Assignment 1 (Unsupervised Learning) - Final report

- Akhila Khanapuri & Yujia Wang

Task 1: Visualization Task

Steps

1. Data Processing

Extracted the content of each twitter account, and remove the URL, punctuations.

2. Calculation

Counted the number of the words and the frequency of each word. Pick the Top 10 words and check if they related to health.

3. Elaborated Processing

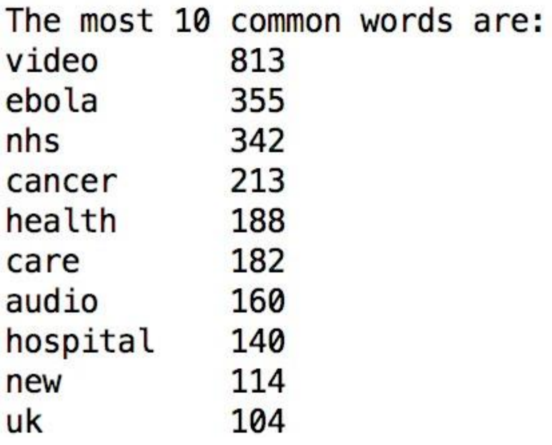
Using the nltk package, Removes the stop words.
Tokenized the text

4. Visualization

Using matplotlib, plotted graph of the probability of occurrence for the 10 most common words.

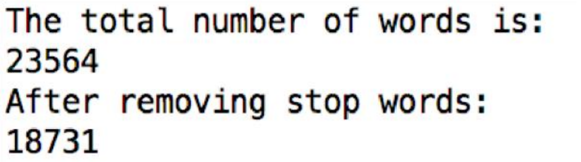
- **Results**

The 10 most common words and their word count (Take 'bbchealth' for example）



```
The most 10 common words are:
video       813
ebola       355
nhs         342
cancer      213
health      188
care        182
audio       160
hospital    140
new         114
uk          104
```

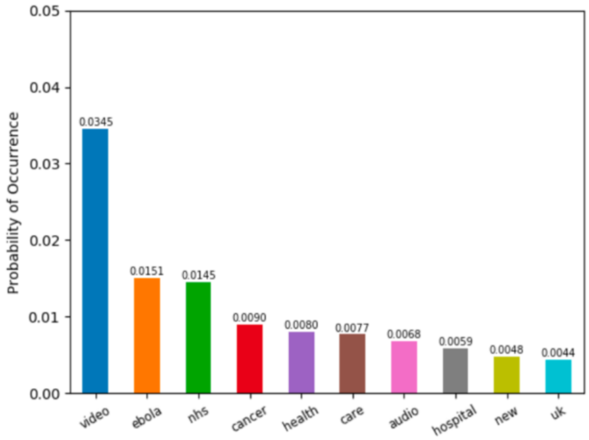
The total number of the words and after removing stop words （Take 'bbchealth' for example）



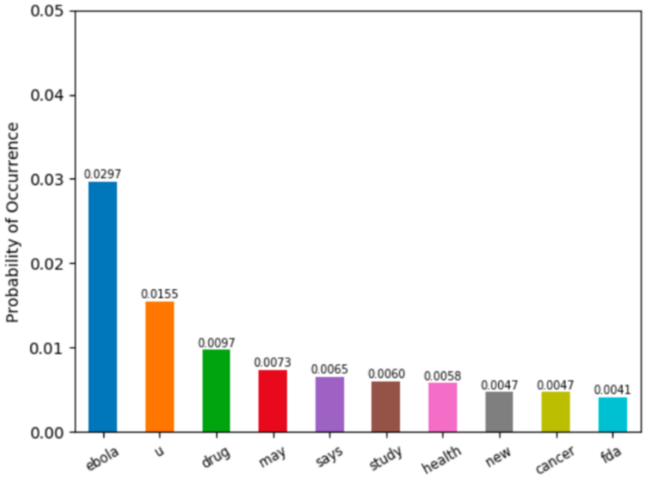
```
The total number of words is:
23564
After removing stop words:
18731
```

The probability of occurrence for the 10 most common words

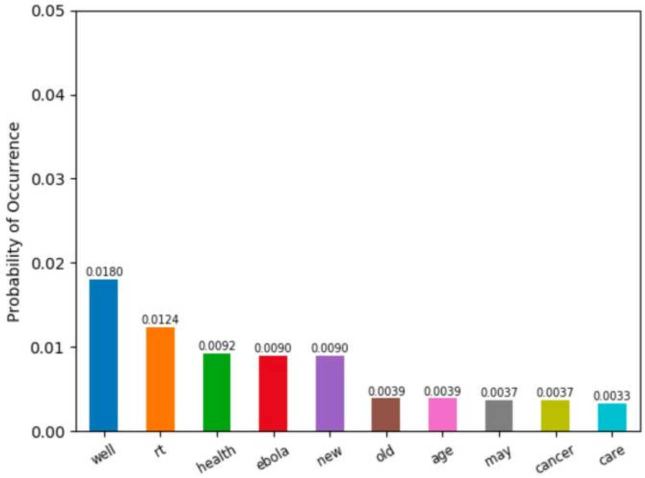
Bbchealth：



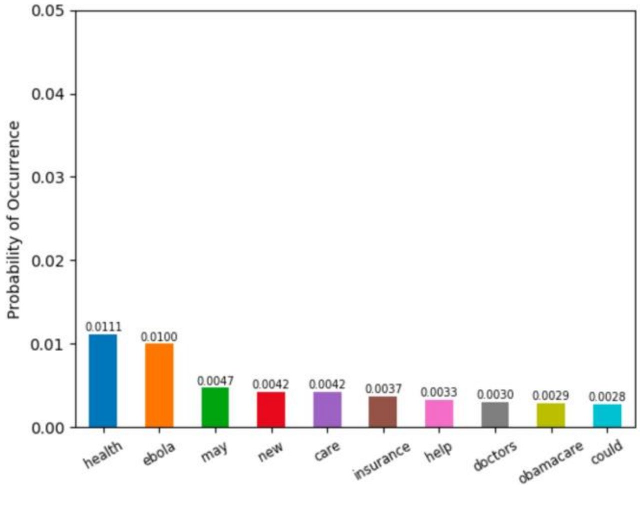
Reuters_health：



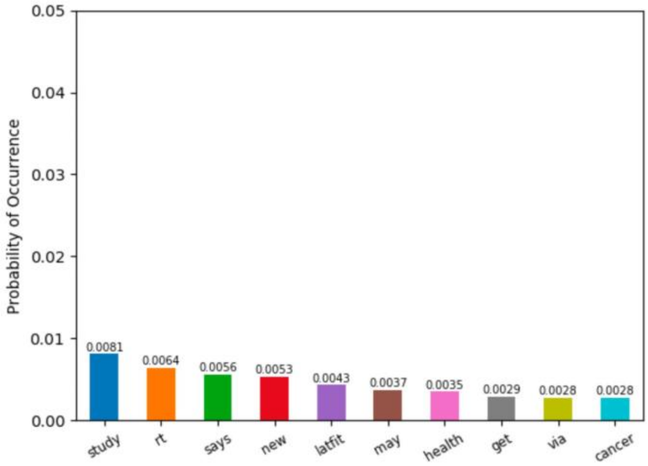
Nytimeshealth:



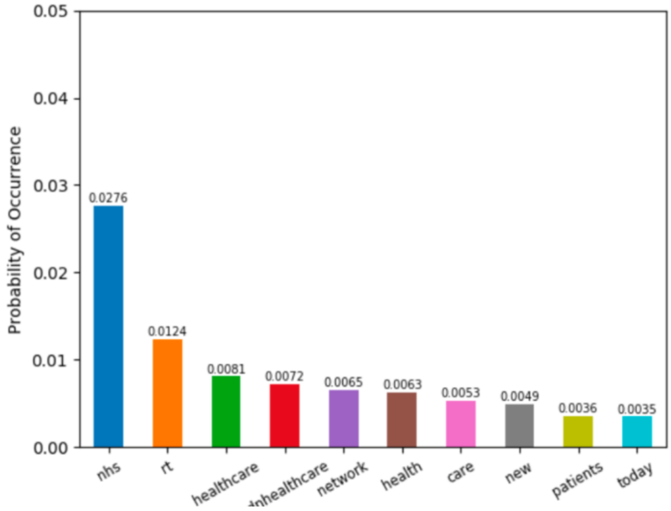
Nprhealth:



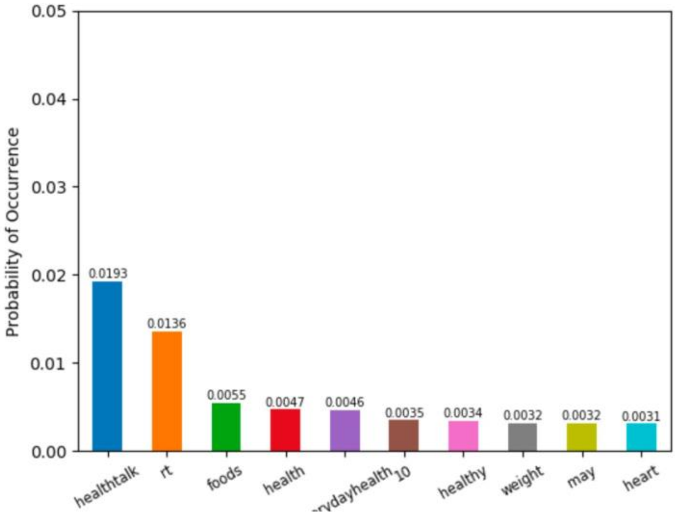
Latimeshealth:



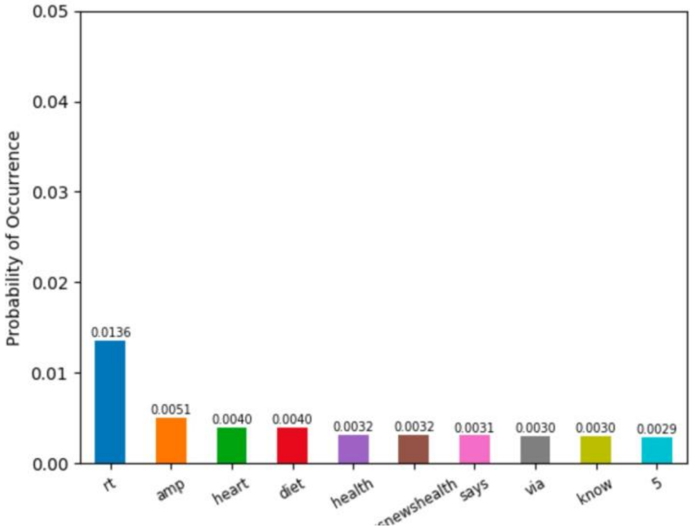
gdnhealthcare:



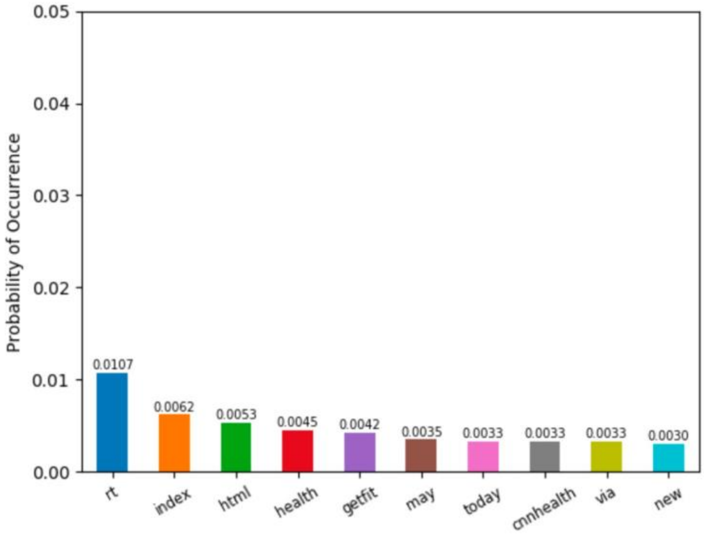
Everydayhealth:



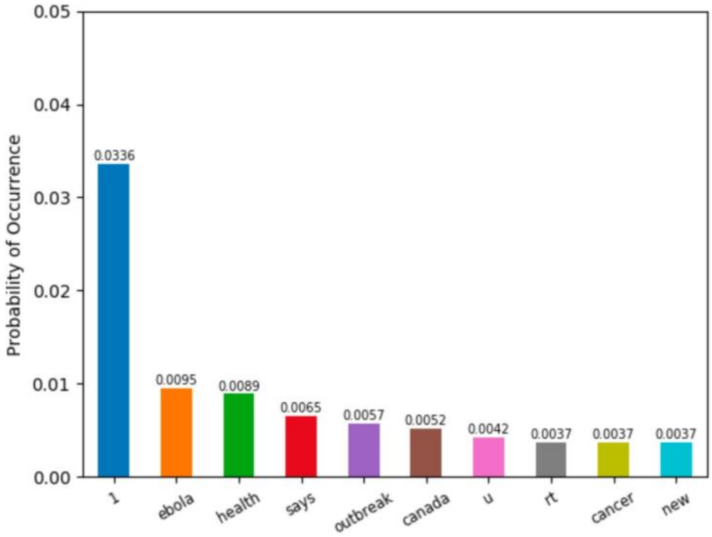
Usnewshealth:



Cnnhealth:



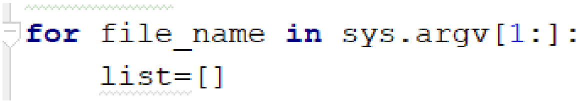
Cbchealth:



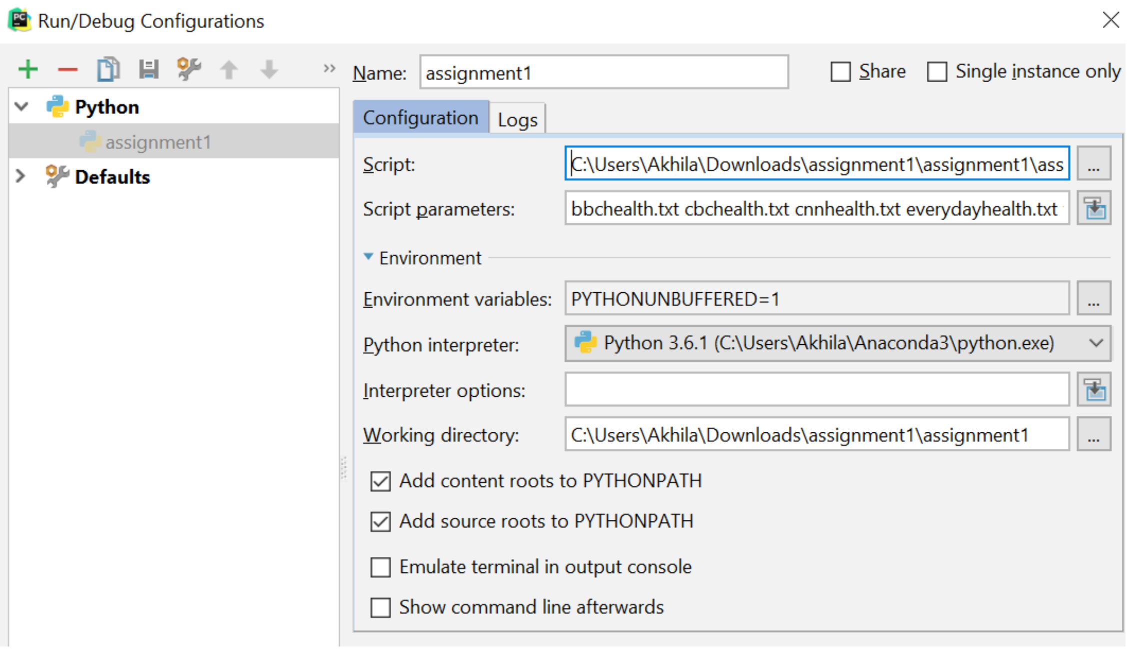
Task 2: Clustering Task

Steps

1. Files import
   Inputted all the files as script parameters to the program.



```
for file_name in sys.argv[1:]:
    list=[]
```



Name: assignment1 □ Share □ Single instance only

Script: C:\Users\Akhila\Downloads\assignment1\assignment1\ass

Script parameters: bbchealth.txt cbchealth.txt cnnhealth.txt everydayhealth.txt

Environment variables: PYTHONUNBUFFERED=1

Python interpreter: Python 3.6.1 (C:\Users\Akhila\Anaconda3\python.exe)

Interpreter options:

Working directory: C:\Users\Akhila\Downloads\assignment1\assignment1

☑ Add content roots to PYTHONPATH

☑ Add source roots to PYTHONPATH

□ Emulate terminal in output console

□ Show command line afterwards

2. Data preprocessing

   Extracted the content of each twitter account, and remove the URL, punctuations.
   Added original labels to assign value to tweets belonging to one Twitter account.

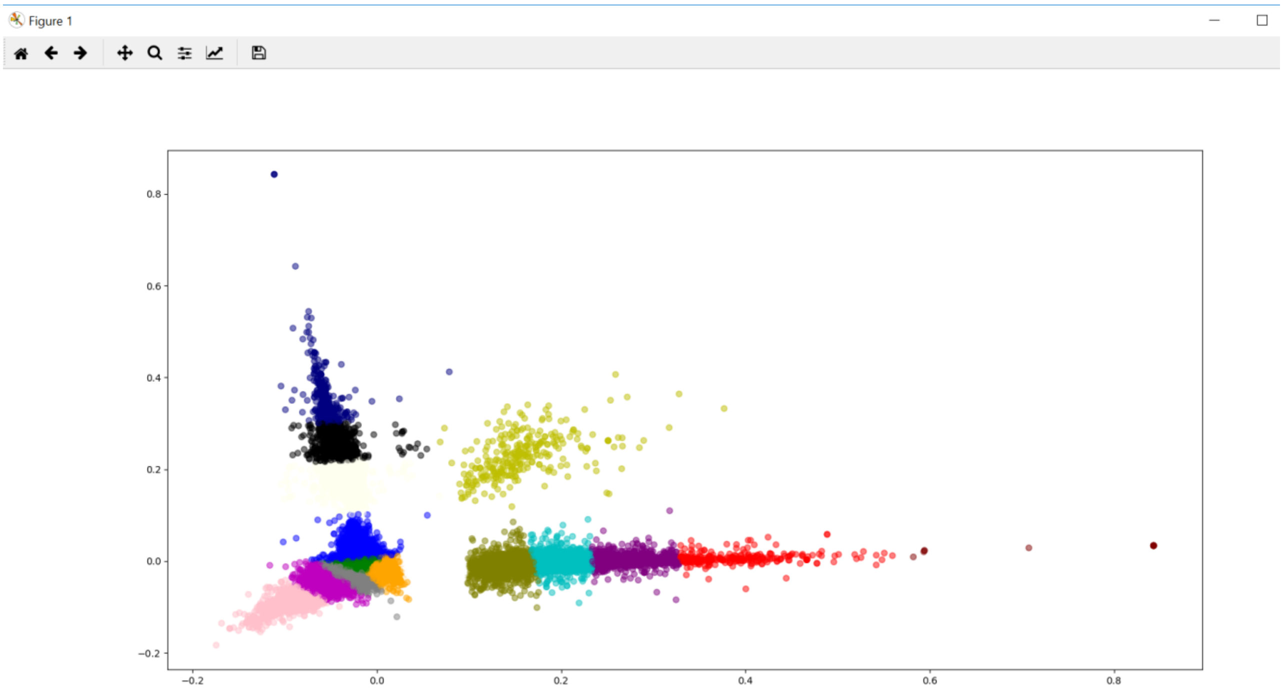
3. Vectorization of texts - TfidfVectorizer

   Made a vector respresentation of all transformed tweets in the data set, with tfidf technique.
   Limited the features to 5000.
   Converted to array for dimensionality reduction

4. Principal Component Analysis & K means
   As mentioned in the assignment document, fixed 2 dimensions for clustering.
   Used tfidf matrix as input to K means.
   The number of clusters were fixed to 16 as mentioned in the assignment document.

5. Comparison of Orginal Label vs Cluster Label
   Compared the labeling of the tweets of cluster and original text.

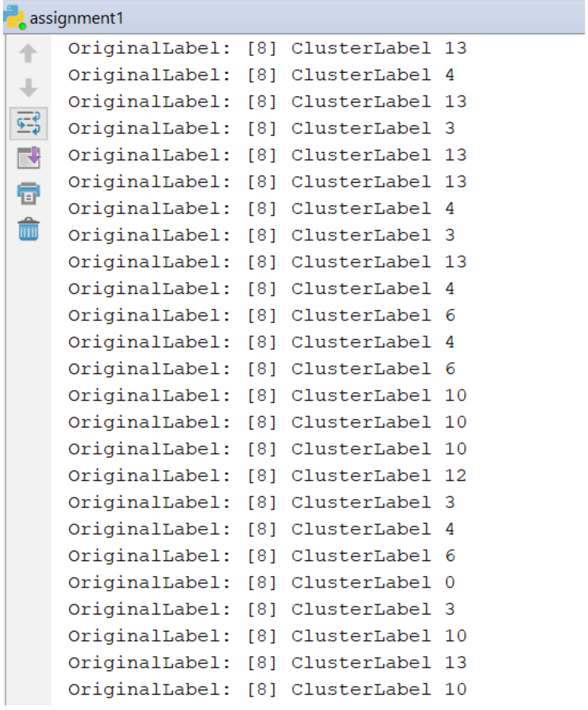
6. Visualization
   Defined color map.
   Using matplotlib, plotted the clusters in 2 dimensions.



Analysis:

1.) Most common words were ebola, cancer, hospital, health.
All these words were related to health
Note: The word "Video" makes it to the list of top 10 words as its been used by every tweet to inform the followers of a link to Video.
The results would have been a bit different if "Video" would have been treated as a stop word.

2.) We analyzed that each twitter account doesn't form its own cluster.
As we compare the results of the labeling of cluster to the original data we can conclude that not all the tweets from one account formed a part of one cluster.



```
OriginalLabel: [8] ClusterLabel 13
OriginalLabel: [8] ClusterLabel 4
OriginalLabel: [8] ClusterLabel 13
OriginalLabel: [8] ClusterLabel 3
OriginalLabel: [8] ClusterLabel 13
OriginalLabel: [8] ClusterLabel 13
OriginalLabel: [8] ClusterLabel 4
OriginalLabel: [8] ClusterLabel 3
OriginalLabel: [8] ClusterLabel 13
OriginalLabel: [8] ClusterLabel 4
OriginalLabel: [8] ClusterLabel 6
OriginalLabel: [8] ClusterLabel 4
OriginalLabel: [8] ClusterLabel 6
OriginalLabel: [8] ClusterLabel 10
OriginalLabel: [8] ClusterLabel 10
OriginalLabel: [8] ClusterLabel 10
OriginalLabel: [8] ClusterLabel 12
OriginalLabel: [8] ClusterLabel 3
OriginalLabel: [8] ClusterLabel 4
OriginalLabel: [8] ClusterLabel 6
OriginalLabel: [8] ClusterLabel 0
OriginalLabel: [8] ClusterLabel 3
OriginalLabel: [8] ClusterLabel 10
OriginalLabel: [8] ClusterLabel 13
OriginalLabel: [8] ClusterLabel 10
```

The clusters were formed due to familiarity in the text in tweets and since all of them had health related tweets, Its possible that most of content of the Tweets were similar.

- **Problems encountered:**

1. There are still some digitals and single letters in the Top10 common words list, as well as some special words used in social media, we need the further data processing and cleansing.
2. Some file cannot be opened in Pycharm due to the type of decoding, in the following days we are going to cope with it.