Assignment 7:
Natural language Processing:

a)Explore each text file using Python and the NLTK library.
Generate word frequency counts and plots (eliminating stopwords) for each article.
#importing all the necessary libraries

```
In [1]: import os
        import nltk
        import nltk.corpus

In [2]: nltk.download()
        showing info https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/index.xml
Out[2]: True

In [32]: e1 = 'syncorona.txt'

In [31]: e1 = open('syncorona.txt','rt')

In [30]: allwords = e1.read()

In [6]: print(allwords)
        ion to get worse. Most importantly, don't make this decision on your own. It's always best not to adjust the dose or
        stop taking a prescription medication without first talking to the doctor who prescribed the medication.
```

#printing all the words in the text document

```
In [6]: print(allwords)
        ion to get worse. Most importantly, don't make this decision on your own. It's always best not to adjust the dose or
        stop taking a prescription medication without first talking to the doctor who prescribed the medication.
        Will a pneumococcal vaccine help protect me against coronavirus?
        Vaccines against pneumonia, such as pneumococcal vaccine and Haemophilus influenza type B (Hib) vaccine, only help pr
        otect people from these specific bacterial infections. They do not protect against any coronavirus pneumonia, includi
        ng pneumonia that may be part of COVID-19. However, even though these vaccines do not specifically protect against th
        e coronavirus that causes COVID-19, they are highly recommended to protect against other respiratory illnesses.
        I'm older and have a chronic medical condition, which puts me at higher risk for getting seriously ill, or even dying
        from COVID-19. What can I do to reduce my risk of exposure to the virus?
        Anyone 60 years or older is considered to be at higher risk for getting very sick from COVID-19. This is true whether
        or not you also have an underlying medical condition, although the sickest individuals and most of the deaths have be
        en among people who were both older and had chronic medical conditions, such as heart disease, lung problems or diabe
        tes.
        The CDC suggests the following measures for those who are at higher risk:
        * Obtain several weeks of medications and supplies in case you need to stay home for prolonged periods of time.
        * Take everyday precautions to keep space between yourself and others.
        * When you go out in public, keep away from others who are sick, limit close contact, and wash your hands often.
        * Avoid crowds.
        * Avoid cruise travel and nonessential air travel.
```

In [7]:
```
from nltk.tokenize import RegexpTokenizer
tokenizer = RegexpTokenizer(r'\w+')
tokens = tokenizer.tokenize(allwords.lower())
print(tokens)
```

```
['what', 'are', 'the', 'symptoms', 'of', 'covid', '19', 'some', 'people', 'infected', 'with', 'the', 'virus', 'have',
'no', 'symptoms', 'when', 'the', 'virus', 'does', 'cause', 'symptoms', 'common', 'ones', 'include', 'low', 'grade',
'fever', 'body', 'aches', 'coughing', 'nasal', 'congestion', 'runny', 'nose', 'and', 'sore', 'throat', 'however', 'co
vid', '19', 'can', 'occasionally', 'cause', 'more', 'severe', 'symptoms', 'like', 'high', 'fever', 'severe', 'cough',
'and', 'shortness', 'of', 'breath', 'which', 'often', 'indicates', 'pneumonia', 'how', 'long', 'is', 'it', 'between',
'when', 'a', 'person', 'is', 'exposed', 'to', 'the', 'virus', 'and', 'when', 'they', 'start', 'showing', 'symptoms',
'because', 'this', 'coronavirus', 'has', 'just', 'been', 'discovered', 'the', 'time', 'from', 'exposure', 'to', 'symp
tom', 'onset', 'known', 'as', 'the', 'incubation', 'period', 'for', 'most', 'people', 'has', 'yet', 'to', 'be', 'dete
rmined', 'based', 'on', 'current', 'information', 'symptoms', 'could', 'appear', 'as', 'soon', 'as', 'three', 'days',
'after', 'exposure', 'to', 'as', 'long', 'as', '13', 'days', 'later', 'recently', 'published', 'research', 'found',
'that', 'on', 'average', 'the', 'incubation', 'period', 'is', 'about', 'five', 'days', 'how', 'does', 'coronavirus',
'spread', 'the', 'coronavirus', 'is', 'thought', 'to', 'spread', 'mainly', 'from', 'person', 'to', 'person', 'this',
'can', 'happen', 'between', 'people', 'who', 'are', 'in', 'close', 'contact', 'with', 'one', 'another', 'droplets',
'that', 'are', 'produced', 'when', 'an', 'infected', 'person', 'coughs', 'or', 'sneezes', 'may', 'land', 'in', 'the',
'mouths', 'or', 'noses', 'of', 'people', 'who', 'are', 'nearby', 'or', 'possibly', 'be', 'inhaled', 'into', 'their',
'lungs', 'coronavirus', 'can', 'also', 'spread', 'from', 'contact', 'with', 'infected', 'surfaces', 'or', 'objects',
'for', 'example', 'a', 'person', 'can', 'get', 'covid', '19', 'by', 'touching', 'a', 'surface', 'or', 'object', 'tha
t', 'has', 'the', 'virus', 'on', 'it', 'and', 'then', 'touching', 'their', 'own', 'mouth', 'nose', 'or', 'possibly',
'their', 'eyes', 'how', 'deadly', 'is', 'covid', '19', 'the', 'answer', 'depends', 'on', 'whether', 'you', 're', 'loo
```

#length of tokens

In [8]:
```
len(tokens)
```

Out[8]: 1725

#The frequency distribution

In [9]:
```
from nltk.probability import FreqDist
fdist = FreqDist()
for word in tokens:
    fdist[word]+=1
fdist
```

Out[9]:
```
FreqDist({'the': 79, 'and': 37, 'to': 37, 'a': 31, 'of': 28, 'or': 25, 'is': 23, 'covid': 22, '19': 22, 'are': 20,
...})
```

In [10]:
```
fdist_top10 = fdist.most_common(10)
fdist_top10
```

Out[10]:
```
[('the', 79),
 ('and', 37),
 ('to', 37),
 ('a', 31),
 ('of', 28),
 ('or', 25),
 ('is', 23),
 ('covid', 22),
 ('19', 22),
 ('are', 20)]
```

In [11]:
```
from nltk.tokenize import blankline_tokenize
L_blank = blankline_tokenize(allwords)
len(L_blank)
```

Out[11]: 2

In [12]:
```
from nltk.corpus import stopwords
stopwords.words('english')
len(stopwords.words('english'))
```

Out[12]: 179

In [11]:
```
from nltk.tokenize import blankline_tokenize
L_blank = blankline_tokenize(allwords)
len(L_blank)
```

Out[11]: 2

In [12]:
```
from nltk.corpus import stopwords
stopwords.words('english')
len(stopwords.words('english'))
```

Out[12]: 179

In [31]:
```
stoptokens =[token for token in tokens if token not in stopwords.words('english')]
```

In [32]:
```
print(stoptokens)
```

```
ssion', 'new', 'coronavirus', 'clear', 'spread', 'infected', 'person', 'food', 'handled', 'prepared', 'would', 'likel
y', 'exception', 'rule', 'said', 'new', 'coronavirus', 'respiratory', 'virus', 'known', 'spread', 'upper', 'respirato
ry', 'secretions', 'including', 'airborne', 'droplets', 'coughing', 'sneezing', 'virus', 'causes', 'covid', '19', 'al
so', 'detected', 'stool', 'certain', 'people', 'currently', 'cannot', 'rule', 'possibility', 'infection', 'transmitte
d', 'food', 'infected', 'person', 'thoroughly', 'washed', 'hands', 'case', 'hot', 'food', 'virus', 'would', 'likely',
'killed', 'cooking', 'may', 'case', 'uncooked', 'foods', 'like', 'salads', 'sandwiches', 'prevention', 'protect', 'ot
hers', 'covid', '19', 'following', 'actions', 'help', 'prevent', 'spread', 'covid', '19', 'well', 'coronaviruses', 'i
nfluenza', 'avoid', 'close', 'contact', 'people', 'sick', 'avoid', 'touching', 'eyes', 'nose', 'mouth', 'stay', 'hom
e', 'sick', 'cover', 'cough', 'sneeze', 'tissue', 'throw', 'tissue', 'trash', 'clean', 'disinfect', 'frequently', 'to
uched', 'objects', 'surfaces', 'using', 'regular', 'household', 'cleaning', 'spray', 'wipe', 'wash', 'hands', 'ofte
n', 'soap', 'water', 'need', 'know', 'washing', 'hands', 'effectively', 'wash', 'hands', 'often', 'soap', 'water', 'l
east', '20', 'seconds', 'especially', 'going', 'bathroom', 'eating', 'blowing', 'nose', 'coughing', 'sneezing', 'soa
p', 'water', 'readily', 'available', 'use', 'alcohol', 'based', 'hand', 'sanitizer', 'least', '60', 'alcohol', 'cover
ing', 'surfaces', 'hands', 'rubbing', 'together', 'feel', 'dry', 'always', 'wash', 'hands', 'soap', 'water', 'hands',
'visibly', 'dirty', 'cdc', 'handwashing', 'website', 'detailed', 'instructions', 'video', 'effective', 'handwashing',
'procedures', 'wear', 'face', 'mask', 'follow', 'public', 'health', 'recommendations', 'live', 'february', '26', '202
0', 'face', 'masks', 'recommended', 'general', 'public', 'us', 'health', 'facilities', 'require', 'people', 'wear',
'mask', 'certain', 'circumstances', 'respiratory', 'symptoms', 'like', 'coughing', 'sneezing', 'experts', 'recommen
d', 'wearing', 'mask', 'protect', 'others', 'may', 'help', 'contain', 'droplets', 'containing', 'type', 'virus', 'inc
luding', 'flu', 'protect', 'anyone', 'within', 'three', 'six', 'feet', 'infected', 'person', 'cdc', 'offers', 'inform
```

In [15]:
```
len(stoptokens)
```

Out[15]: 944

In [33]:
```
from nltk.probability import FreqDist
fdist = FreqDist()
for word in stoptokens:
    fdist[word]+=1
fdist
```

Out[33]:
```
FreqDist({'covid': 22, '19': 22, 'coronavirus': 20, 'virus': 16, 'people': 12, 'symptoms': 10, 'person': 10, 'risk':
9, 'infected': 8, 'spread': 8, ...})
```

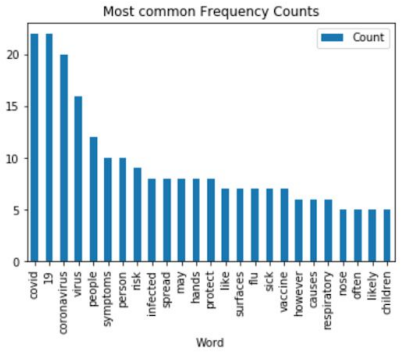
In [34]:
```
first = fdist.most_common(25)
```

In [35]:
```
first
```

Out[35]:
```
[('covid', 22),
 ('19', 22),
 ('coronavirus', 20),
 ('virus', 16),
 ('people', 12),
 ('symptoms', 10),
 ('person', 10),
 ('risk', 9),
 ('infected', 8),
 ('spread', 8),
 ('may', 8),
 ('hands', 8),
 ('protect', 8),
 ('like', 7),
 ('surfaces', 7),
 ('flu', 7),
 ('sick', 7),
 ('vaccine', 7),
 ('however', 6),
```

In [44]:
```
import pandas as pd
df = pd.DataFrame(first,columns = ['Word','Count'])
df.plot.bar(x='Word',y='Count',title = 'Most common Frequency Counts')
```

Out[44]: <matplotlib.axes._subplots.AxesSubplot at 0x1a2ad917d0>



In [ ]:

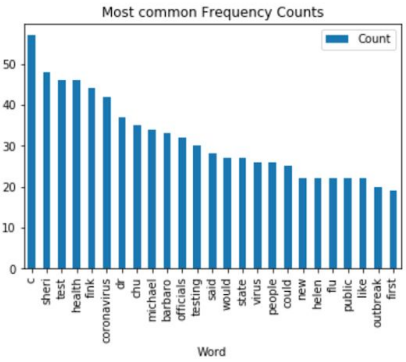
ARTICLE 2:

The most common words used in articles

```
('helen', 22),
('flu', 22),
('public', 22),
('like', 22),
('outbreak', 20),
('first', 19)]
```

In [43]:
```
import pandas as pd
df = pd.DataFrame(second,columns = ['Word','Count'])
df.plot.bar(x='Word',y='Count',title = 'Most common Frequency Counts')
```

Out[43]: <matplotlib.axes._subplots.AxesSubplot at 0x1a1d9819d0>



PART B:
Determine if there is a difference in vocabulary or sentiment
between the two articles, and explain any observed differences.
Include example word counts or sentences to support your observations.

FROM article virus.txt we have observed in depth idea of coronavirus
And in article syncorona.txt we observed the symptoms of coronavirus.
The common words in syncorona.txt

```
In [45]: fdist_top25 = fdist.most_common(25)
         fdist_top25

Out[45]: [('covid', 22),
          ('19', 22),
          ('coronavirus', 20),
          ('virus', 16),
          ('people', 12),
          ('symptoms', 10),
          ('person', 10),
          ('risk', 9),
          ('infected', 8),
          ('spread', 8),
          ('may', 8),
          ('hands', 8),
          ('protect', 8),
          ('like', 7),
          ('surfaces', 7),
          ('flu', 7),
          ('sick', 7),
          ('vaccine', 7),
          ('however', 6),
          ('causes', 6),
          ('respiratory', 6),
          ('nose', 5),
          ('often', 5),
          ('likely', 5),
          ('children', 5)]
```

The common words in virus.txt

```
In [44]: fdist3_top25 = fdist3.most_common(25)
         fdist3_top25

Out[44]: [('c', 57),
          ('sheri', 48),
          ('test', 46),
          ('health', 46),
          ('fink', 44),
          ('coronavirus', 42),
          ('dr', 37),
          ('chu', 35),
          ('michael', 34),
          ('barbaro', 33),
          ('officials', 32),
          ('testing', 30),
          ('said', 28),
          ('would', 27),
          ('state', 27),
          ('virus', 26),
          ('people', 26),
          ('could', 25),
          ('new', 22),
          ('helen', 22),
          ('flu', 22),
          ('public', 22),
          ('like', 22),
          ('outbreak', 20),
          ('first', 19)]
```

The common words in two articles are people,virus,coronavirus,flu.The vocabulary and symptoms used by the two articles are different.The Vocabulary used by the virus.txt describes about the virus.The article syncorona describes about the symptoms of coronavirus
The negative words used by the syncorona.txt are symptoms,risk,infected,spread,flu,,virus.
The Only positive words in the article are vaccine.
The article virus describes the depth of coronavirus.The negative words used by the article are virus,outbreak,coronavirus,flu. The positive words in the article are people,health.

From the above plots ,we can see that there are a greater number of negative sentiment words which makes the articles have more Negative sentiment than Positive Sentiment.

.

References:

Fink, Sheri, and Mike Baker. “'It's Just Everywhere Already': How Delays in Testing Set Back the U.S. Coronavirus Response.” The New York Times. The New York Times, March 11, 2020. https://www.nytimes.com/2020/03/10/us/coronavirus-testing-delays.html.

Harvard Health Publishing. “Coronavirus Resource Center.” Harvard Health. Accessed March 16, 2020. https://www.health.harvard.edu/diseases-and-conditions/coronavirus-resource-center.