ANALYSIS ON STREAMING LIVE TWITTER DATA FOR EVENT PREDICTION

K. lokesh (AP18110010634)

K. Manideep (AP18110010637)

G. Likhith Chandra (AP18110010608)

D. Narshima (AP18110010648)

Y. Varshith(AP18110010640)

ABSTRACT:

* In now-a-days Because of usage of social media, it produces large amounts of data every minute. previous years Industry innovations have allowed new forms of communication between people and started opening many of the new business opportunities.
* People use Twitter to send short and instant messages, and it has become one of the most popular sites. Because of Twitter's widespread acceptance, occurrences such as breaking news and the release of viral videos have become commonplace. can easily attract the interest of others and spread quickly on Twitter. As a result, the number of tweets covering an occurrence can be used to estimate its visibility and significance.
* Furthermore, related tweets show public sentiment and reactions to events. As a result, it's critical to recognise and interpret Twitter events. We present our work in this dissertation, which aims to (1) classify events from a Twitter stream and (2) examine personal topics and events. in addition

users on Twitter, and (3) have a summary of the events discovered via Twitter. First we will concentrate on identifying events on Twitter. We found that the textual content of tweets, combined with their temporal trends, offers valuable insight into the public's attention and priorities. A sudden rise in topically similar tweets normally signals a spike in interest in recent events. We will extract the tweets from streaming api with help of tweepy module and preprocess the extracted tweets and we finally conclude the event prediction .

INTRODUCTION:

The way people access knowledge has radically changed as a result of the exponential growth of social media on the Internet and the rapid adoption of smart mobile devices.

Traditional media such as newspapers, television, and radio have been replaced by digital media such as Twitter among the younger generation. Furthermore, social media makes it possible to

Users should be able to actively engage in the creation of content.

Twitter had over 600 million active users as of July 2014, and these users sent out about 58 million tweets every day. As a result of Twitter's widespread acceptance, major events can now be broadcast live on the platform.

Easily attract the attention of the majority and elicit heated debate. As a result, the number of tweets covering an occurrence can be used to estimate its visibility and significance.

These occurrences may be newsworthy (e.g., a traffic accident) or not.

election) or entirely on the internet (e.g. the spread of a viral video). Furthermore, related tweets show public sentiment and responses to events like elections and scandals.

When it comes to tweeting about events, Twitter users play a crucial role. Each consumer has her own personal interests, which can be deduced from her previous tweets about those interests. The type of event she is concerned with would be influenced by these user-specific interests. Personal preferences and activities are orthogonal in this case since certain events come under specific topics.

Finally, when a significant event occurs, it will spark heated debate. However, reading all of the related tweets to interpret the event takes time. As a result, we intend to summarise the identified events in order to assist users in better interpreting these events.

METHODOLOGY:

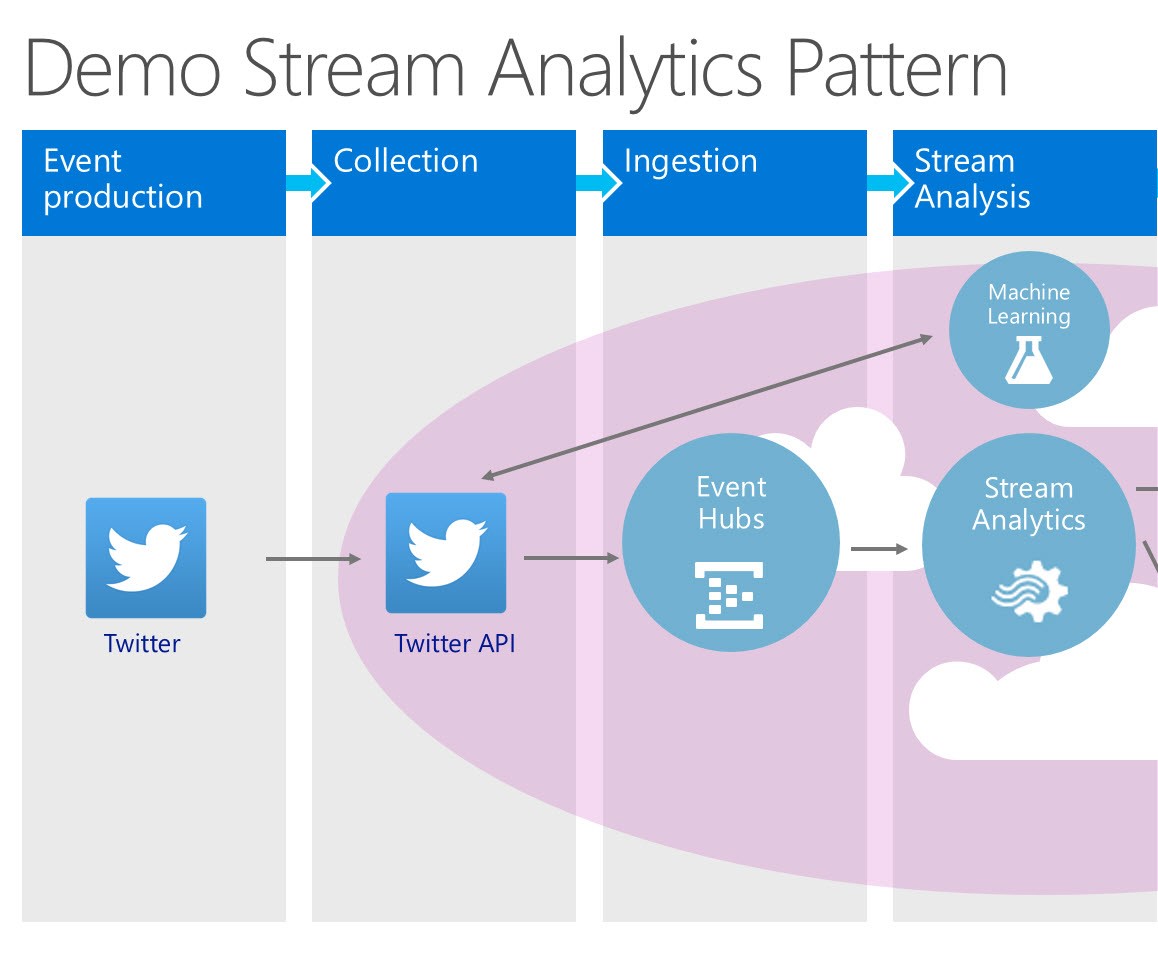
The next step in the methodology was to collect Twitter data and analyse it. Since Twitter provides a publicly available Application Programming Interface (API), my aim was to examine various approaches to APIs, tokens, and their use.

Implementation from a theoretical standpoint, followed by the collection of tweets for sentiment analysis in order to create a training and testing classification process.

Following the collection of a sample data set from Twitter, I began data analysis in order to better understand the data structure of the Twitter stream. It is important to conduct data analyses on tweet attributes in order to

to know its structure and plan tweets for sentiment analysis, pre-processing, and

classification.



PROPOSED SOLUTION:

An API (Application Programming Interface) is a user interface for interacting with web services. It allows developers and the general public to build service products on top of the API and integrate them into their own service solutions. Usage of the Twitter service

Streaming API and REST API are two separate types of approaches that can be used.

When tweets are retrieved as a continuous stream of content, the Streaming API provides access to the service. It is up to the developer's algorithm to determine what types of tweets are retrieved from API, such as keyword-based tweets; but, in the case of a security breach, it is up to the developer's algorithm to determine what kind of tweets are retrieved from API.

REST API would work best for single searches for user profile information.

The Representational State Transfer (REST) API stands for Representational State Transfer.

Its design is based on network concepts, which determine how users can access data resources.

RESTful services use the HTTP request-response protocol to communicate, however

In contrast to the Streaming API, it does not require the use of a persistent HTTP link.  Only API requests are made by certain service applications.  the Twitter service provides many forms of streaming API endpoints, each with its own set of capabilities. The data from Twitter is retrieved in JSON format.

As a result, the Streaming API design architecture allows for the collection of required data for the pre-processing stage of the methodology.

We need to build a Twitter application that will receive access tokens in order to make approved calls to the Streaming API and collect data for the preprocessing phase.

Open Authentication (OAuth) is an authentication standard that allows applications to access data from other services without having to show their credentials. We want to link to the Streaming API in our data mining solution, and therefore the Twitter control panel for developers, apps.twitter.com, offers generating our access token.

Important credentials for creation and Streaming API access can be found in the Keys and Access Tokens tab of the newly developed TwitterMining framework settings. There are four qualifications to be aware of: Consumer Key (API Key) and Consumer Secret

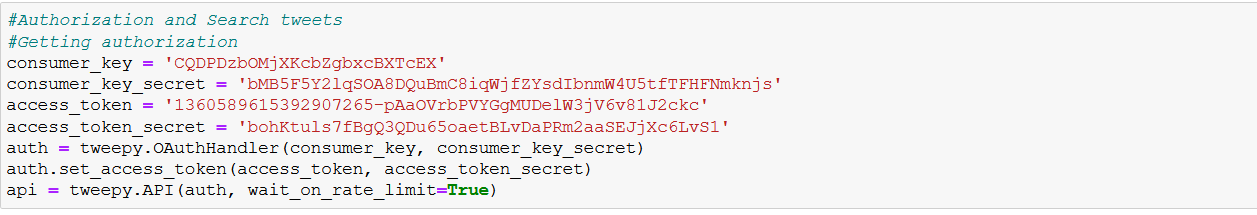
Access Token, Access Token Secret, and User Secret (API Secret).

These credentials give the TwitterMining application all it needs to authorise itself.

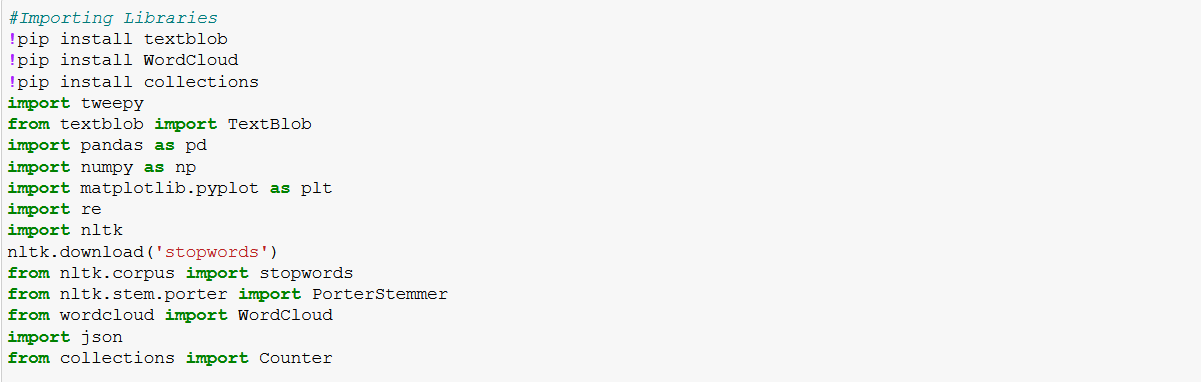
From the account of the user, access tokens are used to make API requests to the Twitter service. Furthermore, for the purposes of this project, the access level is set to Read and Write; however, if necessary, it can be modified according to the permission settings of the project.

submit an application Tokens may also be regenerated or revoked. After obtaining Streaming API credentials, we must establish a link and collect sample tweets in order to analyse the structure of each tweet and the attributes it contains. The API key and API secret must be transferred to OAuthHandler, which will build object auth to allow authentication while the function is running.

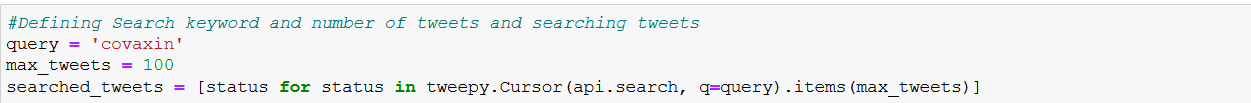
set\_access\_token will setup Access Token and Access Token Secret.



In juypter notebook first of all we have to download all related required modules for analysis of tweets.



Since data analysis involves a sample partition of tweets in order to become familiar with their content, the following code will stream near-real-time tweets based on user feedback. Both tweets are printed in the console, but they are also piped into a text document for a better summary. As an example of user input, consider the following: (Covaxin) is a type of vaccine.



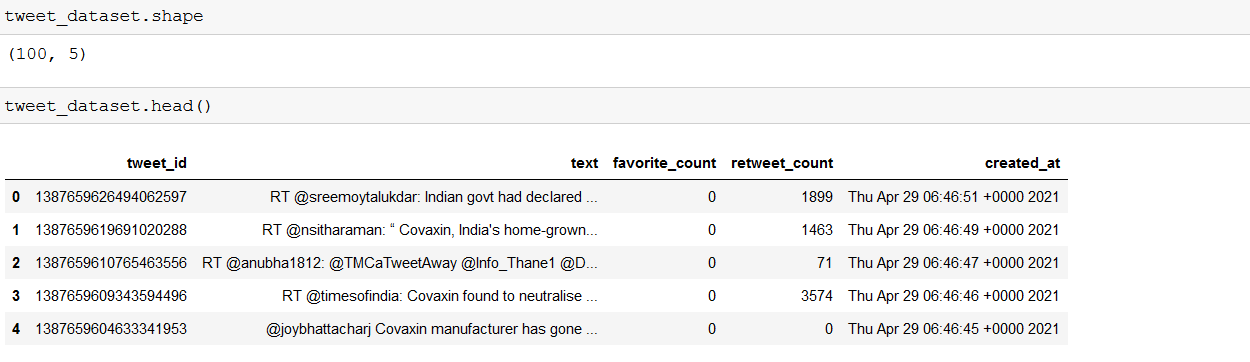
A json format file(where tweets related information will be present) will be downloaded and also CSV file(['tweet\_id',’text’, 'favorite\_count', 'retweet\_count', 'created\_at'])) will be downloaded.



Using the pandas library we can display the dataframe of csv file

Data collection based on data structures plays an important role in data analyses. Effectively structured data into tables provides data scientists with a deeper understanding of the data's structures. We use Data in our tweet data analysis model.

Python Pandas is used to allow frames. A data frame can store information in tables and mark the rows and columns.



The text attribute includes covaxin-related statements from the Twitter Streaming API. The reason for this study domain is that it has gotten a lot of coverage from the media and social media all over the world. It is extremely important.

discussed case, which is required for near-real-time streaming data to be obtained.

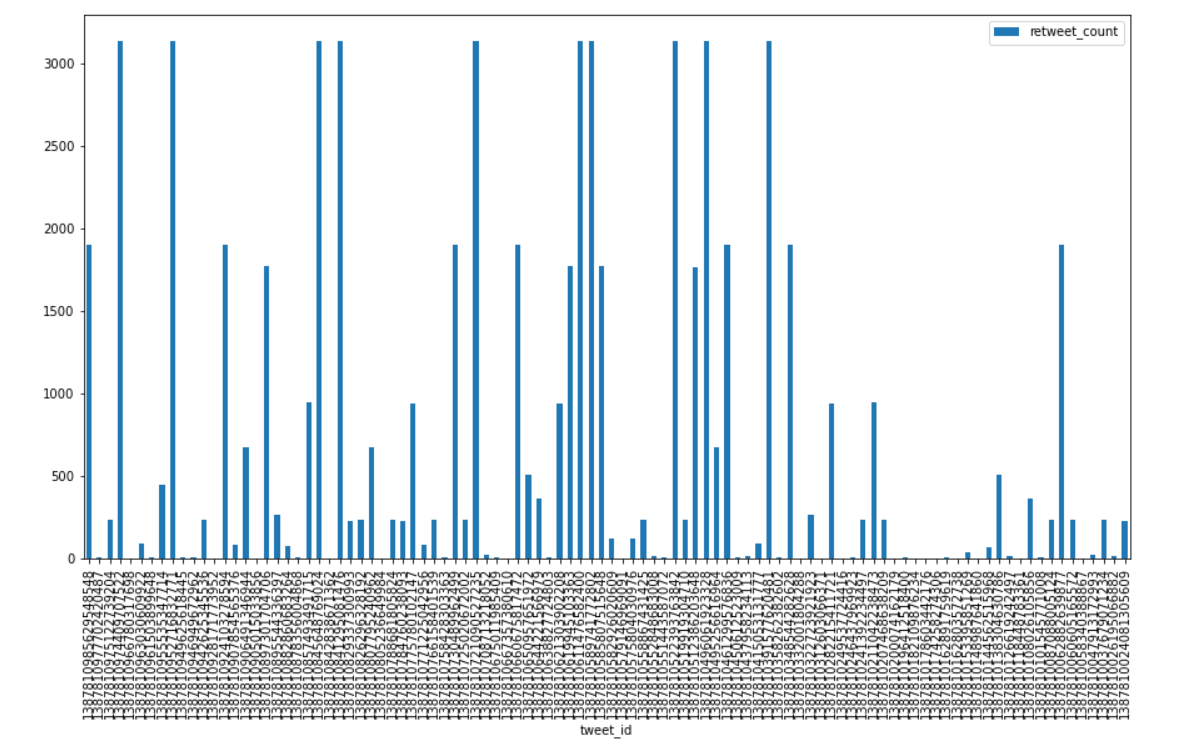
Tweets visualization

Since social media can have a significant effect on perception, the goal is to develop a solution that classifies live tweets based on emotion and converts them into public opinion, which can then be visualised using a graph.

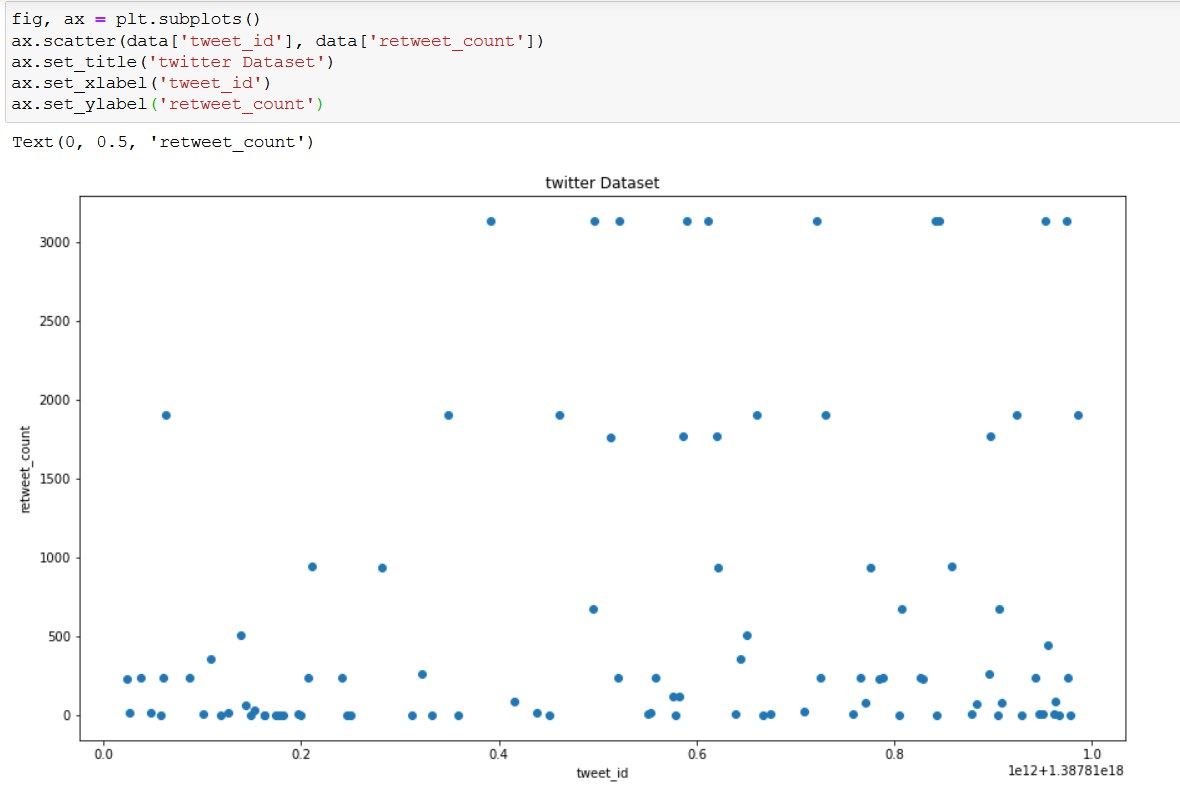
We will plot the distributions based on retweet count and twitter id attributes which are present in the dataframe.

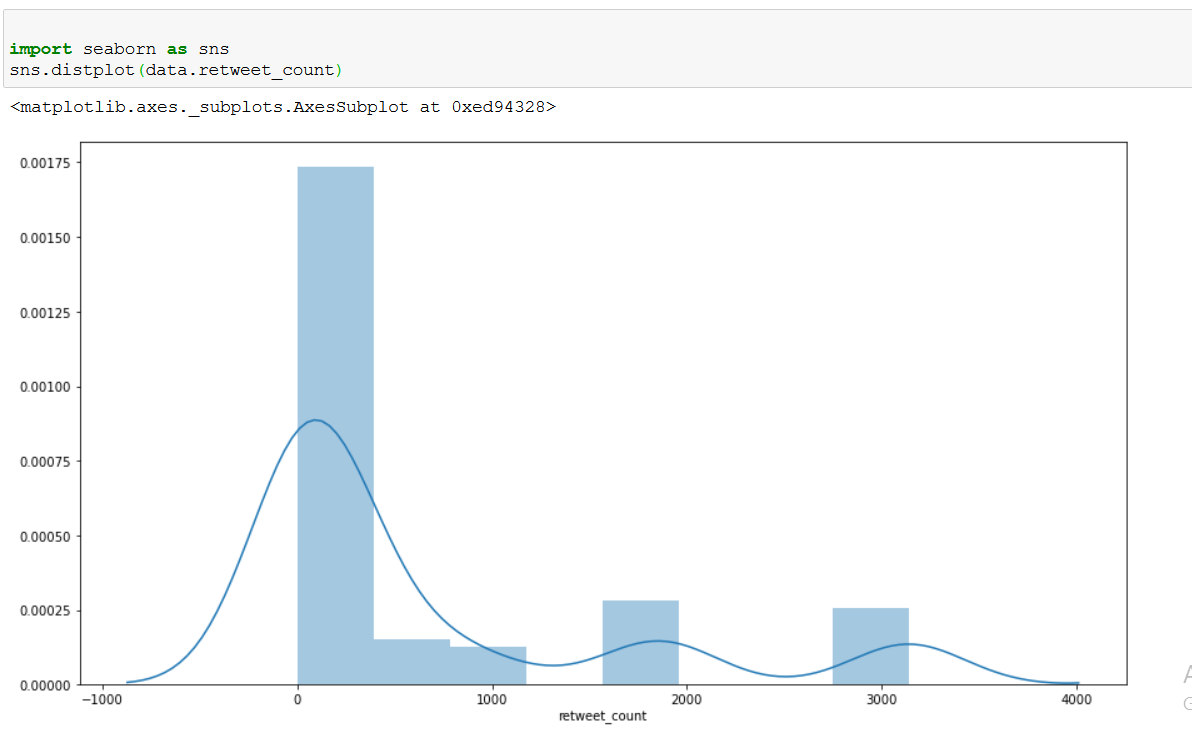
The bar graph shows the distribution of retweet count and users twitter id.





Scatter plot shows the relationship between users tweet id and retweet count.





Natural Language Processing Analyses:

Tokenization is the method of breaking down text strings into tokens, which are words in sentences. This method will aid Named Entity Recognition when analysing Twitter data. We can request samples from the tweet dataset.

To see the results, take the data and tokenize it.  The aim of tokenization is to look at efficient ways of breaking tweets into tokens and comma-separated mentions.

Stop-words are words that have no meaning in context and thus have no significance for sentiment analysis. Natural Language Toolkit includes a stop-words library that can be used in a solution. They've been classified.

As a result, our solution can only connect English stop-words.

NLTK (Natural Language Toolkit) is one of the best library for preprocessing text data

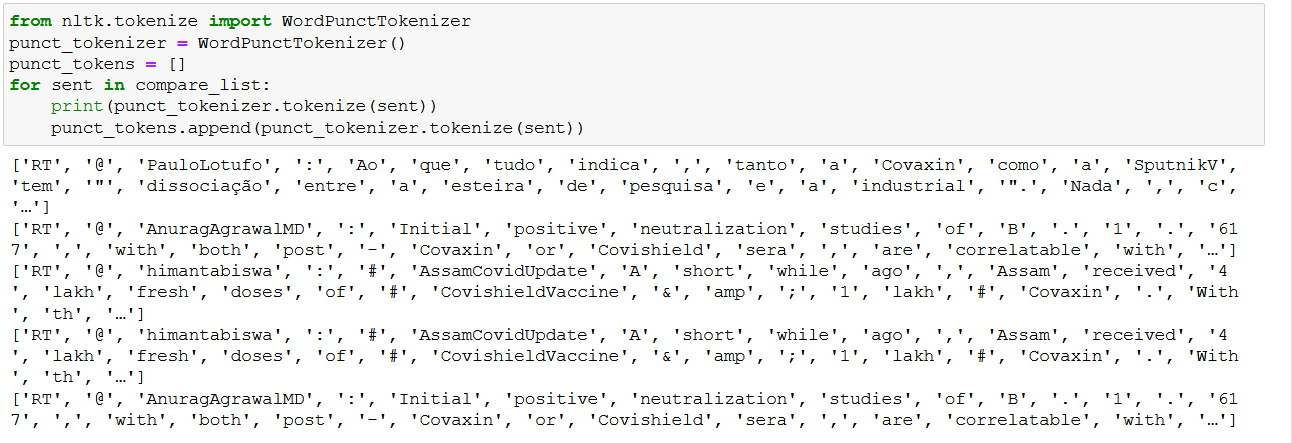
word\_tokenize

When it comes to tokenizing sentences into words, word tokenize is the most common tool. Separate words using spaces and punctuation with word tokenize.



# WordPunctTokenizer

All punctuations are split into different tokens by WordPunctTokenizer.



# RegrexTokenizer

# We can instead use RegrexTokenizer to manage how text is tokenized since there is no tokenizer that splits up words based on space.

# Our pattern is compiled by the RegexpTokenizer class, which then calls re.findall() on our text. This feature could be used to fit alphanumeric tokens with single quotes.

# 

## **Match on Whitespace**

# RegexpTokenizer may also be used to fill in spaces. The matching pattern will be used as separators when the parameter gaps=True is applied. \s+ matches one or more space.

# 

# Rather than spending time analysing the results of each tokenizer, we can combine everything into a single pd.dataframe for quick and accurate interpretation.

# 

# CALCULATING THE WORD FREQUENCY:From the twitter data ,in that for the text attribute we calculate the word frequency.We can use the built-in Python library collections to get a count of how many times each word appears in the sample, which helps build a special form of Python dictionary. . The most common built-in method of the Counter object returns the most frequently used terms and the number of times they are used.

# 

# we can  build a Pandas Dataframe for review and plotting based on the counter that only contains the top 15 most popular terms.

# 

# With the help of pandas data frame we can draw the histogram of top 15 most common words in tweets text.

# 

# WORD CLOUD

# Following the pre-processing steps, we removed all place names and abbreviations from the tweets because they served as a leakage vector, and then we built a word cloud using a frequency distribution of the most frequently used hashtags.

# 

# 

# Sentiment Analyse:

# Now we can measure polarity, subjectivity, emotion, negative, positive, neutral, and compound parameters using cleaned text once more. I add new functionality to our data frame for all measured parameters.

# 

# 

# CONCLUSION:

# The methods described above will detect events by locating all tweets connected to them. However, reading all relevant tweets makes it difficult for users to understand these events. As a result, it is preferable to create a succinct summary to assist the reader.

# Users gain a greater understanding of each event's existence. To summarise the events on Twitter, we suggested a word graph-based summarization framework, and by the plot distributions from the twitter data we can analyze the twitter attributes(retweet count and twitter id) , using the word frequency graph we can know the top repeated words by this we can know the things around us which was trending , Twitter sentiment analysis helps you to monitor what people are saying about your product or service on social media, and it can help you spot disgruntled consumers or negative mentions before they become a big crisis.

# From the pie diagrams of sentiment analysis. We can know the twitter tweets whether it is positive , negative, or neutral ,we have applied sentiment analysis on twitter tweets data.

# REFERENCES:

# <https://towardsdatascience.com/basic-tweet-preprocessing-in-python-efd8360d529e>

# <https://www.researchgate.net/publication/327160507_Sentiment_Analysis_on_Twitter_Data_Using_Machine_Learning_Algorithms_in_Python>

# <https://www.researchgate.net/publication/340523957_Word_frequency_and_sentiment_analysis_of_twitter_messages_during_Coronavirus_pandemic>

# <https://www.geeksforgeeks.org/twitter-sentiment-analysis-using-python/>