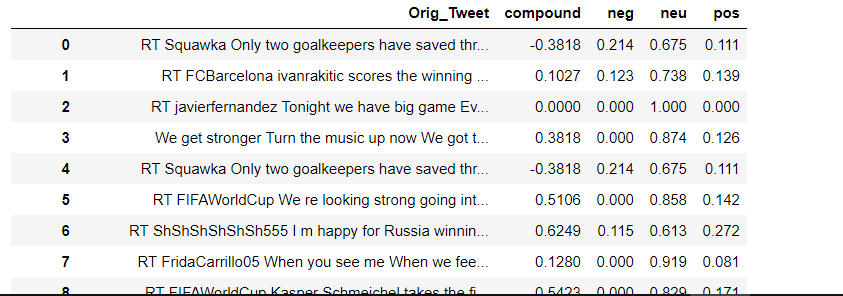
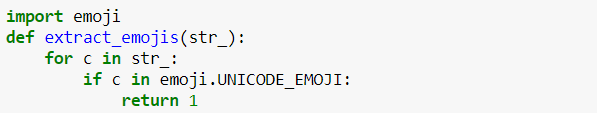
Emoji Data Science

**Downloading and Reading file**

In this document I’ll explain that how i have done twitter analysis of all the ( around 530000 ) tweets which was posted on the topic or with the topic #fifaworldcup2018 . For this analysis I have used the twitter api to get all the tweets related to my point of concern ( i.e. tweets with the #fifaworldcup2018 ), one another option of this analysis of the tweets is the dataset which is available on the kaggle (with the title FIFA World cup 2018 Tweets, this file contains an FIFA.csv file which is the required set of data which is needed ) , after I have downloaded the dataset (or scrapped from the twitter using its api) the next task is data cleaning, one benefit of using kaggle dataset is that cleaning of data is already done by the data provider. But for the sake of efficiency I have checked and verified its cleaness by myself.

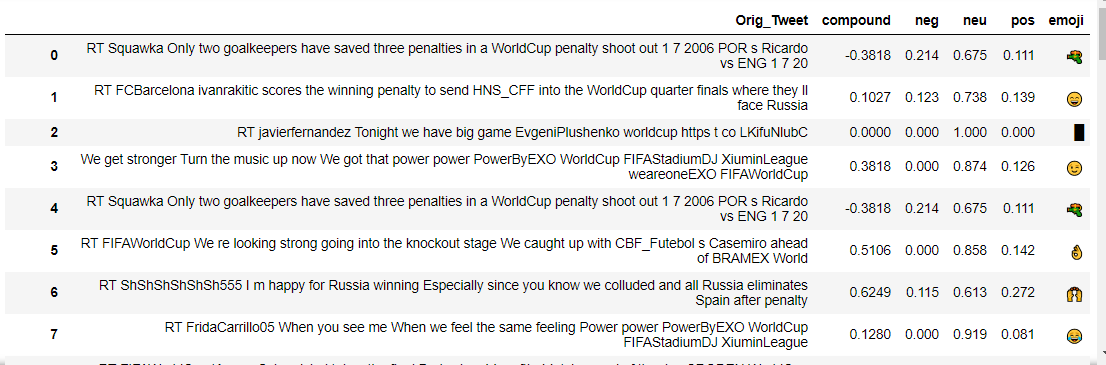


After all of these task I have moved to the next step ( i.e. the main steps all of the above steps are the basics ) now I have to scrap all the emoji’s from the original tweet data , to do this task of scraping the emoji’s from the text I have used the emoji module (this modules contains special classes and methods which help us to scrap all the emoji’s from the string ) for my case I have used a for loop and if statement to loop through each word of string ( tweets ) and using if statement check that it is in emoji.UNICODE\_EMOJI if it is in emoji.UNICODE\_EMOJI it will return that emoji . and finally we will get the emoji .



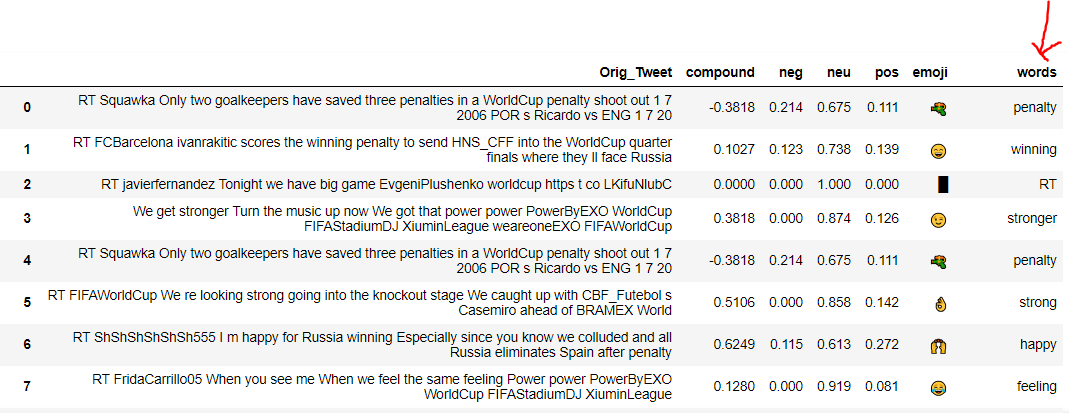
Now I have the original tweet and also I have the emoji , the next turn is to do the sentiment analysis of the tweet and assign a polarity score and compare that polarity score with the emoji Sentiment score and then assign an sentiment to my dataset . To do this all of the task I have used nltk library which is designed for natural language processing . The main reason why I have done sentiment analysis of the tweets by the people with #fifaworldcup2018 is because during this analysis part my main aim was to predict the people view and thought on the fifa world cup which was the most trending topic during the time when world cup is in progress . And also this analysis will help to know the sentiments of people to my point of concern . And also this analysis will definitely help me to predict that which emoji is the most people intrest ( i.e. the emoji which is being used by most of the public , that is the most popular emoji which is most people prefer or like to use most often ) , combing this emoji analysis with the sentiment analysis of the tweets we can do better visualization for the data prediction, but before the visualization part I have to do some feature engineering for our data . Feature engineering is the most important factor the data science ( data visualization ) the more feature you add the more better visualization can be done. The FIFA.csv file provides different columns related to the tweets but for my point of concern the only important column is with the name original tweet all the analysis of data visualization is done using only this column . using this column I have add 4 columns named as pos , neg , neu , compound this was done for the sentiment analysis from the sentiment analysis ( sentiment analysis is done using nltk module for the natural language processing ) the pos column contains the value of positive sentiment of the tweet ( positive sentiment is between 0 and 1 ) , neg column contains the value of negative sentiment of the tweet (negative sentiment is between 0 and 1 ) , neu column shows that how neutral the tweet is ( value of neutral is also between 0 and 1 ) and the next comes the last column named compound this shows the compound value of the tweet ( this is between -1 and 1 ) , i.e. pos shows the percentage of positive words in a tweet and neg shows the percentage of negative words in the same tweet and neu gives the percentage of neutral tweet and finally the last compound shows the overall rating of the tweet it is between 0 and 1 . now it’s time to load emoji ranking dataset The emoji ranking dataset is as shown below

This dataset contains 752 rows that is it contains 752 different emoji, the emoji is shown in the column named Char despite this column the emojiranking dataset contain 4 more columns, named as Neg , Neu , Pos , Sentiment score , this all the 4 column is made using sentiment analysis by using nltk module, the four column has different meaning pos shows the percentage of positive words in a tweet ( it is between 0 and 1 ) and neg shows the percentage of negative words in the same tweet ( it is between 0 and 1 ) and neu gives the percentage of neutral tweet ( it is between 0 and 1 ) and finally the last compound shows the overall rating of the tweet it is between 0 and 1 . now using all of this data present in the FIFA world cup dataset and the emoji ranking dataset I have created an another dataset which contains the 6 column named as original tweet , compound , neg , neu , pos , emoji , the dataset is as shown below :

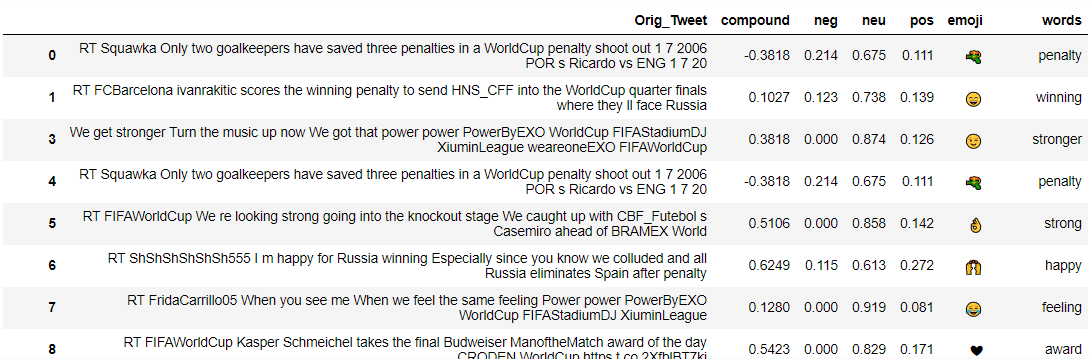


This final dataset contains 530000 rows each with the different row ( that is the content of the tweet in each row is different that is there is no any duplicate row in the this dataset ) and this dataset the Orig\_Tweet contains the original tweets of the peoples the pos column contains the value of positive sentiment of the tweet ( positive sentiment is between 0 and 1 ) , neg column contains the value of negative sentiment of the tweet (negative sentiment is between 0 and 1 ) , neu column shows that how neutral the tweet is ( value of neutral is also between 0 and 1 ) and the next comes the last column named compound this shows the compound value of the tweet ( this is between -1 and 1 ) , i.e. pos shows the percentage of positive words in a tweet and neg shows the percentage of negative words in the same tweet and neu gives the percentage of neutral tweet and finally the last compound shows the overall rating of the tweet it is between 0 and 1 . generating this dataset took my device a long time ( over night ). So , it is very important for me to save this generated dataset to my local drive , to save I have used pandas command and saved it to my local drive so that when ever needed I can easily excess . I have saved this file with the name fifasentimfile.csv and every time when I have to work I simply read this file from my local drive using pandas , and saving the file to my local drive saved me a lot of time .

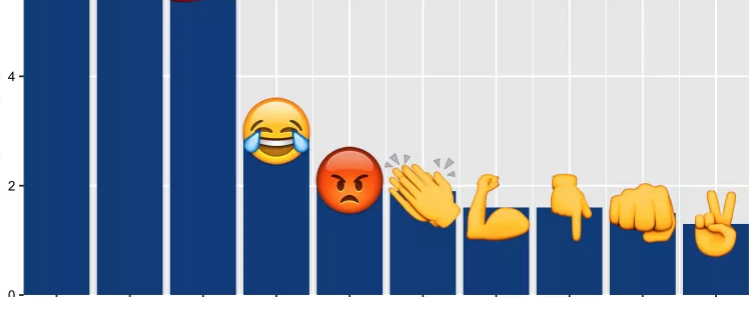
Now let’s do text analysis of all the tweets and detect the most common word with the emojis according to the sentiment score and make a new column named word to store them.



Now its time for natural language processing on the words column. We will convert all the words into lowercase, correct the spelling mistakes and convert the words into its root word(lemitization) and doing all this we will delete all the words with zero sentiment score because these are of no use. After doing all of this we will get the final result as.



**Visualizing emojis**



Now its high time to visualize the emoji using ploting technique available in matplotlib , because visualization is the most important part of this project and I think that it is important part of all the projects because a simple plot can explain a lots of thing that 100’s of lines of word explanation can not explain . due to this reason I always tries to visualize most of the things for proper explanation . and a plot can also be understood by an very less educated people and also an uneducated people, so I have decided to plot the visualization using bar plot with the emoji above it so that it will clearly show that how which emoji is used to which extent and also this graphical visualization will explain that which is emoji is used most of the public that is which is the most popular emoji . the tables in the data set is nice to look at but is there a more prettier way to which we can visualize the data in the emoji , to do all of these task I have to first create an grid and on that grid I will put all the things that I need ( I have decide to draw grid because once a grid is created we can put all thing we need on that grid easily , that is we can easily add and modify the grid and can get the desired plot ) , in this case I need to draw an emoji plot which will show the number of emoji used per 1000 tweets of the data provided in the FIFA.csv data . that is the y axis shows the count of the emoji and the x axis shows the type of emoji and the whole plot shows that which emoji is used most in the analysis of the #fifaworldcup2018 tweets and this will help us to know the people sentiment using the sentiment analysis of the emoji using emoji ranking . And finally my goal of the sentiment analysis of the tweets of the #fifaworldcup2018 is completed.

