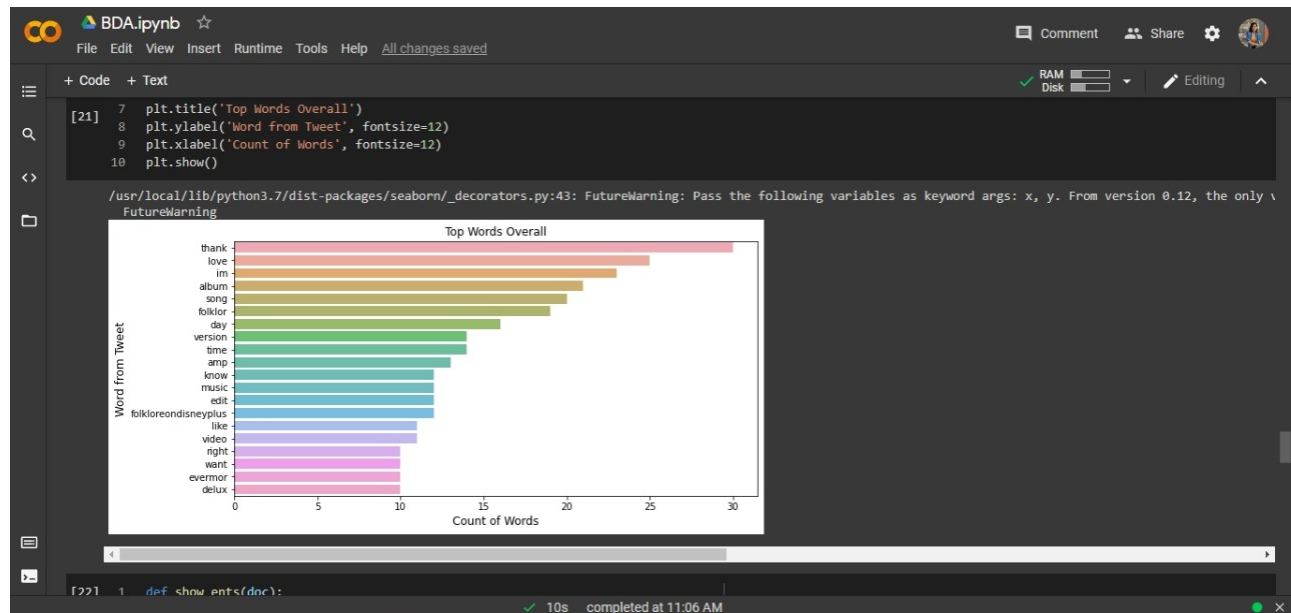
As we have installed spacy library which is used to break the words into categories like people, place, things, organizations etc.

Now next we are going see what type of organization's Taylor swift mentioned the most in her tweets.
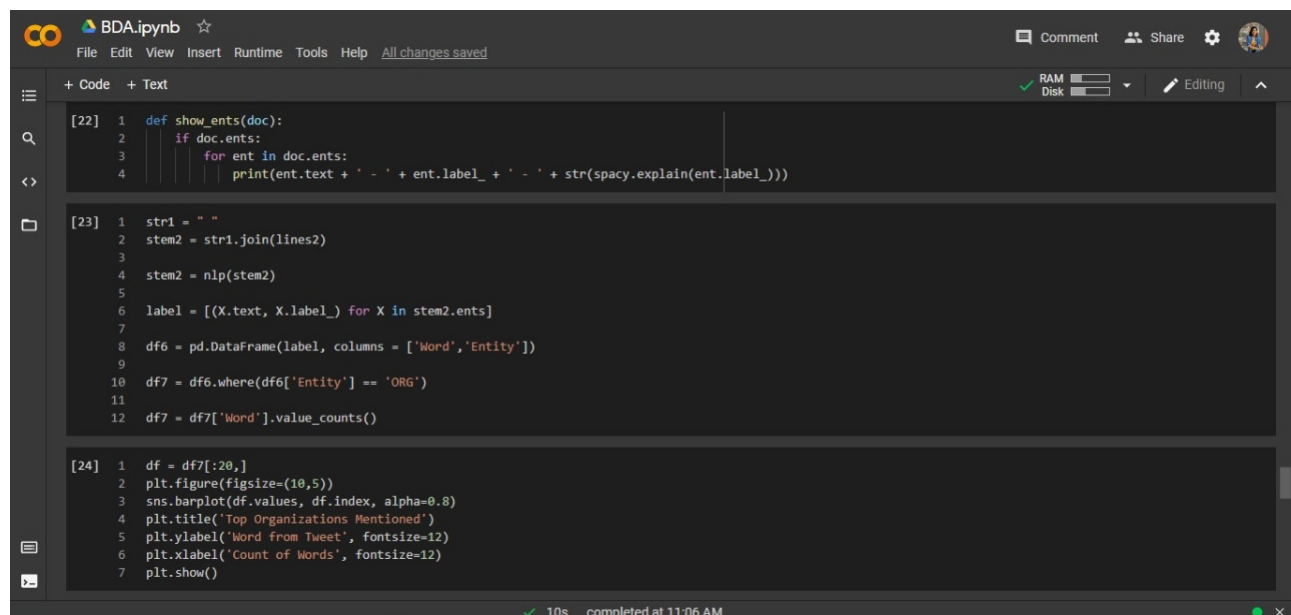
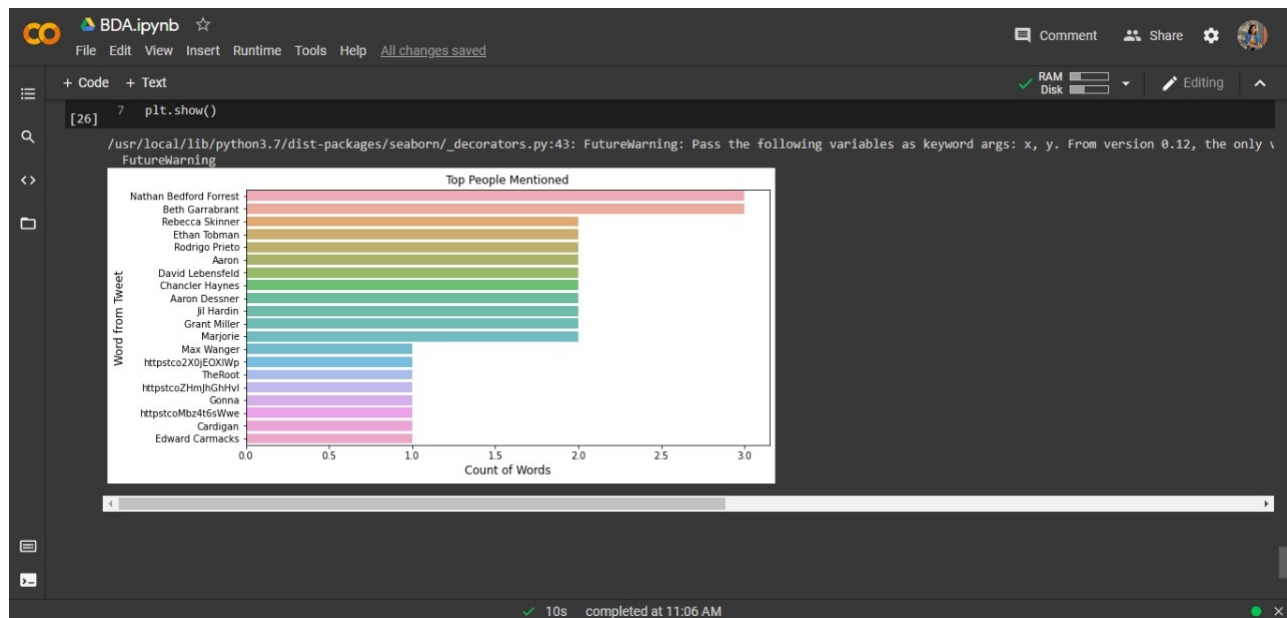The above 24th Code gives out the Top Organizations mentioned by the user.

Graph Representation of Top Organizations Mentioned

Similarly the above code gives out the Top people Mentioned by the user.

Graph representation of Top People Mentioned

<u>**CONCLUSION**</u>

Twitter analytics weren't always so easy to use. In the past, if a marketer wanted to learn about their audience, they'd usually have to use a suite of tools like Buffer, Hootsuite, Commun.it and other services to know what was happening. Twitter was once considered a part of what marketers call the "dark social" --meaning that it wasn't so easy to track engagement in this platform. It seems as though Twitter responded, and turned a light on.

Social media dependency is inevitable, which has resulted in the generation of an abundant amount of data sets, making the method of processing and analyzing of data a challenge. Extensive dependence on social media data such as Twitter data, e-commerce data, etc. have gained much attention in the area of data analysis. In this Mini Project using the tools of Twitter Developers account and Python Packages we have created a small system where you can get to know a particular users interest and get their stats accordingly. This Data can be used for various purposes.

REFERENCE

➢ https://docs.tweepy.org/en/latest/auth_tutorial.html

➢ https://developer.twitter.com/en/docs/twitter-api/getting-started/getting-access-to-the-twitter-api#:~:text=API%20Key%20Secret%20%2D%20This%20is,behalf%20of%20that%20Twitter%20account.