Step 1:

Importing data from json and put it into a dictionary:

import json  
import csv  
  
  
  
f = open('C:/Users/gg/Desktop/SPRING2021/IST664/hw1/01.json',encoding="utf-8")  
p = f.readlines()  
data1 = [json.loads(i) for i in p]  
f = open('C:/Users/gg/Desktop/SPRING2021/IST664/hw1/02.json',encoding="utf-8")  
p = f.readlines()  
data2 = [json.loads(i) for i in p]  
  
data3=data1+data2  
  
  
dict1={}  
dict1['author']=[d['author'] for d in data3]  
dict1['facebook']=[d['thread']['social']['facebook'] for d in data3]  
dict1['title']=[d['title'] for d in data3]  
dict1['published']=[d['published'] for d in data3]  
dict1['url']=[d['url'] for d in data3]  
dict1['replies\_count']=[d['thread']['replies\_count'] for d in data3]  
dict1['country']=[d['thread']['country'] for d in data3]  
dict1['facebook']=[str(d)for d in dict1['facebook']]  
  
[dict2={}  
dict2['text']= d['text'] for d in data3]]

Step 2:

Importing nltk’s tag system

from nltk.tokenize import sent\_tokenize  
from nltk.tokenize import word\_tokenize  
from nltk.corpus import brown  
from nltk.tag import pos\_tag, map\_tag

Step3:

Write functions to determine whether a sentence contain Adj or exclamation marks:

def adjor(w):  
 j=word\_tokenize(w)  
 k=[(word, map\_tag('en-ptb', 'universal', tag)) for word, tag in pos\_tag(j)]  
 for (a,b) in k:  
 if a == '!' or b == 'ADJ':  
 return True  
 else:  
 return False

I choose to use this method because using map.tag can quickly show us whether a word is adj. and if we can identify a word containing adj. The sentence certainly contains adjP. Also, word\_tokenize() methods return “!” individually so a determination of adjP and “!” can be write in one lane to reduce the complicity of the code.

Step3:

Write functions to sperate text into sentence and then use the method wrote in last step to process the data.

sentanalyze=[]

def extractsent(a):  
 for sent in sent\_tokenize(a):  
 if adjor(sent):  
 sentanalyze.append(sent)

This method is using sent\_tokenize() to quickly divide text into sentences for analyze. Then, if a sentence contain AdjP or ‘!’, add it into the list of sentence we are analyzing.

Step4:

Analyze the data:

Calculate the average length of sentence. Prediction is that these sentence would be short based on the strong emotion they are trying to convey.

calc=0  
for d in sentanalyze:  
 calc+=len(d)  
calc/len(sentanalyze)

the average length of these sentences are: 150.01638477249557. In contrast to the prediction I made, these sentence are relatively long.

50 most common Adj.:

Because of that I can use tagged information when trying to get the 50 most common Adj., no cleaning method other than word\_tokenize() and map.tag() would be used.

def adjorize(w):  
 j=word\_tokenize(w)  
 k=[(word, map\_tag('en-ptb', 'universal', tag)) for word, tag in pos\_tag(j)]  
 return k  
  
wordanalyze=[]  
for d in sentanalyze:  
 wordanalyze.append(adjorize(d))

flatword = []  
for b in wordanalyze:  
 for c in b:  
 flatword.append(c)

After processed map.tag() for each sentence, I had flatten the list of list wordanalyze into one list called flatword.

word\_tag\_fd = nltk.FreqDist(flatword)

mostfreqAdj=[wt[0] for (wt, \_) in word\_tag\_fd.most\_common() if wt[1] == 'ADJ']

showing the 50 most common Adj.:

mostfreqAdj[:50]

['Chinese',

'“',

'new',

'more',

'other',

'More',

'last',

'Many',

'first',

'global',

'Most',

'many',

'such',

'due',

's',

'Last',

'medical',

'Other',

'economic',

'public',

'positive',

'—',

'novel',

'[',

'’',

'major',

'\*',

'Related',

'international',

'local',

'Japanese',

'central',

'same',

'Asian',

'least',

'good',

'higher',

'Several',

'British',

'much',

'financial',

'latest',

'Copyright',

'likely',

'social',

'next',

'confirmed',

'recent',

'epidemic',

'second']

As from the result of 50 most common adj. you can see that, first, the covid-19 virus was discovered in China. Also, due to the major position of Chinese mask industry in global supply chain. The Adj ‘Chinese’ was used a lot. There are words like ‘new’, ‘first’ getting used a lot which is related to the covid-19’s never seem before impact of the world. Words like ’Global’ or ‘local’ is indicating that people view the pandemic globally and care about the status locally(for example, I kept track for the cases in the town I am living at daily. But I read a lot of news about pandemic globally.)

50 most common verb:

Using the same technique of 50 most common adj.

mostfreqVerb=[wt[0] for (wt, \_) in word\_tag\_fd.most\_common() if wt[1] == 'VERB']  
mostfreqVerb[:50]

['is',

'have',

'are',

'said',

'be',

'has',

'was',

'been',

'will',

'were',

's',

'had',

'can',

'’',

'including',

'would',

'could',

'reported',

'confirmed',

'may',

'being',

'do',

'according',

'infected',

'should',

'”',

'told',

"'s",

'tested',

'says',

'take',

'get',

'made',

'died',

'make',

'say',

'announced',

'expected',

'see',

'going',

'help',

'go',

'did',

'come',

'work',

'spread',

'know',

'—',

'“',

'does']

After analyzing the 50 most common verb, I got to say, this doesn’t contain much of information like Adj. But if we can perform an analyze between emotion and passive tense, we might found that the major usage of passive verb indicating negative emotion when people tweet about pandemic.

50 most common noun

Same technique like Adj. and Verb

mostfreqNoun=[wt[0] for (wt, \_) in word\_tag\_fd.most\_common() if wt[1] == 'NOUN']  
mostfreqNoun[:50]

['China',

'coronavirus',

'’',

'people',

'virus',

'”',

'%',

'cases',

'outbreak',

'Wuhan',

'health',

'year',

'week',

's',

'“',

'U.S.',

'February',

'Health',

'countries',

'government',

'authorities',

'country',

'ship',

'market',

'Coronavirus',

'Hong',

'world',

'time',

'officials',

'Kong',

'number',

'growth',

'disease',

'New',

'Hubei',

'—',

']',

'passengers',

'companies',

'patients',

'deaths',

'days',

'spread',

'economy',

'travel',

'company',

'US',

'impact',

'news',

'Japan']

The majority noun usage of ‘China’ and ‘Wuhan’ indicate that this dataset is collected around 2020 April. Because I didn’t see words like ‘ventilator’ or ‘vaccine’ as popular but ‘ventilator’ and ‘vaccine’ is probably the top 50 common nouns when people discuss about pandemic 2020 summer to 2021 spring. People is thinking about what the government can do and should do in this specific time as we can clear see that government and countries are the top 10 most common nouns.