Tweet Clustering Preliminary Analysis and Clustering Task (Combined):

By: Meghna Shrivastava

September 18, 2019

(This PDF contains a combined description of the preliminary analysis done previously and the clustering results obtained for the second part of the question)**

STEP 1: <u>PREPROCESSING</u>

- I first converted all the 16 text files into the pandas data frames. In order to take care of the encoding, I made use of the 'encoding = 'unicode_escape'' command.
- Performed concatenation of the dataframes to convert into a single dataframe.
- Reduced the dataframe features to the 'FileName' and 'Tweet'. A quick check of the dataframe:

Out[19]:

	FileName	Tweet
0	bbchealth.txt	Breast cancer risk test devised http://bbc.in/...
1	bbchealth.txt	GP workload harming care - BMA poll http://bbc...
2	bbchealth.txt	Short people's 'heart risk greater' http://bbc...
3	bbchealth.txt	New approach against HIV 'promising' http://bb...
4	bbchealth.txt	Coalition 'undermined NHS' - doctors http://bb...
5	bbchealth.txt	Review of case against NHS manager http://bbc....
6	bbchealth.txt	VIDEO: 'All day is empty, what am I going to d...

- I then tried removing all the unnecessary characters and cleaned the tweets using the string command to get tweets free of any undesired/special characters.

Out[21]:

	FileName	Tweet
0	bbchealth.txt	Breast cancer risk test devised httpbbcinCimpJF
1	bbchealth.txt	GP workload harming care BMA poll httpbbcinCh...
2	bbchealth.txt	Short peoples heart risk greater httpbbcinChTANp
3	bbchealth.txt	New approach against HIV promising httpbbcinEjAjt
4	bbchealth.txt	Coalition undermined NHS doctors httpbbcinCnLwK
5	bbchealth.txt	Review of case against NHS manager httpbbcinFfjci
6	bbchealth.txt	VIDEO All day is empty what am I going to do h...
7	bbchealth.txt	VIDEO Overhaul needed for endoflife care httpb...
8	bbchealth.txt	Care for dying needs overhaul httpbbcinFdSGrl
9	bbchealth.txt	VIDEO NHS Labour and Tory key policies httpbbc...

- I then tried combining the rows of the dataframe in such a way that for each Text file we have just one row of tweet string (All the tweets for that text file get concatenated into one row)

Out[22]:

	FileName	Tweet
0	KaiserHealthNews.txt	Tougher Vaccine Law In Calif Clears First Hurd...
1	NBChealth.txt	Ebola Exposure CDC Worker Remains Well httpnbc...
2	bbchealth.txt	Breast cancer risk test devised httpbbcinCimpJ...
3	cbchealth.txt	Drugs need careful monitoring for expiry dates...
4	cnnhealth.txt	An abundance of online info can turn us into e...
5	everydayhealth.txt	FastFood Makes Up Percent of Calories in US D...
6	foxnewshealth.txt	Injury prevention programs unpopular with high...
7	gdnhealthcare.txt	Are you a member of the network Sign up here f...
8	goodhealth.txt	It's not hard to get a little cardio in at hom...
9	latimeshealth.txt	Five new running shoes that aim to go the extr...
10	msnhealthnews.txt	Heavy Coffee Intake May Affect Fertility Treat...
11	nprhealth.txt	Would You Like Health Insurance With Those Sto...
12	nytimeshealth.txt	Risks in Using Social Media to Spot Signs of M...
13	reuters_health.txt	Los Angeles closes medical marijuana shops bu...
14	usnewshealth.txt	Planning to hire a personal trainer Read these...
15	wsjhealth.txt	Parents Start Companies to Cure Children httpo...

Step 2: PERFORM TOKENIZATION

I performed tokenization to convert the string into the unique identification symbols

Out[23]:

	FileName	Tweet
0	KaiserHealthNews.txt	[tougher, vaccine, law, in, calif, clears, fir...
1	NBChealth.txt	[ebola, exposure, cdc, worker, remains, well, ...
2	bbchealth.txt	[breast, cancer, risk, test, devised, httpbbci...
3	cbchealth.txt	[drugs, need, careful, monitoring, for, expiry...
4	cnnhealth.txt	[an, abundance, of, online, info, can, turn, u...
5	everydayhealth.txt	[fastfood, makes, up, percent, of, calories, i...
6	foxnewshealth.txt	[injury, prevention, programs, unpopular, with...
7	gdnhealthcare.txt	[are, you, a, member, of, the, network, sign, ...

STEP 3: STOPWORD

In order to get rid of the common words and other undesired words using

```
from nltk.corpus import stopwords
```

Out[25]:

	FileName	Tweet
0	KaiserHealthNews.txt	[tougher, vaccine, law, calif, clears, first, ...
1	NBChealth.txt	[ebola, exposure, cdc, worker, remains, well, ...
2	bbchealth.txt	[breast, cancer, risk, test, devised, httpbbci...
3	cbchealth.txt	[drugs, need, careful, monitoring, expiry, dat...
4	cnnhealth.txt	[abundance, online, info, turn, us, ehypochond...
5	everydayhealth.txt	[fastfood, makes, percent, calories, us, diet,...
6	foxnewshealth.txt	[injury, prevention, programs, unpopular, high...

In order to get the frequency counts of the top 10 words for each Tweet Account file, I made use of the following package command:

```
In [29]: from collections import Counter
```

To get the dataframe as:

Out[29]:

	FileName	Tweet	Words
0	KaiserHealthNews.txt	[tougher, vaccine, law, calif, clear, first, h...	[('health', 990), ('report', 520), ('insurance...
1	NBChealth.txt	[ebola, exposure, cdc, worker, remains, well, ...	[('new', 386), ('study', 347), ('ebola', 275),...
2	bbchealth.txt	[breast, cancer, risk, test, devised, httpbbci...	[('video', 814), ('nh', 349), ('ebola', 349), ...
3	cbchealth.txt	[drug, need, careful, monitoring, expiry, date...	[('ebola', 455), ('say', 327), ('health', 327)...
4	cnnhealth.txt	[abundance, online, info, turn, u, ehypochondr...	[('rt', 655), ('getfit', 257), ('health', 251)...

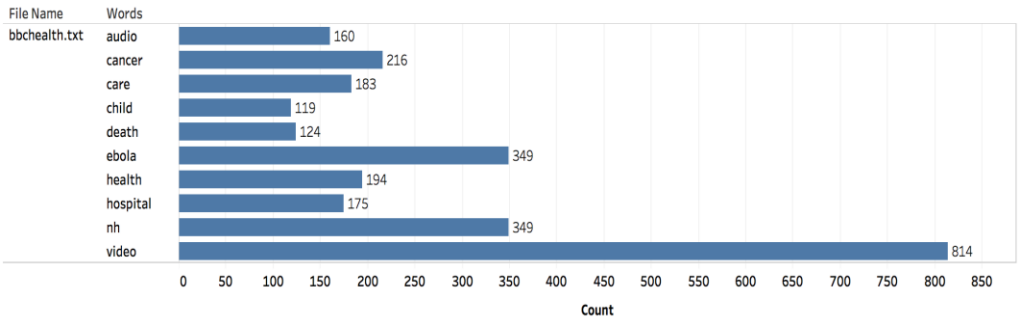
In the above image, you can see that I was able to derive the frequency count of words for each File. I then modified the code to consider top 10 words based on the frequency count.

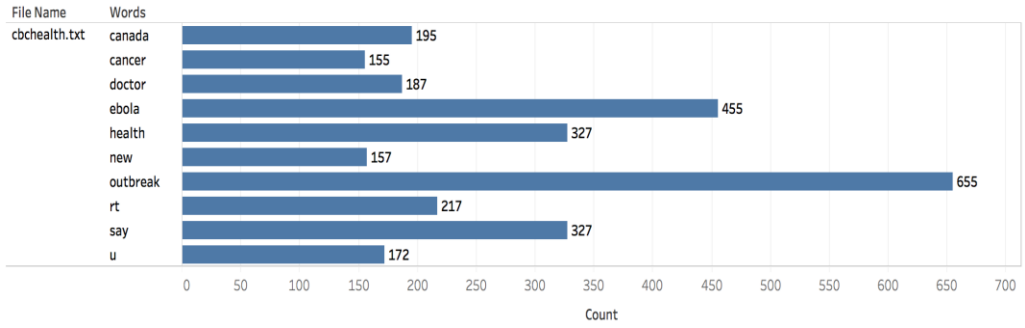
I then converted the final data frame into the CSV file to visualize the words using Tableau.

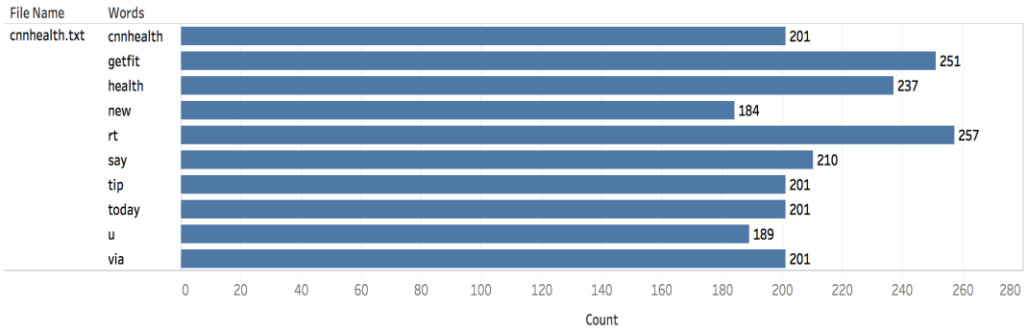
Below are some of the file's visuals that were generated

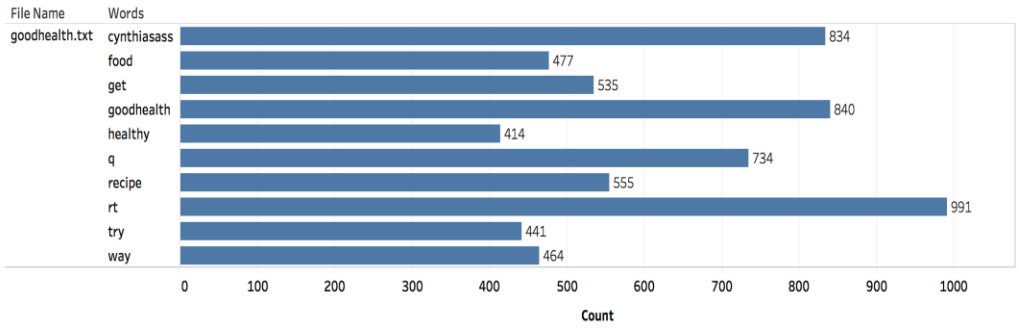
Visualization:

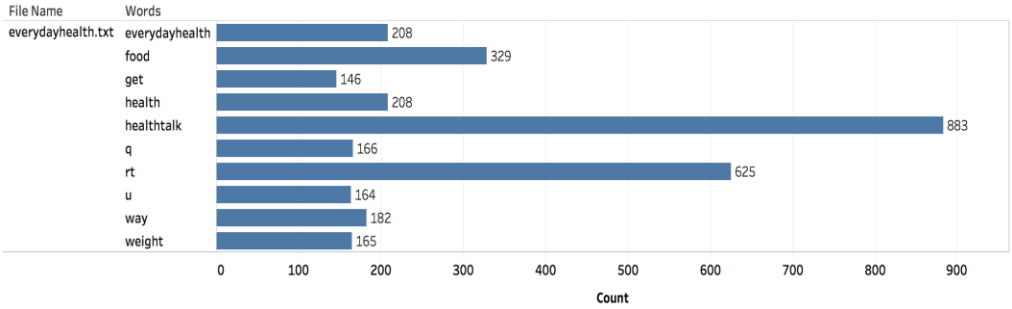
1a. Below are the most common top 10 words visualizations from all the files using a bar plot.

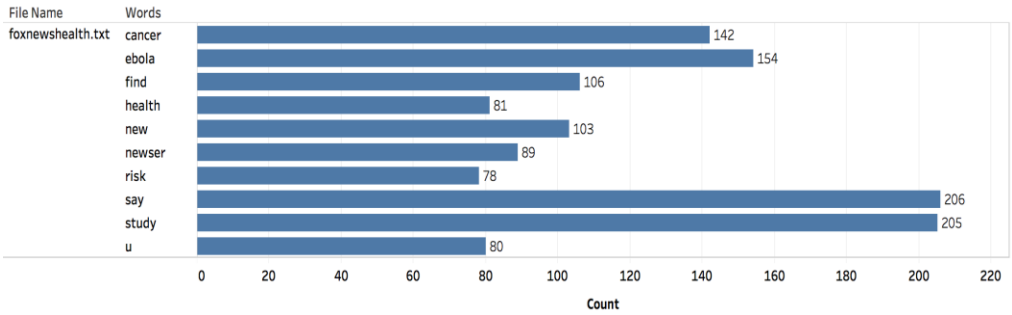


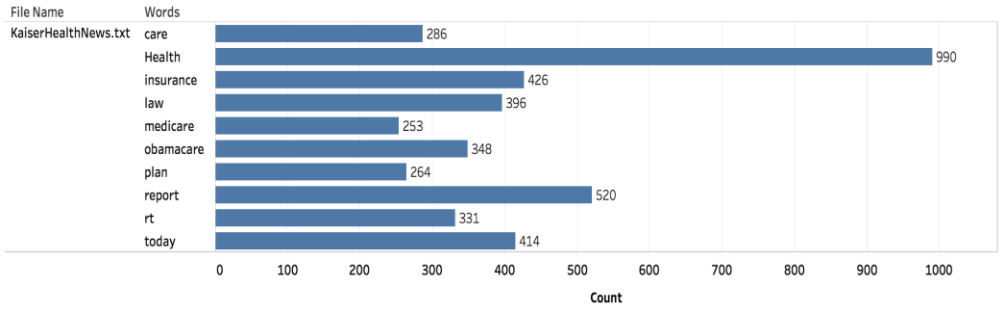


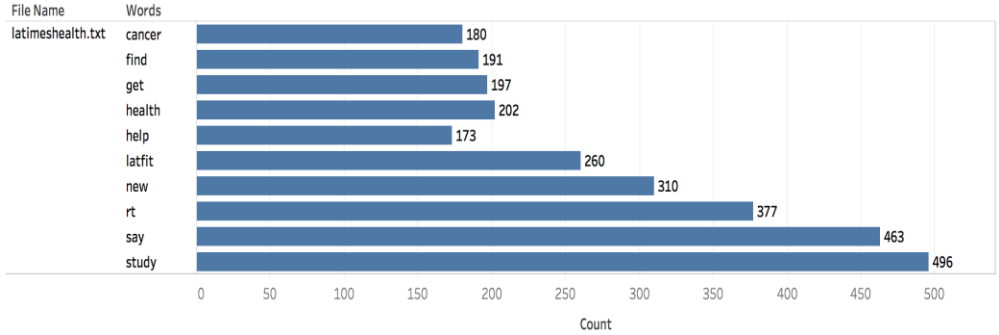


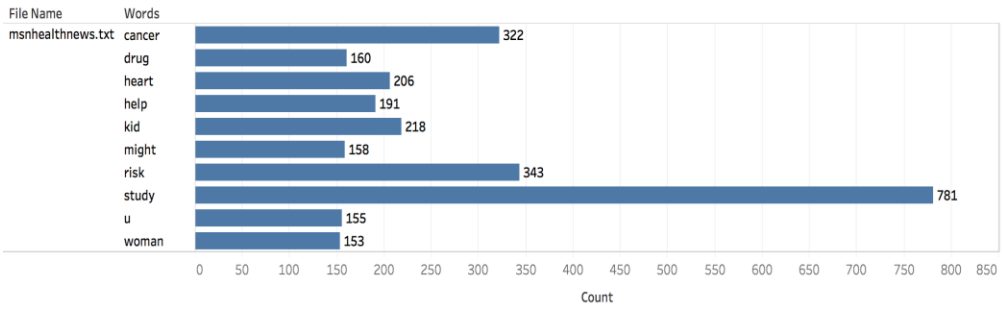

File Name
Words
goodhealth.txt
cynthiasass 834
food 477
get 535
goodhealth 840
healthy 414
q 734
recipe 555
rt 991
try 441
way 464
0 100 200 300 400 500 600 700 800 900 1000
Count

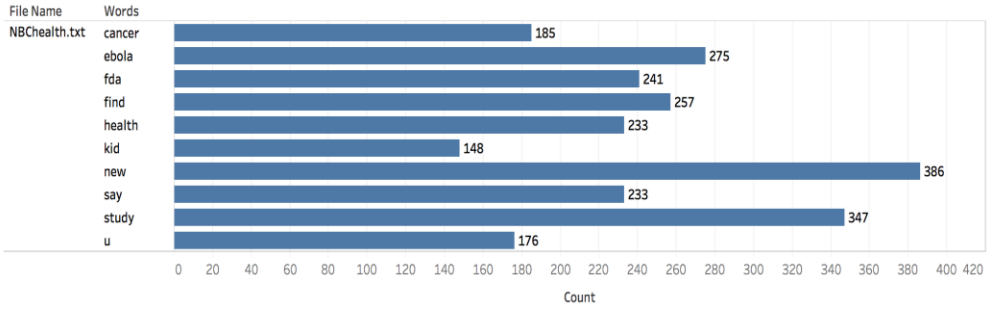

File Name
Words
everydayhealth.txt
everydayhealth 208
food 329
get 146
health 208
healthtalk 883
q 166
rt 625
u 164
way 182
weight 165
0 100 200 300 400 500 600 700 800 900
Count

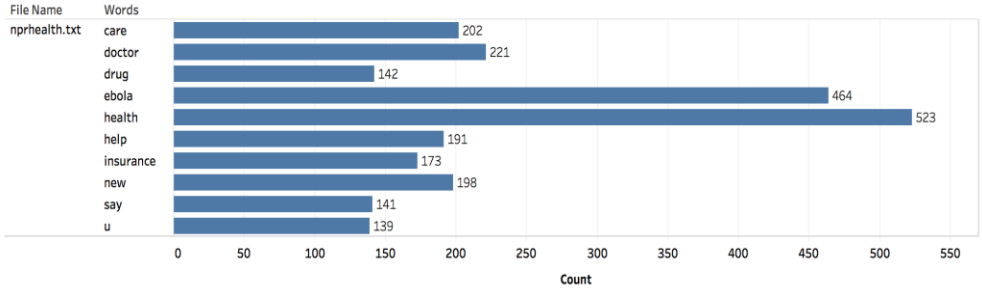

File Name
Words
foxnewshealth.txt
cancer 142
ebola 154
find 106
health 81
new 103
newser 89
risk 78
say 206
study 205
u 80
0 20 40 60 80 100 120 140 160 180 200 220
Count

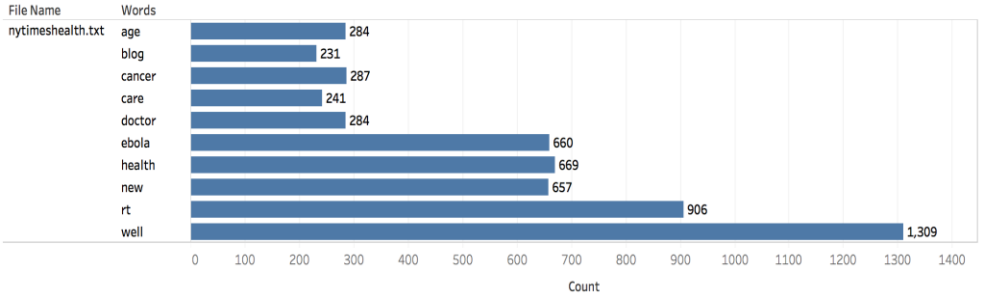

File Name
Words
KaiserHealthNews.txt
care 286
Health 990
insurance 426
law 396
medicare 253
obamacare 348
plan 264
report 520
rt 331
today 414
0 100 200 300 400 500 600 700 800 900 1000
Count

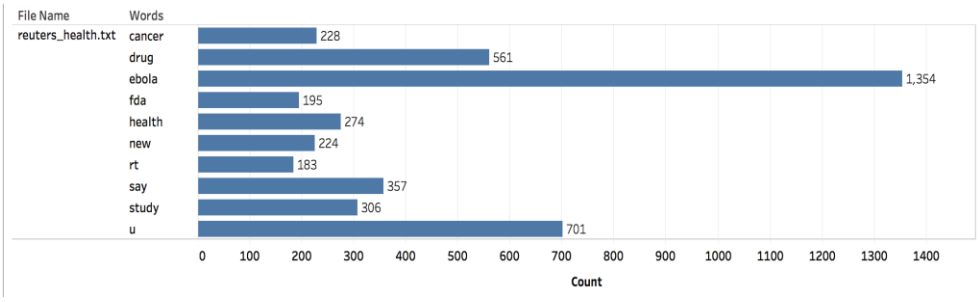

File Name
Words
latimeshealth.txt
cancer 180
find 191
get 197
health 202
help 173
latfit 260
new 310
rt 377
say 463
study 496
0 50 100 150 200 250 300 350 400 450 500
Count

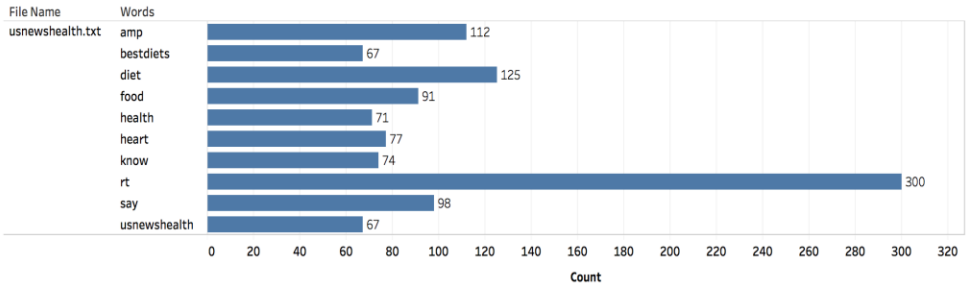

File Name
Words
msnhealthnews.txt
cancer
drug
heart
help
kid
might
risk
study
u
woman
322
160
206
191
218
158
343
781
155
153
0 50 100 150 200 250 300 350 400 450 500 550 600 650 700 750 800 850
Count

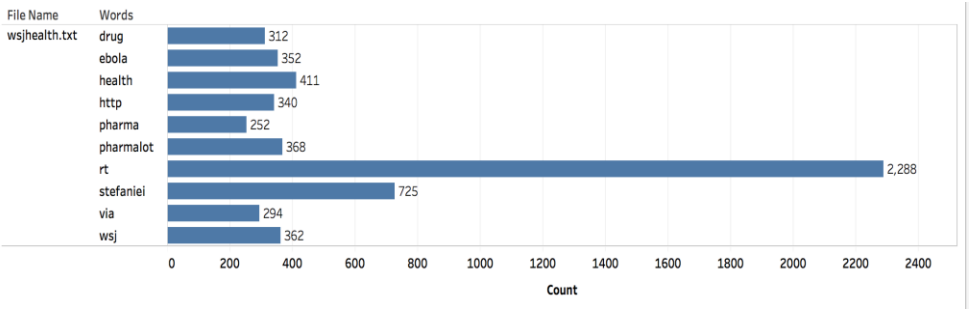

File Name
Words
NBChealth.txt
cancer
ebola
fda
find
health
kid
new
say
study
u
185
275
241
257
233
148
386
233
347
176
0 20 40 60 80 100 120 140 160 180 200 220 240 260 280 300 320 340 360 380 400 420
Count


File Name
Words
nprhealth.txt
care
doctor
drug
ebola
health
help
insurance
new
say
u
202
221
142
464
523
191
173
198
141
139
0 50 100 150 200 250 300 350 400 450 500 550
Count


File Name
Words
nytimeshealth.txt
age
blog
cancer
care
doctor
ebola
health
new
rt
well
284
231
287
241
284
660
669
657
906
1,309
0 100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400
Count


File Name
Words
reuters_health.txt
cancer
drug
ebola
fda
health
new
rt
say
study
u
228
561
1,354
195
274
224
183
357
306
701
0 100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400
Count





1b. Are these most probable words related to health?

Not every word is related to health.

1c. If not, can you propose a way to improve the results?

I possibly would like to invest more time in creating stopwords to get rid of the words that doesn't have any relevance in addition to making use of the CounterVectorizer and Bigram techniques to improve the results.

Task 2: Clustering

2a. Combine tweets from all 16 accounts and cluster the tweets with K-means clustering, using K = 16. Graph these clusters in two dimensions. You may choose to use principal components analysis (PCA) or another method of your choice to choose two dimensions for visualization.

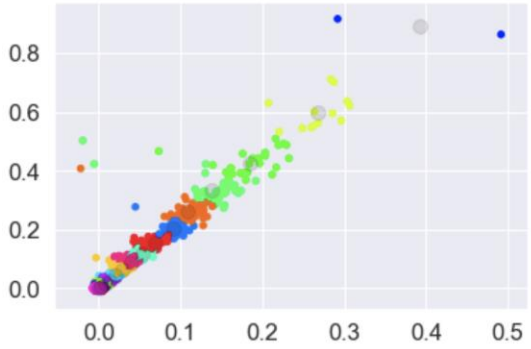
Solution: In my preliminary analysis of all the 16 files given, I had made use of the CounterVectorizer method to count the number of times a word occurs in a corpus.

I made of the 'gensim' package to import 'WordtoVec' in order to convert the words into the vector form.

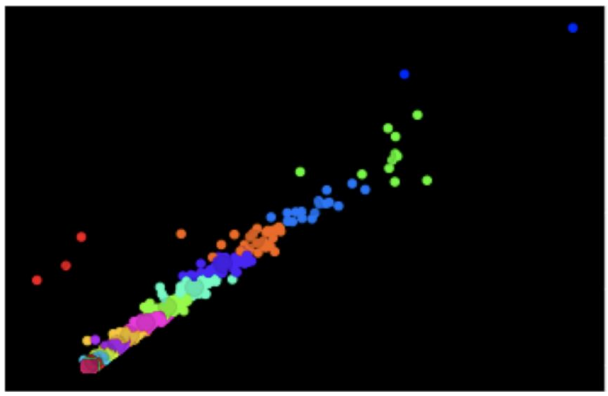
I then performed K-means clustering and obtained the following results:

```
In [83]: plt.scatter(X[:, 0], X[:, 1], c=y_kmeans, s=20, cmap='hsv')

         centers = km.cluster_centers_
         plt.scatter(centers[:, 0], centers[:, 1], c='black', s=100, alpha=0.1);
```



Which basically looks like (Overlapping of the above results can be clearly seen in the dark background)



2b. Does each Twitter account form its own cluster? No

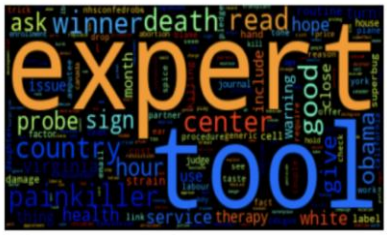
2c. Why or why not is this the case?

Solution:

We can see the overlapping of the data points for different clusters that are obtained after performing K-Means clustering. This means that there are no independent clusters formed for each tweet account. This also tells us that some words are distributed across different clusters, means they are tweeted by different tweet accounts.

I also dived deep into visualizing certain words that are part of each cluster (K=16 clusters) in order to validate the above scatter plot. I tried checking whether there is an existence of the words in more than one cluster or not:

Visuals generated for Clusters 1 to 16: In sequence (Left ---To --- Right)



Observation:

We can see that there are certain words that are repeated in different clusters (E.g. Word **'Obama'** is present is Cluster 1 and Cluster 12 results, Word '**Nomination**' present in Cluster 3 and Cluster 11 and so on...).