

Twitter Data Analysis

Jyoti Chaudhary

INTRODUCTION

Topic 1:

Mental Health



Mental Health condition is stigmatized, no one talks about it and no one wants to be associated with it. Hence it was very difficult in the past to get the data about mental health conversations. But thanks to the social media platforms such as Twitter that people now come forward and freely talk about their conditions and their challenges while keeping their identity anonymous at the same time.

I chose this topic because I envisioned this project as my first step towards the future goal of analyzing mental health conversational data scaled with Natural Language Processing techniques. (I won't delve into the details since its not in scope for this project).

Twitter is a rich source of data as it provides real time streaming on sentiments from all parts of the world. However, careful analysis and a lot of data filtering and preprocessing is required to achieve accurate results.

Topic 2:

Pollution and Climate Change



With this project, I wanted to bring to notice the menace of pollution and climate change today. I believe that this is the most serious threat the mankind is going to face by the end of this century. There are various kinds of pollution - water pollution, air pollution and a new kind of

pollution called plastic pollution. Due to increasing levels of Carbon dioxide in the atmosphere and lowering levels of oxygen in oceans, many marine species are endangered due to a phenomenon called ocean acidification. Personally, while working on this project, I came across disturbing facts that I did not know about earlier; one such fact is that it takes atleast 450 years

for a plastic bottle to completely biodegrade. And because of human activities, each day large amounts of plastic waste is being dumped in ocean.

To start work on project, the first step was to code the setup to stream the twitter data on above two topics.

TABLE OF CONTENTS

- 1. TWITTER DATA SCRAPING**
 - 1.1 KEYWORDS USED**
 - 1.2 LENGTH OF COLLECTION PERIOD**
 - 1.3 DATA VOLUME**
- 2. CANDIDATES FOR ANALYSIS**
 - 2.1 TOP HASHTAGS**
 - 2.2 TOP SIGNIFICANT WORDS**
 - 2.3 TOP URLs**
 - 2.4 TOP RETWEETS**
- 3. WORDCLOUD**
 - 3.1 WORDCLOUD FOR TOP HASHTAGS**
 - 3.2 WORDCLOUD FOR SIGNIFICANT WORDS**
- 4. CONCLUSIONS / FINDINGS**
- 5. CHALLENGES FACED**
- 6. FUTURE WORK**
- 7. ACKNOWLEDGEMENT**

1. TWITTER DATA SCRAPING

1.1. Keywords Used

I used the below listed keywords for scraping the data:

For “Mental Health” ->

*'mentalhealth', 'MentalHealthAwareness', 'SuicidePrevention',
'ptsd', 'alzheimers', 'bipolar', 'Endthestigmaofmentalhealth', 'mentalhealthadvocate', 'psychothera
py', 'psychotherapist', 'NIMHgov', 'mhsm', 'mentalhealth', 'psychcentral', 'SOSChat', 'mhchat',
'PTSDRecovery', 'stressdisorder', 'bipolardepression', 'APAPsychiatric', 'mentalillness',
'dementia', 'schizophrenia'*

For “Pollution and Climate Change” ->

*'pollution', 'airpollution', 'noisepollution', 'stopplastic', 'plastticpollution', 'oceanpollution',
'airquality', 'climatereality', 'DelhiAirKills', 'delhiarquality',
'beatpollution', 'overfishing', 'plasticwaste', 'delhismog', 'delhipollution', 'climatechange', 'cleanseas
, 'globalwarming', 'actonclimate', 'TheMostSeriousSmog'*

1.2. Length of Collection Period

Dec 10th, 10:00 am to 07:00 pm (EST time)

1.3. Data Volume

Mental Health

Count of Tweets collected = 26,710

Data File Size = approx 300MB

Pollution and Climate Change

Count of Tweets collected = 46,469

Data File Size = approx 150MB

2. CANDIDATES FOR ANALYSIS

2.1. TOP HASHTAGS

Batch command for executing mapper and reducer code. The output is sorted with the most frequent Hashtags on the top.

```
C:\Users\Tanvi\Desktop\project2\Hashtag>python mapper.py | sort | reducer.py >>
tophashtags_P.txt
```

Result:

Pollution and Climate Change		Mental Health	
climatechange	7482	mentalhealth	3376
globalwarming	1494	dementia	788
blueplanet2	1270	ptsd	555
belugasolar	1158	depression	462
environment	1010	alzheimers	401
climateaction	707	mentalillness	321
nature	698	anxiety	234
thewalkingdead	683	mhsm	226
pollation	666	bipolar	217
mondaymotivation	640	mentalhealthawareness	202

Fig 2.1.1 Mapper_top_hashtags.py

```
for line in tweets_file:
    try:
        tweet = json.loads(line)
        num_tags = len(tweet['entities']['hashtags'])
        if num_tags > 0:
            for i in range(num_tags):
                tag = tweet['entities']['hashtags'][i]['text']
                if tag:
                    print ("{}\t{}".format(tag.lower(), 1))
        else:
            data = tweet['text'].strip().split(" ")
            for i in range(len(data)):
                if len(data[i]) > 0:
                    if data[i][0]=="#":
                        print ("{}\t{}".format(data[i].lower(), 1))
    except:
        continue
```

In the above Fig 2.1.1, the mapper code for top hashtags directly reads the hashtags from entities object – “`tweet['entities']['hashtags'][i]['text']`”. If no hashtags found under entities object, then program reads the tweet’s text “`tweet['text']`” for hashtags. This approach is more efficient as it requires less processing since most times hashtags are read from entities objects and the program does not read the long tweet text. This approach fetches most relevant data.

In below Fig 2.1.2, the reducer code for top hashtags used dictionary structure to sort the reducer output. Dictionary length limited to 10 to achieve space efficiency.

Fig 2.1.2 Reducer_top_hashtags.py

```
import sys
import operator

Count_of_hashtag = 0
previousTag = None
tweets_data = dict()

for line in sys.stdin:
    data_mapped = line.strip().split("\t")
    if len(data_mapped) != 2:
        continue

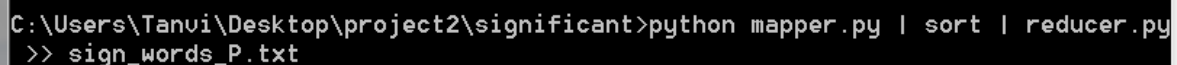
    thisTag, thisCount = data_mapped

    if thisTag in tweets_data.keys():
        tweets_data[thisTag] += int(thisCount)
    else:
        if len(tweets_data) > 10:
            min_key = min(tweets_data, key=tweets_data.get)
            tweets_data.pop(min_key)
            tweets_data[thisTag] = int(thisCount)
        else:
            tweets_data[thisTag] = int(thisCount)

sorted_tweet = sorted(tweets_data.items(), key=operator.itemgetter(1), reverse = True)[0:10]

for data in sorted_tweet:
    print ("{}{}\t{}".format(data[0], data[1]))
```

2.2. TOP SIGNIFICANT WORDS:



```
C:\Users\Tanvi\Desktop\project2\significant>python mapper.py | sort | reducer.py
>> sign_words_P.txt
```

Results:

Pollution and Climate Change		Mental Health	
pollution	8380	bipolar	4271
enforcement	4295	dementia	1754
plastic	3446	please	1712
trump	3344	people	1013
climate	2540	explanation	917
slowdown	1623	mental	902
people	1605	health	808
coincides	1526	don't	600
confidential	1526	family	557
obtained	1526	suffers	524

Fig 2.2.1 mapper_top_words.py

```
#!/usr/bin/python

import sys
import json
import nltk
nltk.download("stopwords")
from nltk.corpus import stopwords

#set the path to the input file
#tweets_data_path = 'C:/Users/Tanvi/Desktop/project2/stream/tweets_MH.txt'
tweets_data_path = 'C:/Users/Tanvi/Desktop/project2/stream/pollution/tweets_P.txt'

#open the input file for reading
tweets_file = open(tweets_data_path, "r")

#process each line in input file
for line in tweets_file:
    try:
        tweet = json.loads(line)
        word_list = []
        data = tweet['text'].strip().split(" ")
        for i in range(len(data)):
            if len(data[i]) > 4:
                if ((data[i][0] == "#") or (data[i][0] == "@") or (data[i][:4]=="http")):
                    continue
                else:
                    word_list.append(data[i].lower())
        filtered_words = [word for word in word_list if word not in stopwords.words('english')]

        if len(filtered_words) > 0:
            for i in range(len(filtered_words)):
                print ("{}{}\t{}\n".format(filtered_words[i], 1))

    except:
        continue
```

In above Fig 2.2.1, the mapper code for Top Significant Words:

- Data extracted from tweet['text']
- Word length greater than 4
- Filtered words beginning with “#”, “@” and “http”
- Filtered stopwords using NLTK library’s STOPWORDS corpus

2.3. Top URLs

```
C:\Users\tanvi\Desktop\project2\topURLs>python mapper.py | sort | reducer.py >> topurls_P.txt
```

Results:

Pollution and Climate Change	
http://cleantechnica.com/2017/12/10/mappair-new-uk-wide-online-high-resolution-air-pollution-map/	1143
http://www.bbc.co.uk/news/science-environment-42264788	598
https://you.38degrees.org.uk/petitions/bring-back-bottle-deposits-to-stop-plastic-pollution-in-our-oceans-1?bucket=email-blast-11_12_2017_kck_meeting_bt&source=twitter-share-button	483
http://aliases.site/tiredearth	442
http://aliases.site/tiredearthfacebook	394
http://aliases.site/twittertiredearth	373
https://www.nytimes.com/2017/12/10/us/politics/pollution-epa-regulations.html	334
http://www.bbc.com/news/science-environment-42264788	311
http://aliases.site/7	259
https://twitter.com/omarvaidd/status/939895892198469632	217

Mental Health	
https://twitter.com/aaronkudi/status/922963875997343744	380
https://twitter.com/real_assange_/status/940321917558333440	257
https://twitter.com/sesthebest/status/940371357245751296	80
https://twitter.com/pyrexpicasso/status/940336632904978432	68
http://atomicdinosaurenemy.com/2017/12/05/this-alzheimers-reversing-oil-is-beating-prescription-drugs/	60
https://twitter.com/igorrf_bsb/status/940051504924319745	53
https://twitter.com/calgaryherald/status/940229556165922816	51
http://nomorebatterywaste.com/2017/12/05/this-alzheimers-reversing-oil-is-beating-prescription-drugs/	50
https://gleam.io/59wsp/huge-makeup-giveaway	50
https://twitter.com/fact/status/939785761553043456	45

In the below Fig 2.3.1, for the mapper code for TopURLs:

Data has been extracted from `tweet['entities']['urls']['expanded_url']`. If “expanded_url” not found for a tweet, then its extracted from `“tweet ['entities']['urls']['url']”`

Fig 2.3.1 Mapper_top_Urls.py

```
for line in tweets_file:
    try:
        tweet = json.loads(line)
        num_urls = len(tweet['entities']['urls'])
        #print("num_urls: ", num_urls)
        if num_urls > 0:
            for i in range(num_urls):
                url = tweet['entities']['urls'][i]["expanded_url"]
                if url:
                    print ("{}\t{}".format(url.lower(), 1))
                else:
                    url = tweet['entities']['urls'][i]["url"]
                    if url:
                        print ("{}\t{}".format(url.lower(), 1))

    except:
        continue
```

- For “Pollution and Climate Change”, the top URLs were about:
 - UK's high-resolution air pollution map
 - Seven charts that explain the plastic pollution problem
 - Bring back bottle deposits to stop plastic pollution in our oceans
 - Under Trump, E.P.A. Has Slowed Actions Against Polluters, and Put Limits on Enforcement Officers
 - Dec 11th World Mountain Day 2017
 - Black lungs and poisoned water: Trump is dismantling the EPA. Key enforcement agents no longer have the authority to order air and water pollution tests. WH has removed the guardrails.
- For “Mental Health”, the top URLs were about:
 - Tweet from an account in the name of Julian Assange about Hillary Clinton’s medical document.
 - A book – “The High Cost of Flowers” by Cynthia Kraack. This book revolves around a troubled family dealing with Alzheimer’s disease
 - This “Alzheimer’s-Reversing Oil” is Beating Prescription Drugs!!!

- No Evidence of Canola Oil Causing Alzheimer's and Dementia
- Dental Problems Seen in Bipolar Disorder Patients
- Police officer with PTSD from Pulse massacre loses his job

2.4. Top ReTweets:

```
C:\Users\tanvi\Desktop\project2\topRetweets>python mapper.py | sort | reducer.py
>> reTweets_MH.txt
```

Results:

Pollution and Climate Change	
1384	There's widespread evidence that the Trump admin has been rolling back federal regulations. How about enforcement t... https://t.co/fi4zk3uvvZ
1128	MappAir — New UK-Wide, Online, High-Resolution Air Pollution Map CleanTechnica #BelugaSolar https://t.co/RcIXswJ1M1
965	When scientists say bears are going extinct, I want people to realize what it looks like, says photographer Paul... https://t.co/QPjM3e6q9a
787	Black lungs and poisoned water: Trump is dismantling the EPA. Key enforcement agents no longer have the authority t... https://t.co/GvE2DkS5dk
342	Climate change is having a negative impact on labor productivity, the spread of infectious diseases and exposure to... https://t.co/s7C7zO9KBh
311	Before Trump's announcement, "Uranium firm lobbied Trump to scale back Bears Ears, which has highest concentrations... https://t.co/n4Y0uDcj9c
288	Under Scott Pruitt, the EPA has abandoned its basic obligation to protect our clean air and water. Special interest... https://t.co/FOkQMvr0wv
212	Air pollution kills 10x as many in the UK as drug overdose, hospital admissions for respiratory conditions risen 3x... https://t.co/qRdBRajlop
179	This is a planetary crisis... we are ruining the ecosystem of the ocean.
176	My mum died of lung cancer this year. She was 62 and she'd never smoked. But she'd lived in London her whole life.... https://t.co/Ox97HSSUCZ

Mental Health	
909	@Twitter just deleted this tweet with no explanation as to why. I'm posting it again. Please retweet!
226	SLC TWITTER: please be on the lookout for one of my grandmas Maata Satini. She suffers from severe dementia and was... https://t.co/ill3w0Ql12
191	"If others are in the same situation, & it helps them to hear about our experiences of dementia, then that's what w... https://t.co/H8ut3U2l94
63	vegas is weird man... buncha heroin addicts, cocaine snorting vegans, high school drop outs, supreme clothing dickr... https://t.co/oc831y8akQ
60	i think im bipolar
49	Do you experience the 'winter blues', SAD, or depression, especially during the winter months? #ad Check out these... https://t.co/YlBtwliEvN
48	The only shame in #MentalHealth is on those who judge others.
45	Thanks. I thought it was early onset dementia. https://t.co/Y2nEs0SfWI
43	idk if im bipolar or i just always notice some bullshit when im happy
43	Damn https://t.co/4GlskakGPc

In the below Fig 2.4.1, the mapper code for Top ReTweets, the data has been extracted from *Tweet['retweeted_status']['id_str']* and *Tweet['retweeted_status']['text']*

Fig 2.4.1 Mapper_top_ReTweets.py

```
for line in tweets_file:
    try:
        tweet = json.loads(line)
        retweet_id = tweet['retweeted_status']['id_str']
        if retweet_id:
            tweet_text = tweet['retweeted_status']['text']
            print ("{0}\t{1}\t{2}".format(retweet_id, 1, tweet_text))
    except:
        continue
```

- For “Pollution and Climate Change”, the top retweets were about:
 - Trump administration’s lack of enforcing environmental regulations on polluters
 - Trump dismantling EPA
 - Trump scaling back Bears Ears National Monument under the influence of lobbying Uranium Companies
 - Polar bears going extinct
 - UK air pollution
 - BBC series blueplanet2 showing shocking images of ocean plastic pollution
 - California Fire
 - World's first Solar-Powered Train ready to be launched in Australia
- For “Mental Health”, the top retweets were about:
 - A tweet from an account in the name of Julian Assange
 - Tweets about Dementia, bipolar and depression

3. WordCloud:

Installed “WORDCLOUD” package using “pip install” command. Fig 3.1 (below) is the screenshot of successful installation:

Fig 3.1 WordCloud installation

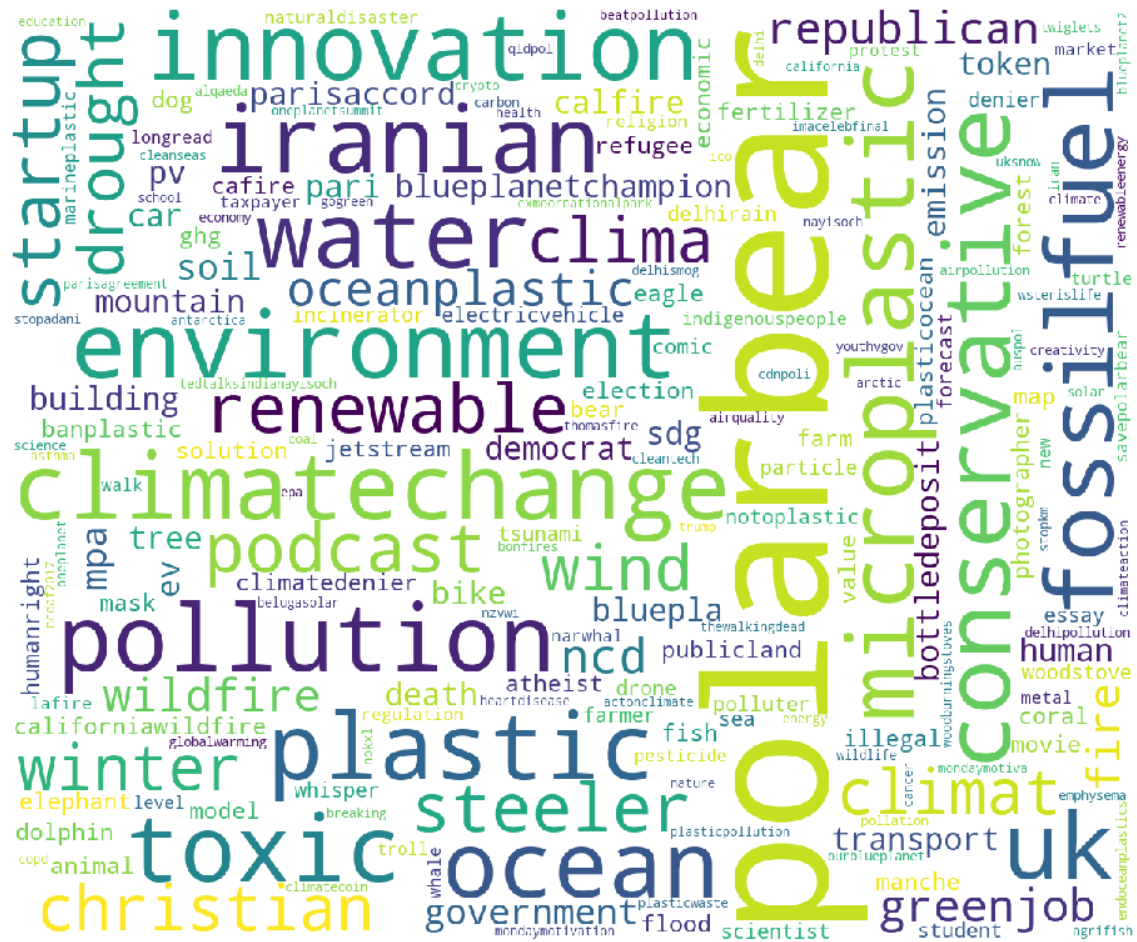
```
C:\Users\tanvi\Desktop>Python -m pip install wordcloud-1.3.2-cp36-cp36m-win32.whl
Processing c:\users\tanvi\desktop\wordcloud-1.3.2-cp36-cp36m-win32.whl
Requirement already satisfied: matplotlib in c:\users\tanvi\appdata\local\programs\python\python36-32\lib\site-packages (from wordcloud==1.3.2)
Requirement already satisfied: pillow in c:\users\tanvi\appdata\local\programs\python\python36-32\lib\site-packages (from wordcloud==1.3.2)
Requirement already satisfied: numpy>=1.6.1 in c:\users\tanvi\appdata\local\programs\python\python36-32\lib\site-packages (from wordcloud==1.3.2)
Requirement already satisfied: pytz in c:\users\tanvi\appdata\local\programs\python\python36-32\lib\site-packages (from matplotlib->wordcloud==1.3.2)
Requirement already satisfied: cycler>=0.10 in c:\users\tanvi\appdata\local\programs\python\python36-32\lib\site-packages (from matplotlib->wordcloud==1.3.2)
Requirement already satisfied: six>=1.10 in c:\users\tanvi\appdata\local\programs\python\python36-32\lib\site-packages (from matplotlib->wordcloud==1.3.2)
Requirement already satisfied: python-dateutil>=2.0 in c:\users\tanvi\appdata\local\programs\python\python36-32\lib\site-packages (from matplotlib->wordcloud==1.3.2)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in c:\users\tanvi\appdata\local\programs\python\python36-32\lib\site-packages (from matplotlib->wordcloud==1.3.2)
Requirement already satisfied: olefile in c:\users\tanvi\appdata\local\programs\python\python36-32\lib\site-packages (from pillow->wordcloud==1.3.2)
Installing collected packages: wordcloud
Successfully installed wordcloud-1.3.2

C:\Users\tanvi\Desktop>cd project2
```

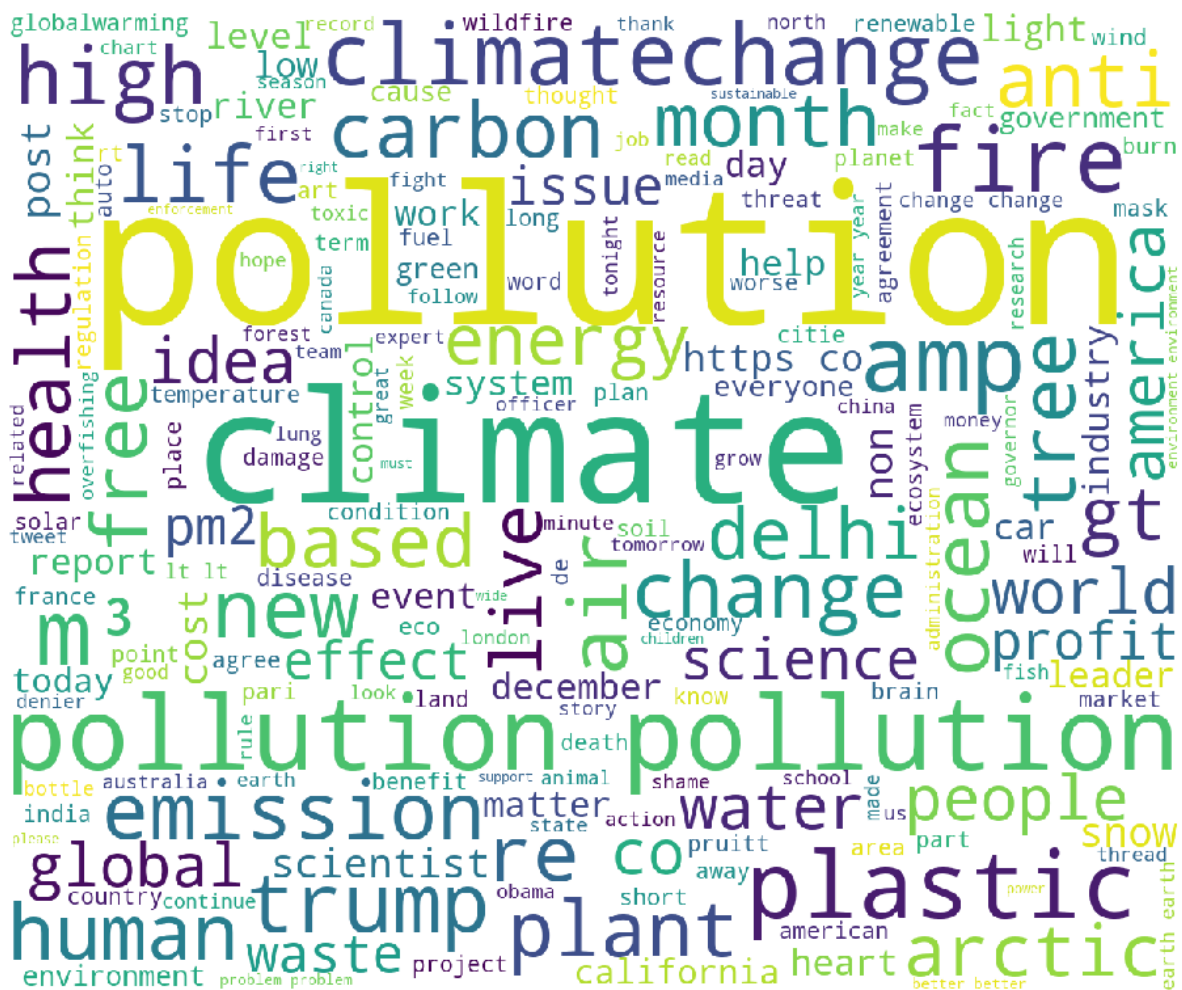
Python command for generating the WordCloud:

```
C:\Users\tanvi\Desktop\project2\wordcloud>python wordcloud.py
```

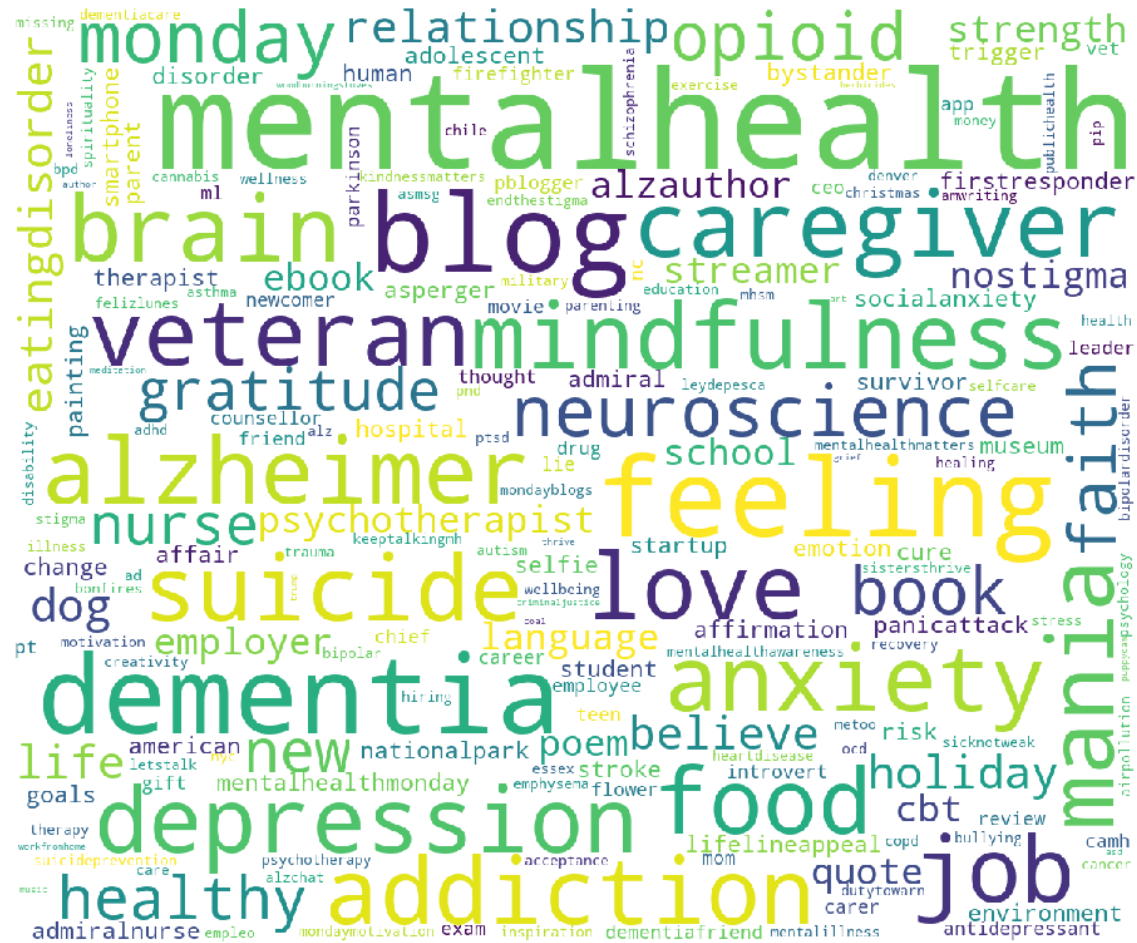
3.1. WordCloud for Top Hashtags (Pollution and Climate Change):



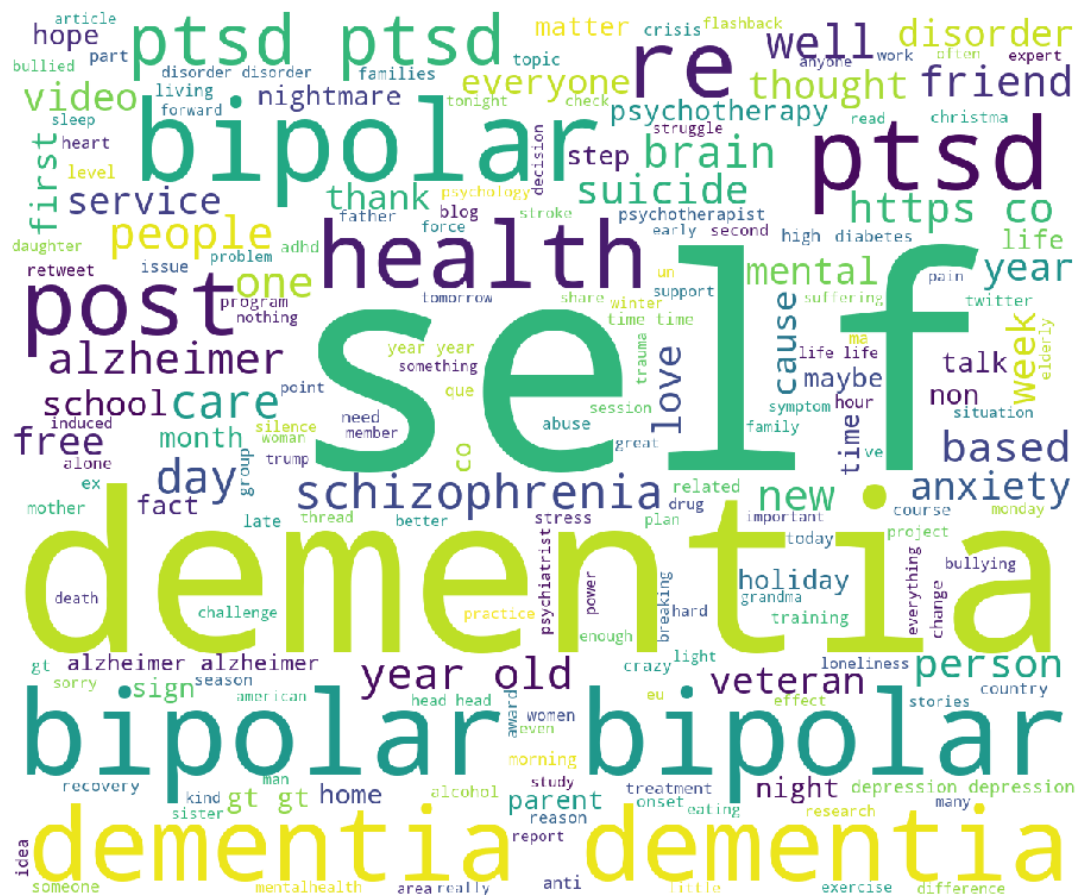
3.2. WordCloud for Significant Words (Pollution and Climate Change):



WordCloud for Top Hashtags (Mental Health):



WordCloud for Significant Words (Mental Health):



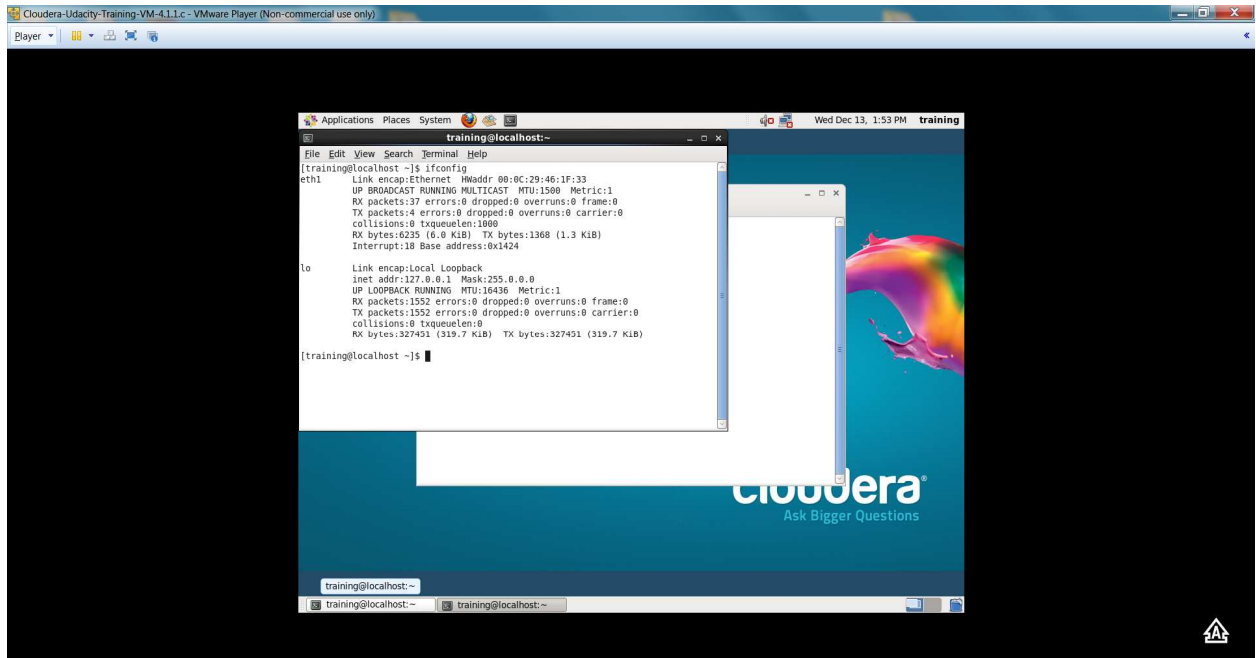
4. CONCLUSIONS / FINDINGS:

- For “Mental Health” topic, what was trending in my data was:
 - a Dec 10th tweet from an account in the name of Julian Assange where the account owner puts out a 2014 dated medical document claiming the document belongs to Hillary Clinton diagnosing her of Subcortical Vascular Dementia.
 - A book was mentioned in 2 different top URLs – “The High Cost of Flowers” by Cynthia Kraack. This book is about a family dealing with Alzheimer’s disease to one of its members.
 - News items about PTSD

- For “Pollution and Climate Change”, the trending topics were -
 - Controversy over Trump dismantling EPA. EPA is short for United States Environmental Protection Agency. Top reTweeted news was:
 "Under Trump, E.P.A. Has Slowed Actions Against Polluters, and Put Limits on Enforcement Officers"
 - Trump scaling back “Bears Ears National Monument” under the influence of lobbying Uranium Companies
 - UK’s air pollution
 - Polar Bears going extinct
- The results I got after running my entire setup were correct. The trending retweets and top URLs in my data were about my two chosen topics.
- This is the first time I have worked with twitter data. After working on this project, I am more curious and interested in exploring and mining the mental health data for sentiment analysis.
- The keywords I gave in my streaming job were specific to the topics. But still there was some data in my files which was not related to my topics but otherwise was trending on Twitter on Dec 10th when I ran my scraping job. This is indicated by the top hashtag such as ‘nzwvi’ for (New Zealand vs West Indies match). Other hashtag, thewalkingdead, may be due to the trending news about “The Walking Dead” series stumbled into its worst mid-season finale ratings since 2011 . I think if I had collected the data over multiple days and if the data size was larger, then my files would have more prominent data about the keywords that I had scraped.
- I did not add the collection of authoritative users because the code was similar to collecting top ReTweets. The users with the higher ReTweets can be categorized as authoritative users. Or the users with the maximum number of followers could have been retrieved for this metric. I believe I need to collect the data over multiple days to accurately determine the authoritative users.

5. CHALLENGES FACED WHILE WORKING ON THIS PROJECT

- Could not run my code on Hadoop. **IP address of VM was not available**. Please see the below screenshot from VM.



- I added too many keywords during my initial scraping efforts that led to irrelevant data in my files and hence irrelevant results. I ran 3-4 iterations of scraping the data and executing the code to identify the right set of keywords to be added to my streaming job.
- Faced “Read TimeOut Error” during initial twitter scraping. But when I ran the job during night, the job ran atleast for 2 hours continuously before connection broke again. I had to restart my job multiple times to scrape the data.
- Spent a lot of time trying to figure out how to sort the reducer output.
- Problems faced while installing WORDCLOUD package and invested around 2 hours to fix the problem.

6. Future Work

Techniques such as Text Mining, Sentiment Analysis and NLP can be applied on this data. I intend to scrape more data and continue working on Mental Health topic for my future work. I will try to source more mental health conversational data from other sources such as Crisistextline.org and will apply Deep Neural Network techniques which I have learned in my “Deep Learning” course this semester. My long term (ambitious) goal is to implement a text based psychotherapy application scaled with Natural Language Processing.

7. Acknowledgement

I want to thank **Prof Ajita John** for her incredible support throughout this project. It was a great learning experience being part of this class and working on this course project. I gained skills on SAS Enterprise Miner, Hadoop MapReduce framework and about Twitter analysis. Thank You!!