End Semester Report

Data collection

For data collection of the tweets asking or sharing their PII, I tried to search tweets with following search queries:

- Address sharing queries:
  1. tweets = twitter.search(q = "meet me -filter:retweets", lang = 'en' , count = "100" , result_type = "mixed",include_entities='True')
  2. tweets = twitter.search(q = "address is -filter:retweets", lang = 'en' , count = "100" , result_type = "mixed", include_entities='True')
- Contact sharing queries:
  1. tweets = twitter.search(q = "contact me -filter:retweets", lang = 'en' , count = "100" , result_type = "mixed",include_entities='True')
  2. tweets = twitter.search(q = "contact is -filter:retweets", lang = 'en' , count = "100" , result_type = "mixed",include_entities='True')
  3. tweets = twitter.search(q = "number is -filter:retweets", lang = 'en' , count = "100" , result_type = "mixed",include_entities='True')
  4. tweets = twitter.search(q = "call me -filter:retweets", lang = 'en' , count = "100" , result_type = "mixed", include_entities='True')
  5. tweets = twitter.search(q = "contact at -filter:retweets", lang = 'en' , count = "100" , result_type = "mixed",include_entities='True')
  6. tweets = twitter.search(q = "call at -filter:retweets", lang = 'en' , count = "100" , result_type = "mixed", include_entities='True')
- Email sharing queries:
  1. tweets = twitter.search(q = "mail me -filter:retweets", lang = 'en' , count = "100" , result_type = "mixed",include_entities='True')
  2. tweets = twitter.search(q = "email me -filter:retweets", lang = 'en' , count = "100" , result_type = "mixed", include_entities='True')
  3. tweets = twitter.search(q = "email is -filter:retweets", lang = 'en' , count = "100" , result_type = "mixed", include_entities='True')
- Address asking queries:
  1. tweets = twitter.search(q = "share your address -filter:retweets", lang = 'en' , count = "100" , result_type = "mixed",include_entities='True')
- Contact asking queries:
  1. tweets = twitter.search(q = "share your contact -filter:retweets", lang = 'en' , count = "100" , result_type = "mixed",include_entities='True')
- Email asking queries:
  1. tweets = twitter.search(q = "share your email -filter:retweets", lang = 'en' , count = "100" , result_type = "mixed",include_entities='True')

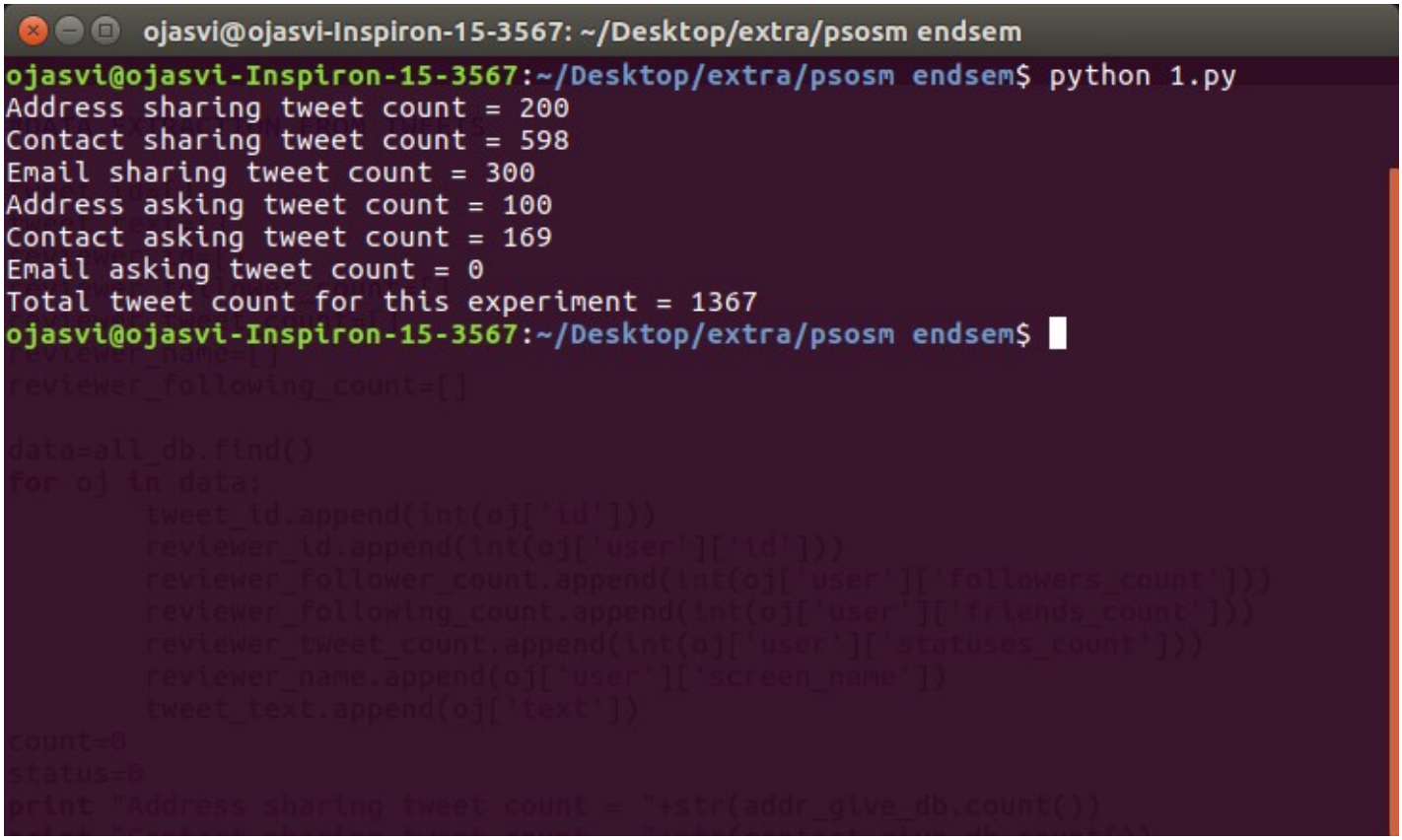
Decided PII's

Decided PII's for this project are:

- Address Details
- Contact Details
- Email Details

Data Collected

After collecting data, it was saved in mongodb database. The amount of collected tweets as per the individual querry type and in total all the tweets collected are as follows:



Feature Set

To train the model, the following set of features have been decided to assume a tweet to be from a spam user or not:

- Follower/Following ratio: For spammers, this ratio is less as spammers try to follow more people as much as they can, so that some of their targets might follow them back.
- Reputation: It is followers/(followers+following). Generally spammers have reputation<1.
- Tweet count: It is a general belief that the tweets posted by the spammers is generally much lower than than the tweets posted by an average twitter account.
- Verified: This parameter is used keeping in mind that spammers can never be verified users.

Machine Learning technique- K Means Clustering

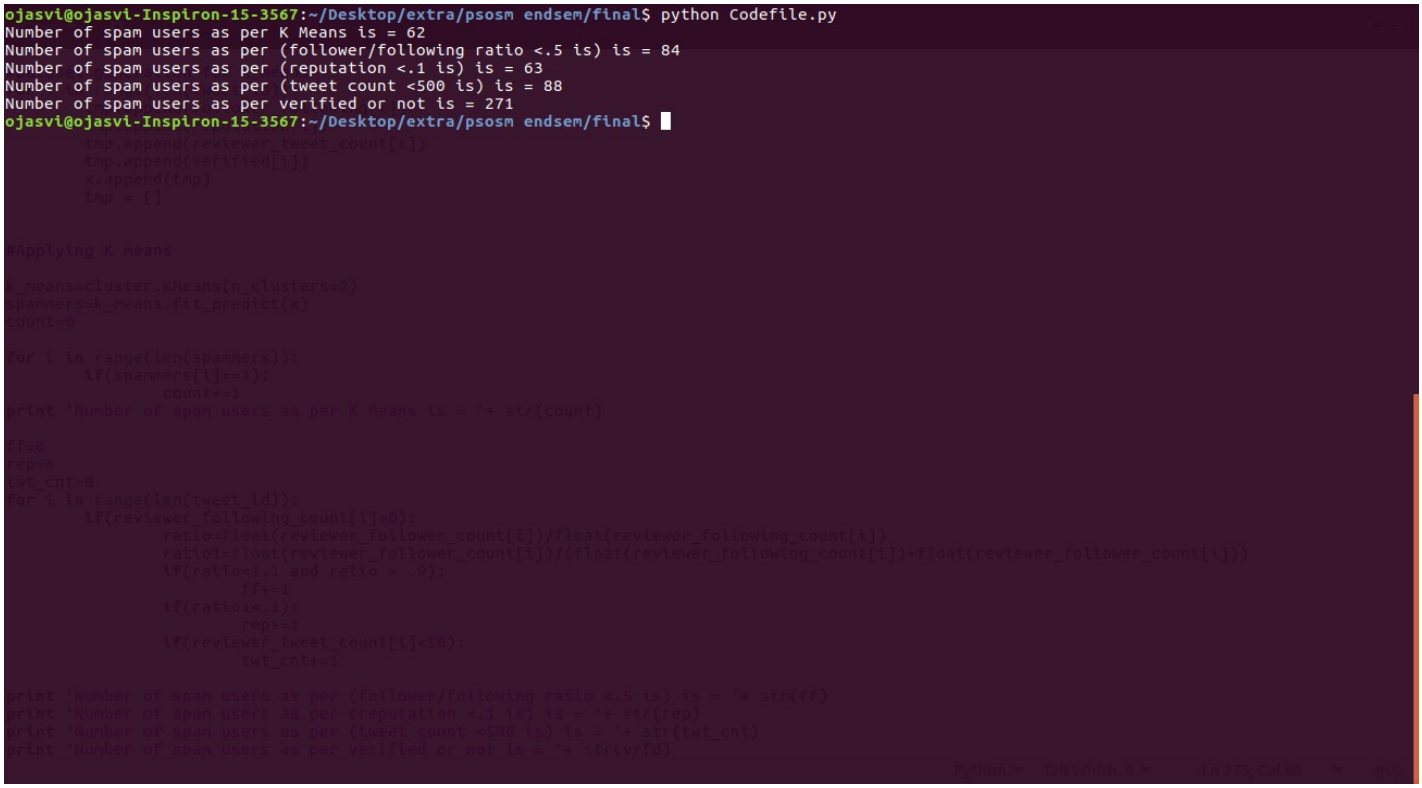
K Means clustering is an unsupervised machine learning technique which tries to bifurcate the testing data on the basis of variations between the input data set passed to the function.

Passed input - a list containing tuples for all tweets, where each tuple consisted 4 values of feature set

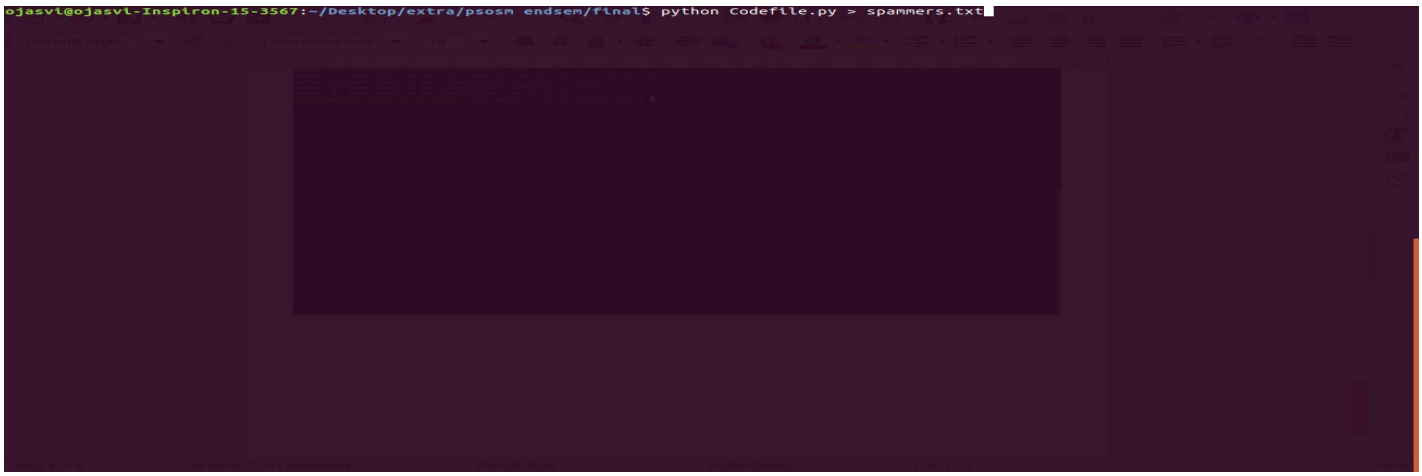
Output – A list containing either 0 or 1 for all the tweets collected marking them as spam(1) or non spam(0).

Output Analysis

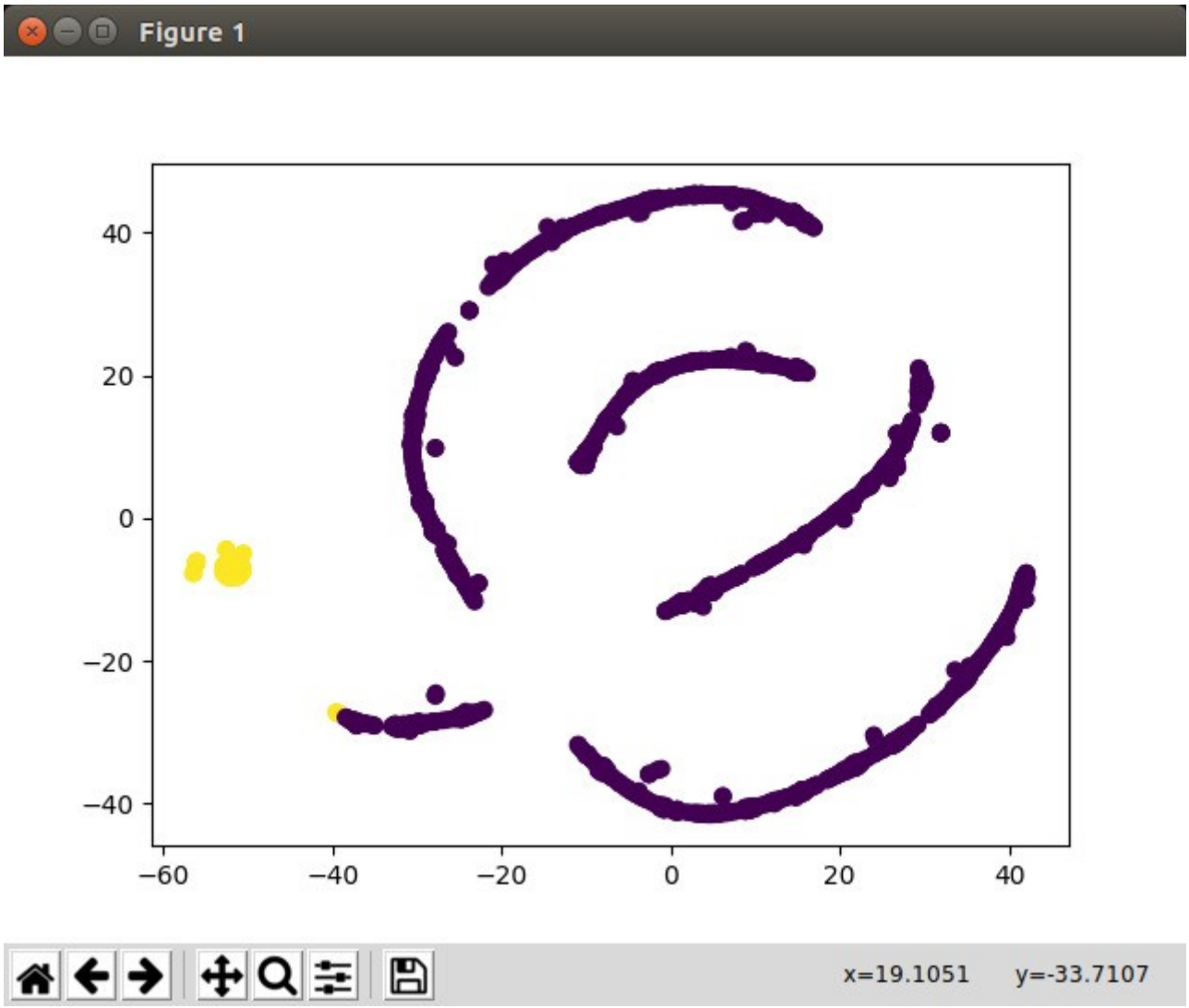
The output after applying the K Means technique, following output were generated and was compared with output from native methods like comparing the following/follower ratio to be less than .5 for the spammers, lesser tweets posted by the spammers etc.



The list of spammer ID's generated according to kmeans is as follows:



Graph of spam tweets vs non spam tweets



purple- non spam
yellow- spam