# BONUS PROJECT REPORT

*SAI BHARATH KUMAR BANDI - UIN: 670244109 - NETID: sbandi3*

**CS 418 – Introduction to Data Science**

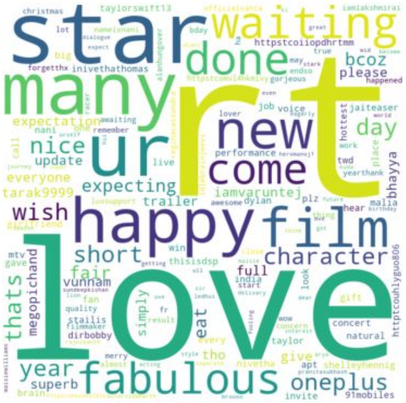
## INTRODUCTION:

The objective of this project is to perform sentiment analysis on tweets fetched from a selected account. We will create a twitter developer account and fetch the API tokens and secret tokens for data collection. Google Cloud Platform is used for analysing the data and Python3 is used for programming.

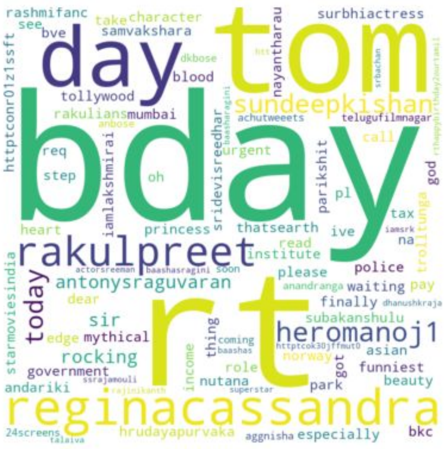
## TASKS AND ANSWERS:

As per the instructions given in the project description, the tasks are completed.

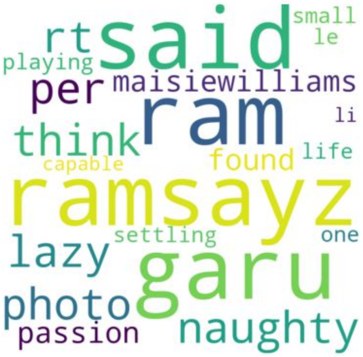
1. The twitter account used for analysis is: Bharathbandi12
2. All the twitter authentication keys are copied from the Twitter developer account and pasted in twitter_auth.py file.
3. In project3.py:
   a. Function twitter_auth() is used to return the twitter handler
   b. Function get_tweets() is to fetch the tweets from a selected account and create a list of text in the tweets
   c. Function make_dataframe() is used to create a dataframe of text in tweets
   d. Function text_preprocess() is used to clean the text and lemmatize the words
      i. Here the words are lemmatized. TextBlob and Word() is used
   e. Function generate_sentiment() is used to return the sentiment of set of tweets
      i. Sentiment is classified based on the value of text.sentiment.polarity value
      ii. If the value > 0, then the tweet is considered as good tweet
      iii. If the value = 0, then the tweet is considered as neutral tweet
      iv. If the value < 0, then the tweet is considered as bad tweet
   f. Function create_word_cloud() is used to create a word cloud for each classified tweets
   g. The word clouds are saved in the files Good.png, Neutral.png and Bad.png
4. The word cloud for good tweets:



5. The word cloud for neutral tweets:



6. The word cloud for bad tweets:



**Analysis:**

- There are more words such as ‘happy’, ‘love’, ‘star’ etc. in positive cloud
- There are more words such as ‘bday’, ‘tom’ etc. in neutral cloud
- There are more words such as ‘lazy’, ‘naughty’ etc. in bad cloud
- Most of the words are classified as neutral
- But the most common word appeared in three clouds is the word ‘rt’ - which is retweet. Most of the tweets contain this word irrespective of the bad, neutral or good. So this word appeared more frequently in the three clouds
- The analysis and classification is mostly depend on the number of tweets retrieved. Here we have retrieved 34 tweets from the account Bharathbandi12.
- If the count of the tweets retrieved is more, then the analysis on the type of tweet will be more precise.

**Reasons for using GCP vs a local environment:**

- The main advantage of using GCP is for processing large datasets. It takes very less time for processing large amounts of data.
- Whereas in local environments, processing of large datasets takes more time.
- In GCP, The data processing can be divided across different nodes of the cluster for parallel processing.
- In our project, the number of tweets retrieved from an account can be large and also can change from time to time.
- We need to fetch the data directly from the twitter from time to time. Where we need scalability
- So, we are using GCP for performing data processing
- Moreover, GCP provide secure and reliable infrastructure irrestive of the workload.