# Introduction:

In this era of big data, where millions of data is generated every second, all that is required is to leverage the data in order to get meaningful insights from them and perform its application to get advantage in all sorts of business. In this task we are going to perform analysis and visualization on the chosen data with the help of Python language. The aim of the project is to connect APIs with the help of Python and fetch required data for visualization. We have chosen the language python as it have a variety of libraries which allows and makes it easy to utilize its various services. The chosen library for this task is Tweepy which is an open source library and is hosted on GitHub. This library allows communication with Twitter platform through with the help of Python language and use its API. The version of tweepy used is 3.7.0 released on November 2018. The second API is chosen for visualization is Github (github). This API helps to connect with Git (version control application) and fetch the data form it and push into table and perform visualization.

# PART 1 – Project Plan

**Group name:**

**Co-students:**

**API used:** For this project we would be collecting data from Twitter API.

**Data collected from twitter API:** We have created a list of famous celebrities’ including bill gates, Barack Obama, Narendra Modi, etc. and we have fetched their details such as followers, friends, status, URL and pushed these details into sqlite database and another table is created in which the data searched for the tweets which has used their names and pushed into the table and same data is used by creating join.

**Data collected from Github API:** The data is fetched for the number of repositories present for the latest technologies on Github. The count is pulled from GIT and pushed to table and PIE chart is created.

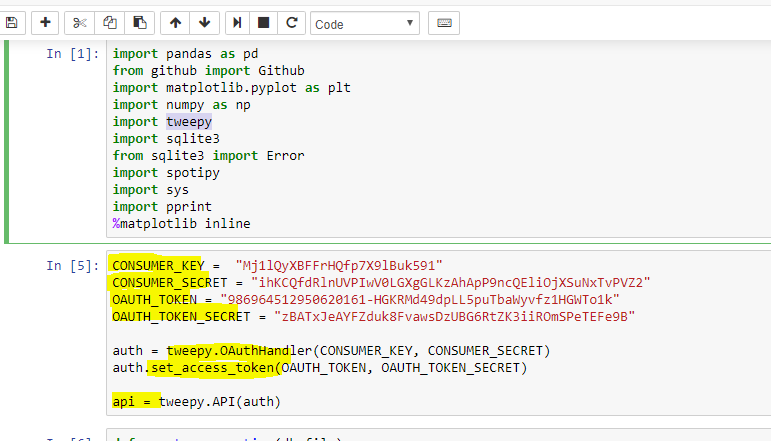
**Visualization package:** MatplotLib

**Graphs/charts:** Pie chart and Bargraph are created using the calculated data. In the bar graph there would comparison of the tweets done by each of the celebrities and the pie chart will compare the number of repositories present for that particular technology.

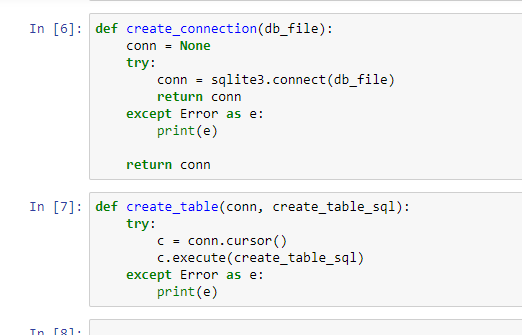
# PART 2 – Gather the data and save it to a database

For this part the data is fetched from the API’s and then that data is pushed into the database. For fetching of data from twitter the api is used as Tweepy. For connecting Tweepy the Consumer key and authentication credentials are passed (which are generated using developer feature in Twitter).

The below snapshot shows the keys used for authentication and import of library in the code.



Connection for Sqlite database created in python using sqlite3 library and for creating table and connection separate methods are created which is used frequently in the code for creating tables.



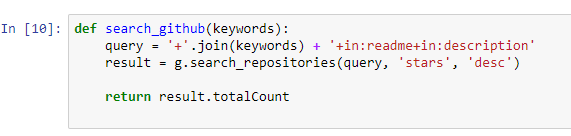


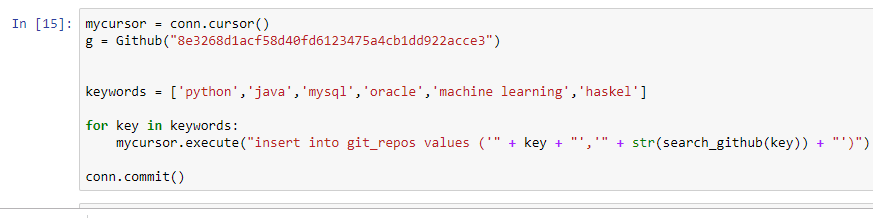
Once tables are created then the data is fetched from twitter and pushed into the two tables which are created for twitter data.

**get\_user** method is used to fetch the data from tweepy:



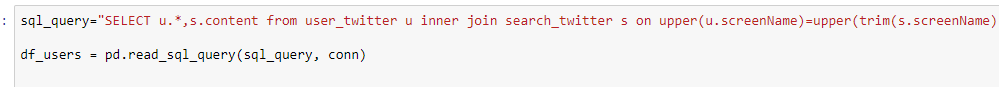
Similarly the connection is made with Github and data is pushed into the git hub table which is created in the above section.



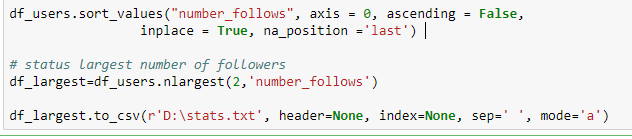


# PART 3 – Process the data

The join is created between the tables’ **user\_twitter** and search**\_twitter.** The user\_twitter table has details about the celebrity and search\_twitter has the details of number of times the celebrity # tag is referred and hence to get the details of user with search then join is used and the result is fetched into pandas dataframe.

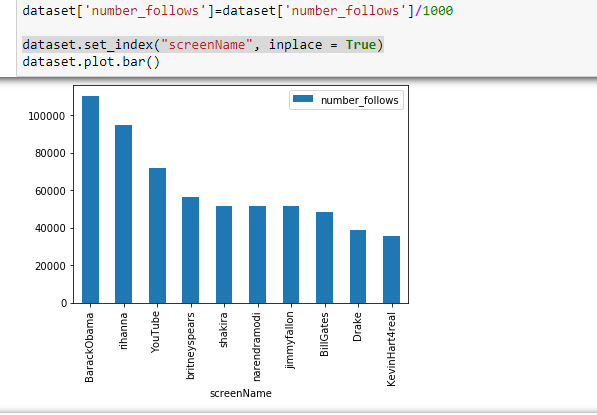


For data calculation we have sorted the data and created a text file which has the nlargest number of followers of the celebrity.

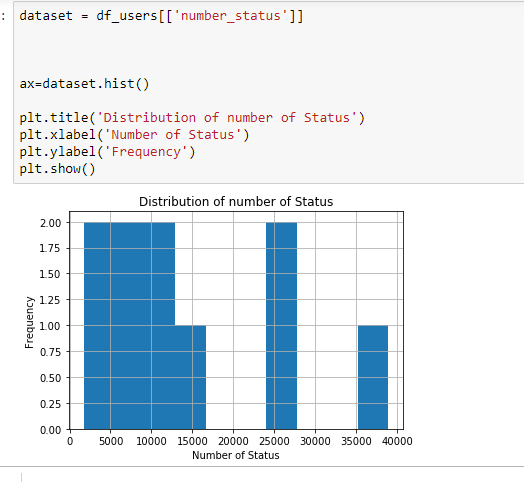


# PART 4 – Visualize the data

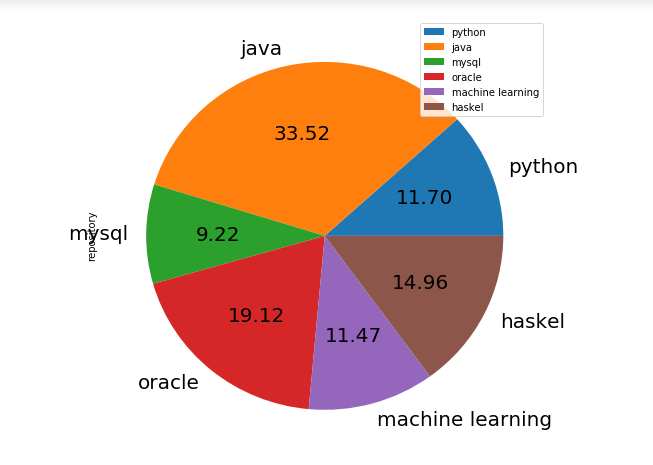
# In the complete assignment total three charts are plotted, 1st is bar graph, 2nd is histogram and 3rd is pie chart. The first bar graph highlights the number of followers of the celebrity and the graph shows that the highest number of followers are of Barack Obama and it has crossed 0.1 million followers and second highest are belongs to rihanna which is around 90k and the lowest number of followers are of kevinHart4Real.



The second is histogram to visualize the frequency of number of status by the celebrities:



The third chart is pie chart which shows the percentage of repositories belongs to particular technology on git gub.



The highest number of repositories are belong to Java and lowest percentage is of MySql. The Haskell also comprises of good number of repositories.

# PART 5: Report

## 1. Goals

The Goal of this project is to connect APIs with the help of Python and fetch required data for visualization. In this task we are going to perform analysis and visualization on the chosen data with the help of Python language. The aim of the project is to connect APIs with the help of Python and fetch required data for visualization. We have chosen the language python as it have a variety of libraries which allows and makes it easy to utilize its various services.

## 2. Goals achieved

We have successfully created the python code which connects with the twitter and Github API and collect the data to visualize them and compare them in order to get insights and knowledge from the data. The data is pushed into the tables in SQLITE and database is created which is used for visualization.

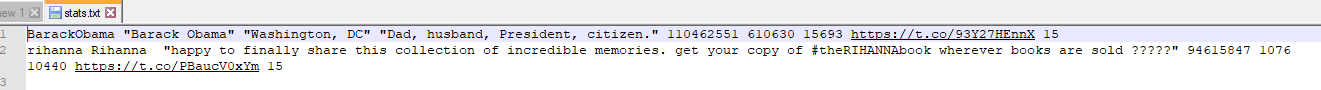
## 3. Problems faced:

Issue faced while trying to connect with credential on GitHub api and to fix it we have used token authentication.

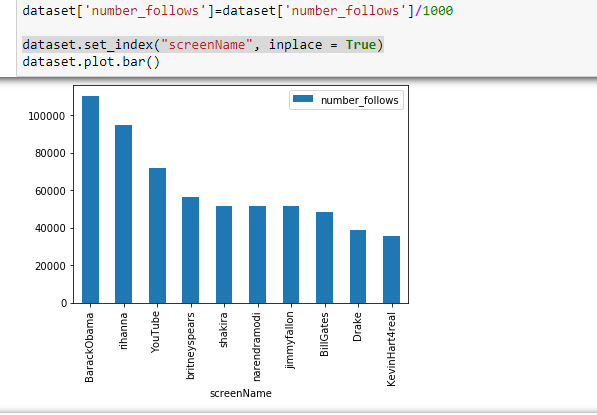
Another challenge is faced while creating the pie chart and it was showing that error about data type, we have checked that the repository column datatype was object. To fix this we have change the data type to int.

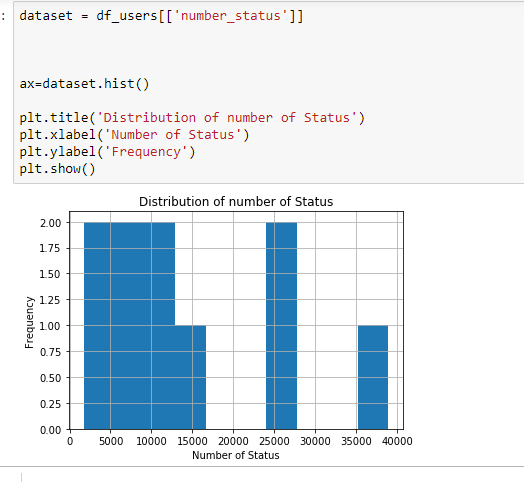
## 4. Calculation file

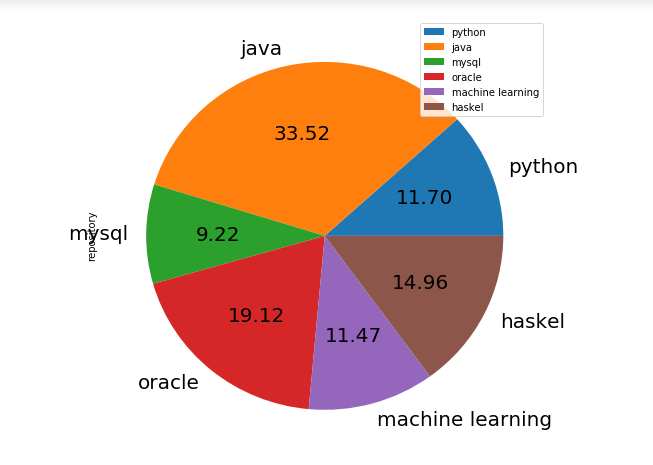
Below shows the snapshot of calculation file.



## 5. Screenshots of Visualization







## 6. Instruction to run the code

The zip submitted has python notebook, this needs to be uploaded to Jupyter and then the graphs and output can be visualized. There are option present in the Jupyter to run the notebook. In this way the code can be executed.

## 7. Documentation of functions

def create\_connection(db\_file)

**Input** – File name of the SQLITE database file.

**Output** – Connection is returned.

def create\_table(conn, create\_table\_sql)

**Input** – Needs connection variable of SQLITE and create table query.

**Output** – Nothing.

def search\_github(keywords)

**Input**: Needs keyword to be searched on Git hub

**Output**: Integer (number of repositories) for particular keyword.

## 8. Resources utilized

We have utilized the python as it have a variety of libraries which allows and makes it easy to utilize its various services.