ASSIGNMENT 1: Clustering

Machine Learning Team Projects - IS590ML || Group1 || Group Members: Paulami Ray, Saumil Shah

Visualization Task:

We imported all 16 datasets in Jupyter Notebook and used Python to analyze the dataset and visualize it.

Steps followed:

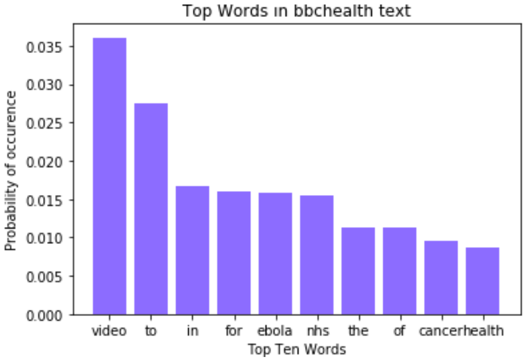
1. Read the csv files one by one and delimited them by '|' operators.
2. We did basic data preprocessing and picked up the tweet text bit from the dataset and discarded the rest of the columns.
3. We used the gensim.utils.simple_preprocess to convert the data into list of tokens. It lowercases the letters, tokenizes them and with deacc property is set to TRUE, it removes punctuations.
4. We applied lemmatization using WordNetLemmatizer() to convert the words to their root words.
5. We save the data words in a list and then create a dictionary to take the count of the individual words.
6. We sorted the counts of the words in a descending order and print the top 10 words from each tweet document.
7. We then visualized the top 10 words in a histogram for each of the tweet document.
   Observation: We observed that most of the words that appeared in the document were common English words and words like {"video","audio" etc}.
   Then we decided to remove the stopwords to check whether the words related to health appears as the top words.
8. We then removed the stopwords from nltk.corpus as well as some hardcoded nonrelevant repetitive words like {"video", "audio","rt"}.
   Observation: We observed that health related words now appear majorly among the top ten words for each of the tweet accounts.

Q. Are these most probable words related to health?

Solution: No, initially without the removal of stop words and frequently occurring non-relevant words, the top 10 words list did not contain many health - related words. We can refer to the below visualizations to get an overview of the words that appeared.

Visualizations:

Fig 1 – Fig 16 - Top Ten Words before removing stop words



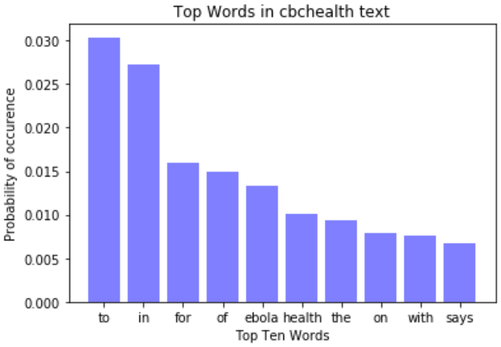
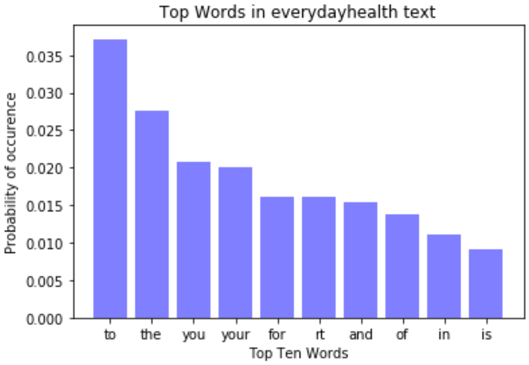


Fig (1)

Fig (2)



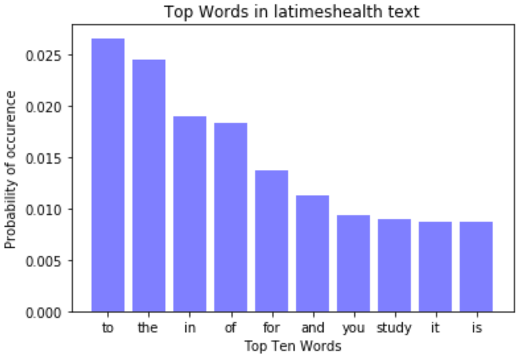
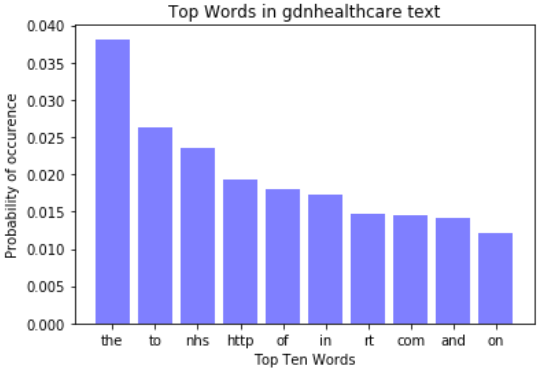


Fig (3)

Fig (4)



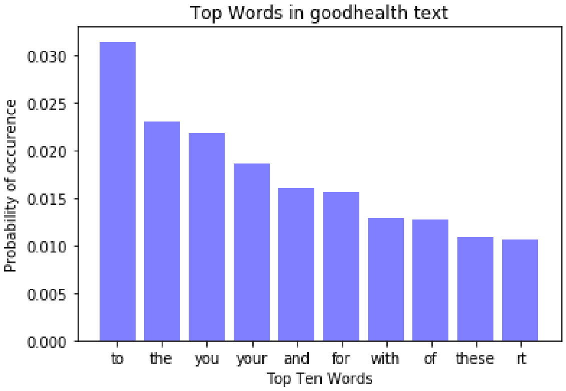


Fig (5)

Fig (6)

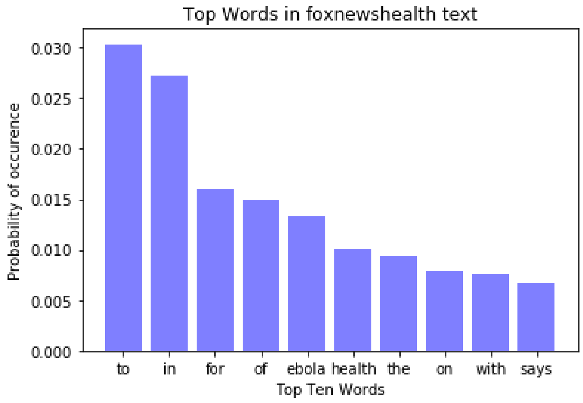


Fig (7)

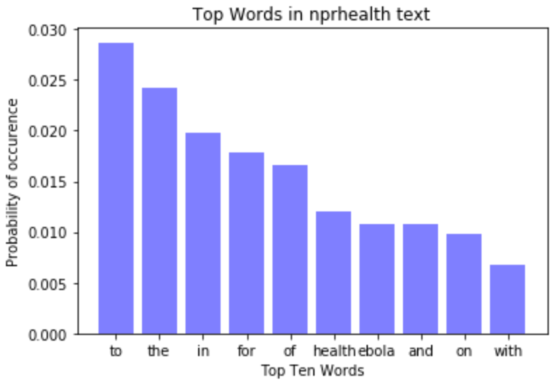


Fig (8)

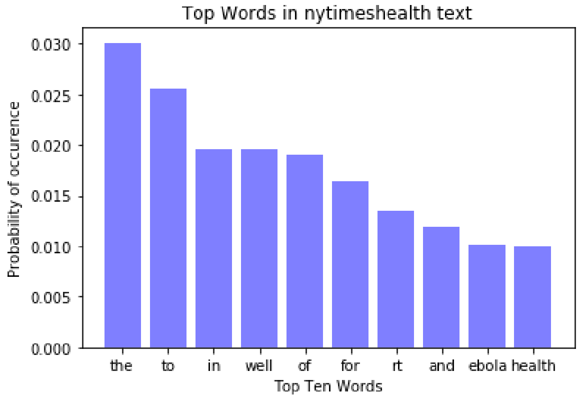


Fig (9)

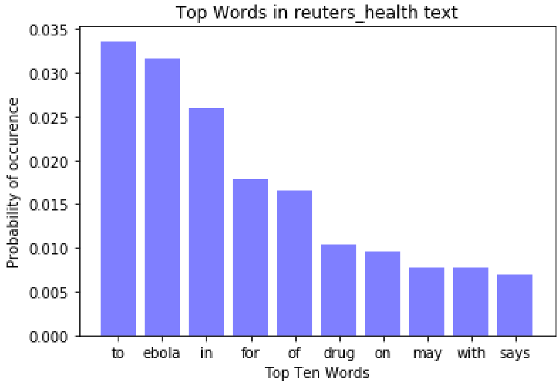


Fig (10)

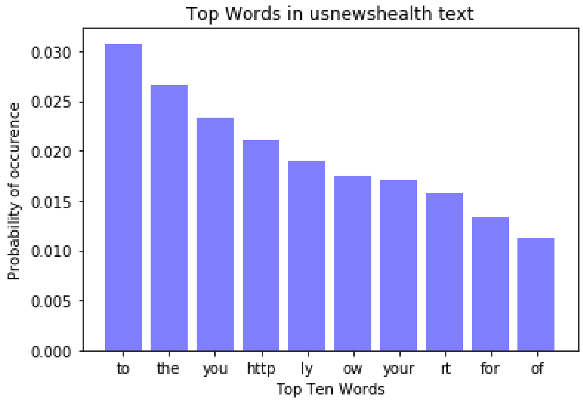


Fig (11)

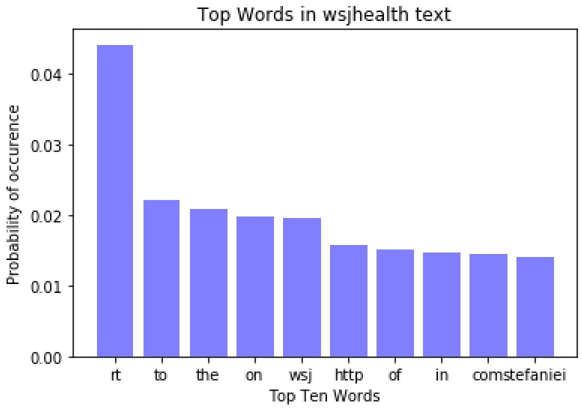


Fig (12)

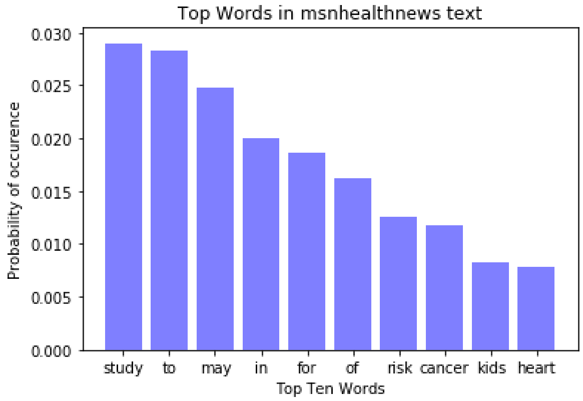


Fig (13)

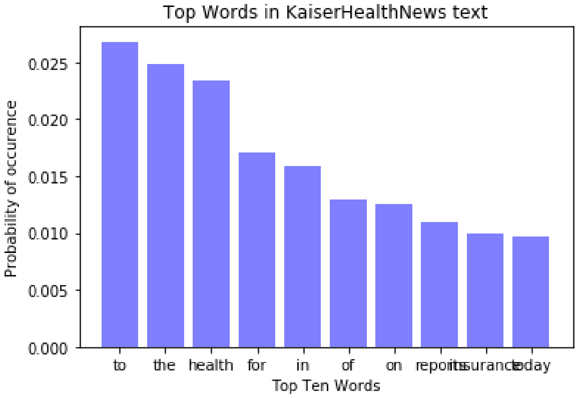


Fig (14)

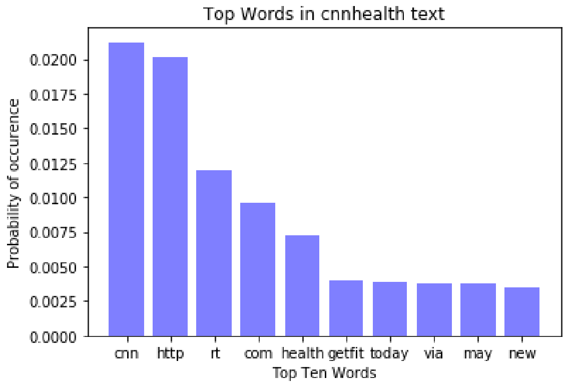


Fig (15)

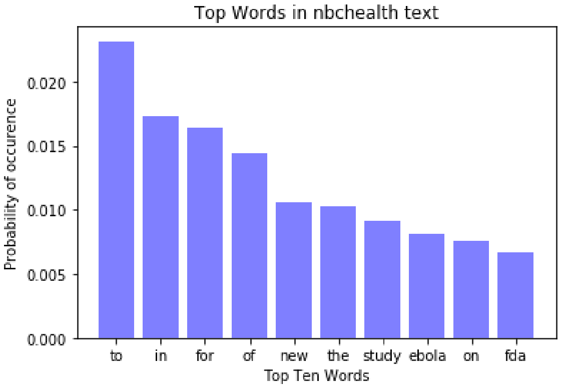


Fig (16)

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

Solution: We removed the stop words and found that most of the words in the top ten list contains health related words. This definitely improved the results.

We can refer to the below visualizations for the same.

Visualizations:

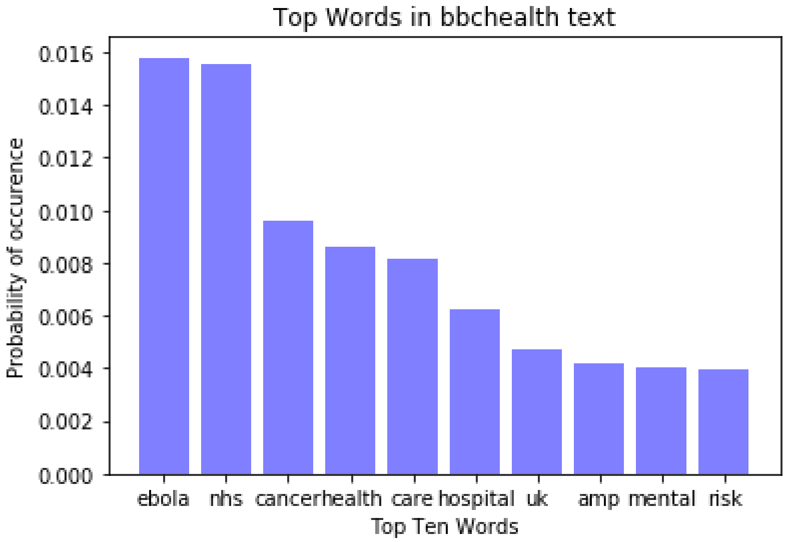


Fig (17)

Here in Fig 17, we can clearly see that the words like {ebola, nhs, cancer, health, care, hospital, mental, risk} are all related to health and these words are coming frequently in the corpus once we removed our stop words.

Similarly, we can see the rest of the tweet documents which has similar results.

Fig 18 – Fig 32 - Top Ten Words after removing stop words

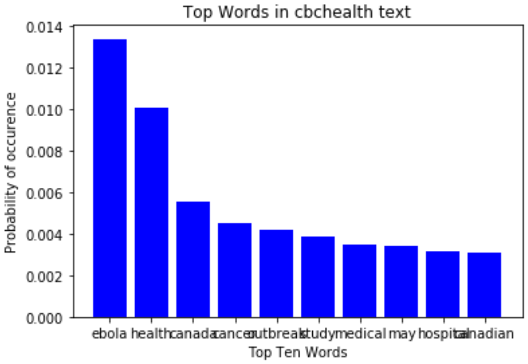


Fig (18)

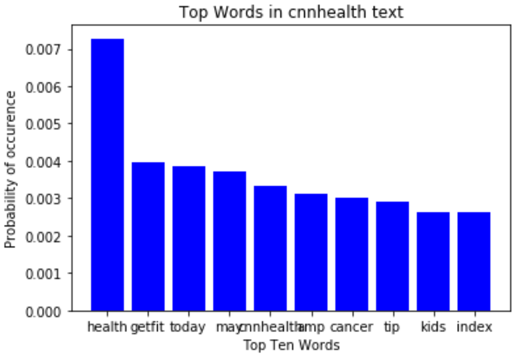
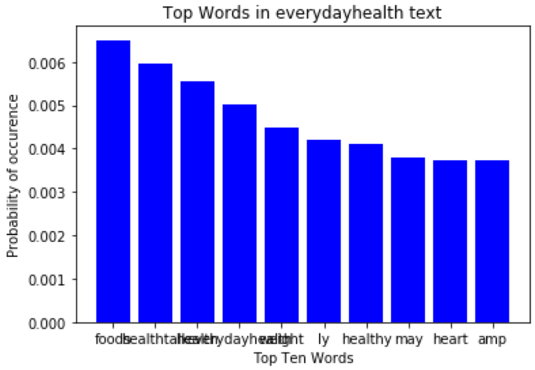


Fig (19)



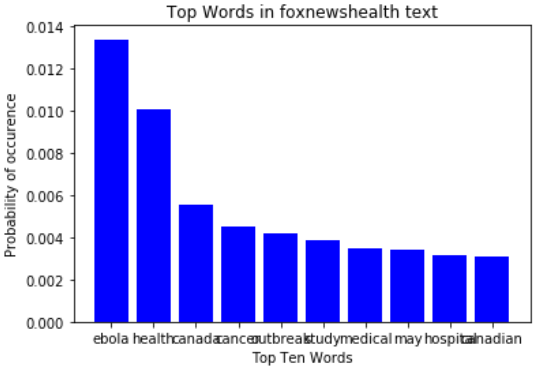
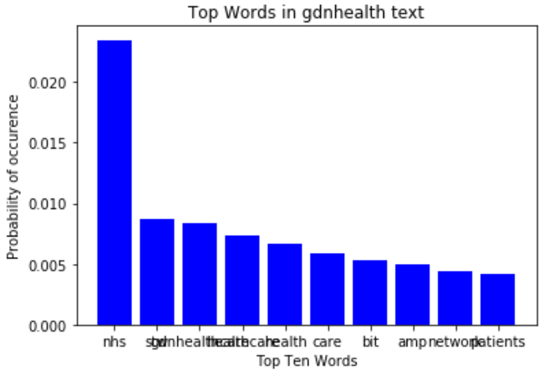


Fig (20)

Fig (21)



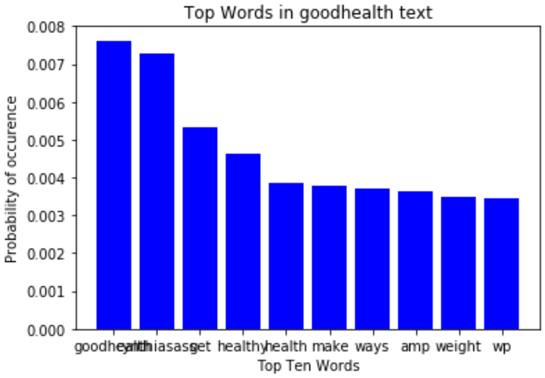
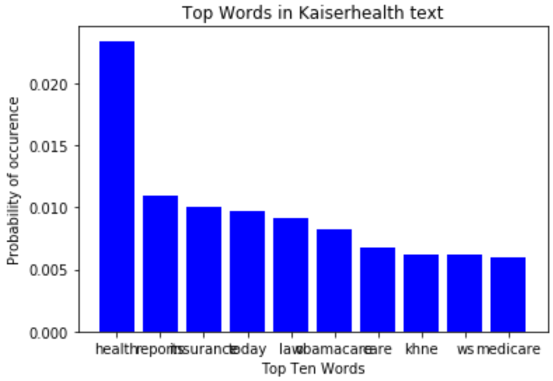


Fig (22)

Fig (23)



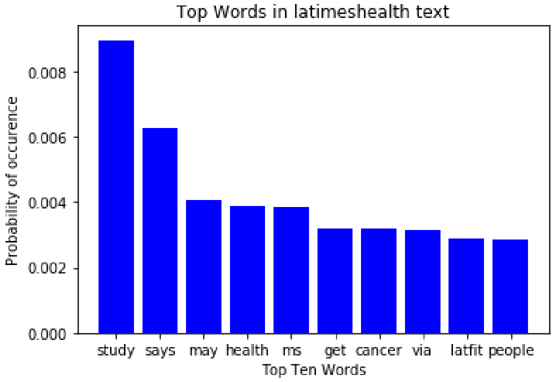


Fig (24)

Fig (25)

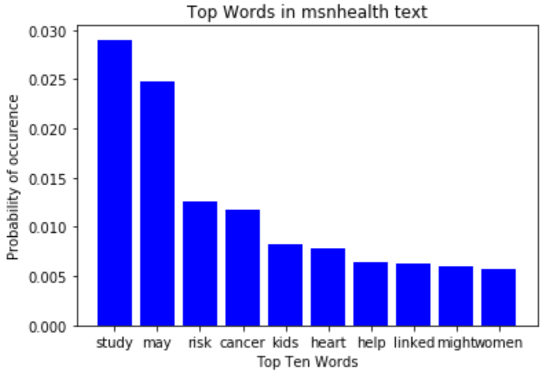


Fig (26)

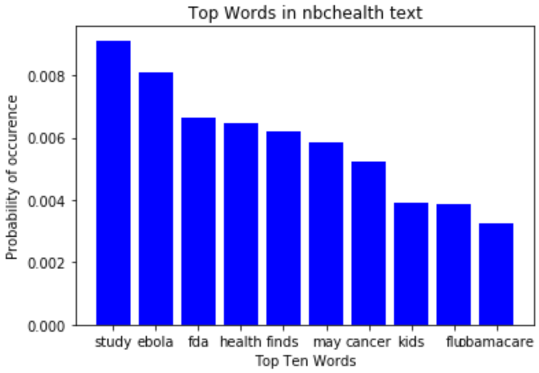


Fig (27)

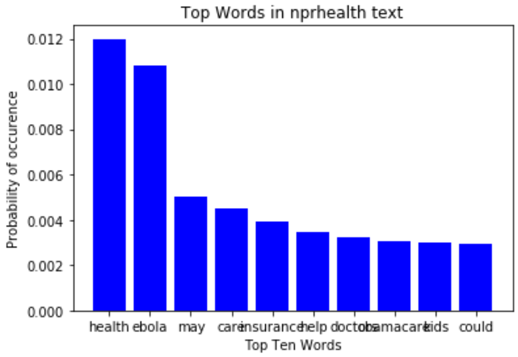


Fig (28)

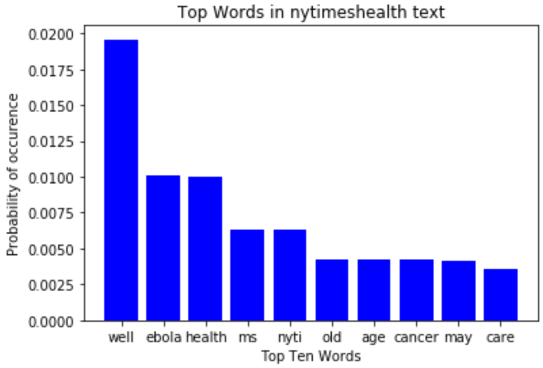


Fig (29)

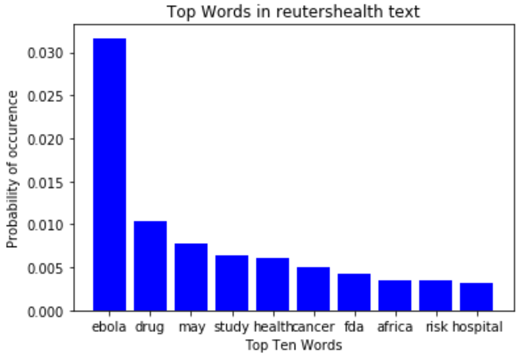


Fig (30)

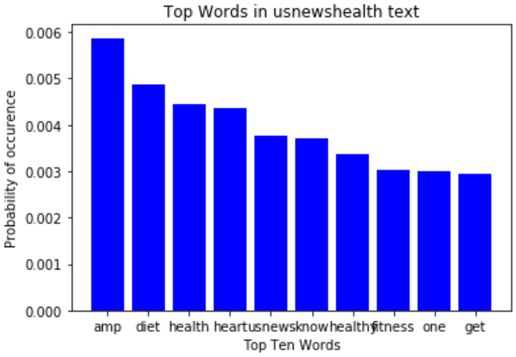


Fig (31)

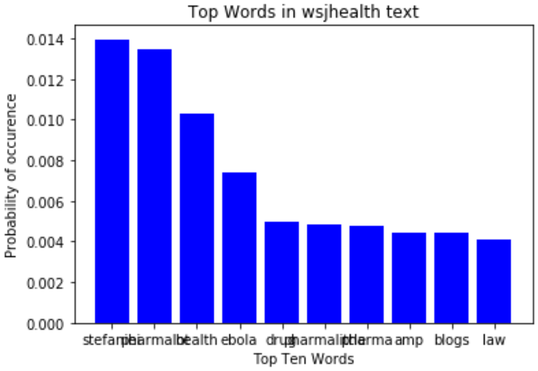
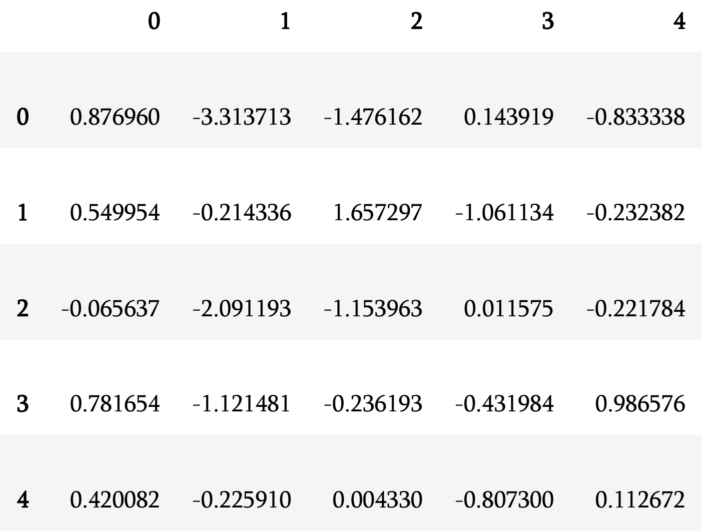


Fig (32)

Clustering Task:

Steps followed:

1. We combined all the tweets from 16 accounts and concatenated them into a data frame.
2. We again performed the basic data preprocessing as we did in the visualization task and removed the stop words.
3. We applied TF-IDF vectorization to understand how important a particular word is in the corpus of the tweets and to minimize on the number of features we removed the unnecessary redundant words by keeping the max_features parameter = 500.
4. Next, we standardized the features by removing the mean and scaling the variance.
5. We understand that the main goal of PCA (Principle Component Analysis) is to reduce the dimensionality of the original feature space by projecting it onto a smaller project space. PCA helps us to find the variables which can be dropped because their influence is negligible. For this model since the number of features are very large we had to do PCA before clustering.
6. For choosing the number of components for PCA we had to understand the tradeoff between the number of features and explained variance. For our model, we finally took the number of features as 5. The below table shows the data frame for the principle components.



	0	1	2	3	4
0	0.876960	-3.313713	-1.476162	0.143919	-0.833338
1	0.549954	-0.214336	1.657297	-1.061134	-0.232382
2	-0.065637	-2.091193	-1.153963	0.011575	-0.221784
3	0.781654	-1.121481	-0.236193	-0.431984	0.986576
4	0.420082	-0.225910	0.004330	-0.807300	0.112672

7. Finally, we did K means clustering with number of cluster =16 as asked in the question. We choose 16 different colors from the palette to clearly distinguish the clusters.
   LABEL_COLOR_MAP = {0:'red', 1: 'green', 2: 'blue', 3:'yellow', 4: 'violet' , 5: 'black', 6:'darkorange', 7: 'olive', 8: 'chocolate', 9:'magenta', 10: 'maroon', 11: 'cyan', 12:'khaki', 13: 'grey', 14: 'wheat', 15: 'red'}

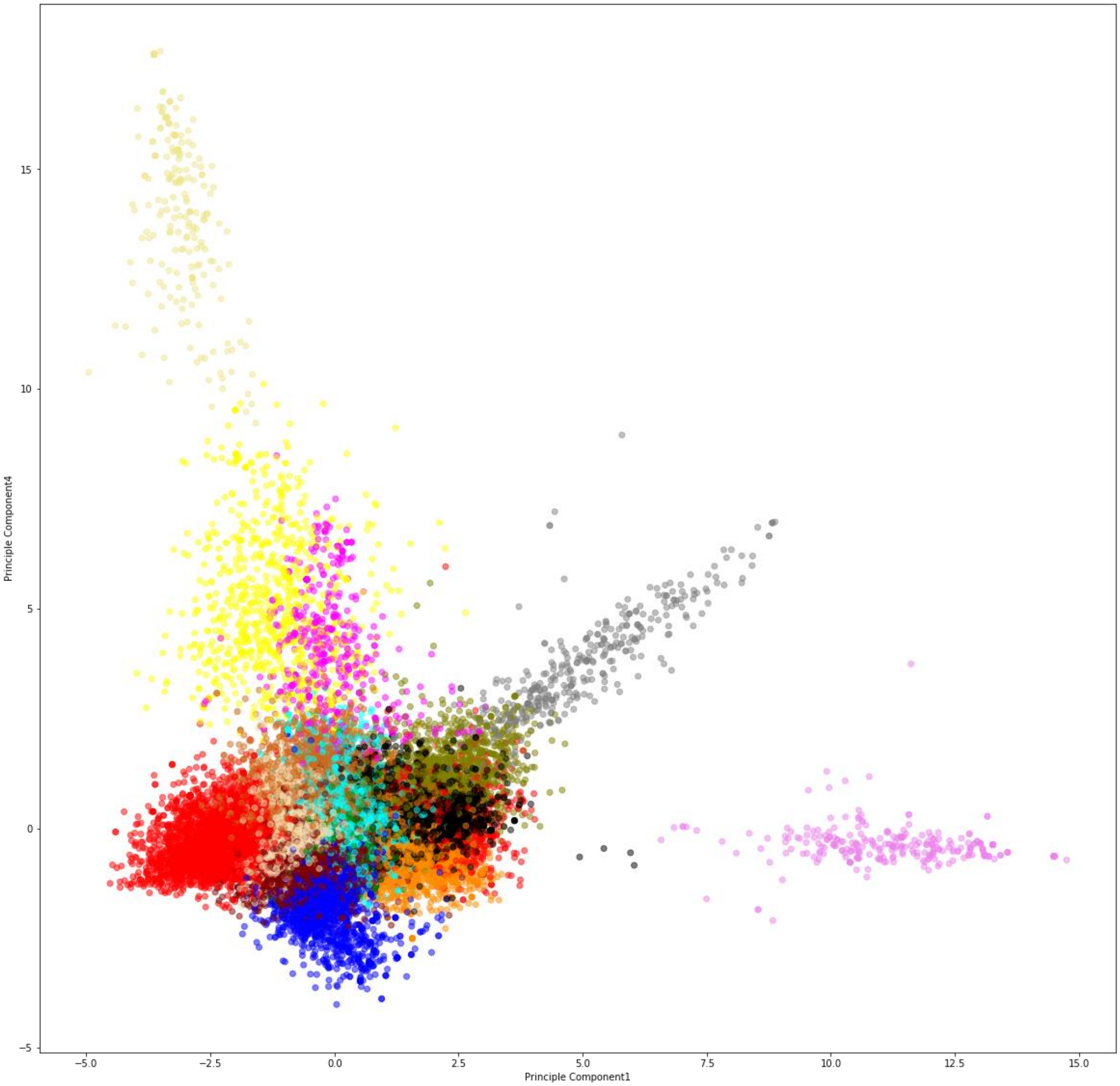


Fig (33) Clustered models

8. We used the PCA components to do the visualization in the clustered model. We tried otherwise as well to do the clustering before doing any principle component analysis, but the result did not vary to a great extent. Therefore, we went with the first approach.

Q. Does each Twitter account form its own cluster? Why or why not is this the case?

Solution: Each Twitter account does not form its own cluster as can be seen from the above plotting. There are words that are common to different clusters, making it difficult to cluster the words based on the twitter accounts. All the tweets are based on health-related topics, so the clusters were formed on the basis of familiarity of the context of the words and the Euclidean distance between these words and not on the basis of the twitter files they belonged to.

Optional bonus task

Q. Change one of the parameters used in the clustering task. How do your results differ, and why?

Solution: We chose the number of PCA components as 5 for our model and after applying the clustering with cluster size as 16 we studied the output in a two-dimensional model. To achieve this, we tried different combinations of our PCA components and for each selection the final cluster representation was different. In the above plot, we selected the combination with principle component 1 and 4 which seemed to be the most accurate one for our model.