Assignment_2_2_Raghuwanshi_Prashant_DSC550

September 11, 2021

0.0.1 Assignment: 2.2 Exercise, Build Your Text Classifiers

0.0.2 Name: Prashant Raghuwanshi

0.0.3 Date: 9/10/2021

0.0.4 Course: DSC550-T301 Data Mining (2221-1)

```
[1]: # import libraries
     import numpy as np
     import pandas as pd
     # import source file to data frame
     ccjsonsrc = pd.read_json('D:\MS_DataScience\DSC550\controversial-comments.
      ↪jsonl', lines = True)
```

A. Convert all text to lowercase letters.

```
[2]: # using lambda function and convert the string to lower case
     ccjsonlower = ccjsonsrc.apply(lambda x: x.astype(str).str.lower())
     # limitting the records in dataframe
     sampledf = ccjsonlower.head(50000)
     sampledf
```

```
[2]:       con                                               txt
     0       0  well it's great that he did something about th…
     1       0                      you are right mr. president.
     2       0  you have given no input apart from saying i am…
     3       0  i get the frustration but the reason they want…
     4       0  i am far from an expert on tpp and i would ten…
     …     ..                                              …
     49995   0  &gt; it's just too bad she sold her soul to fo…
     49996   0                                       /globalists
     49997   0                                         [removed]
     49998   0  i can't disagree that machines will take many …
     49999   0  i disagree. i think if child care were actuall…

     [50000 rows x 2 columns]
```

B. Remove all punctuation from the text.

[3]:
```python
# Create the punctuation dictionary by using unicodedata
import sys
import unicodedata
punctuation = dict.fromkeys(i for i in range(sys.maxunicode)
                      if unicodedata.category(chr(i)).startswith('P'))
```

[4]:
```python
# removing punctuation from each row of dataframe's txt column
for i in range(len(sampledf)) :
    test = [string.translate(punctuation) for string in (sampledf.loc[i,␣
 ↪"txt"])]
    # coverting list to string
    test1 = "".join(str(x) for x in test)
    # updating the row values
    sampledf.loc[i, ["txt"]] = test1
# print dataframe after removing punctuations from txt column
sampledf
```

```
C:\ProgramData\Anaconda3\lib\site-packages\pandas\core\indexing.py:1637:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  self._setitem_single_block(indexer, value, name)
C:\ProgramData\Anaconda3\lib\site-packages\pandas\core\indexing.py:692:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  iloc._setitem_with_indexer(indexer, value, self.name)
```

[4]:
```
       con                                                txt
0        0  well its great that he did something about tho…
1        0                            you are right mr president
2        0  you have given no input apart from saying i am…
3        0  i get the frustration but the reason they want…
4        0  i am far from an expert on tpp and i would ten…
…      ..                                                  …
49995    0  gt its just too bad she sold her soul to fox n…
49996    0                                          globalists
49997    0                                             removed
49998    0  i cant disagree that machines will take many j…
49999    0  i disagree i think if child care were actually…

[50000 rows x 2 columns]
```

C. Remove stop words.

[5]:

```
# load library
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
# nltk.download('stopwords')
# nltk.download('punkt')
#
```

[6]:

```
# load stop words
stop_words = stopwords.words('english')
```

[7]:

```
# remove stop words from each row of dataframe's txt column
for i in range(len(sampledf)) :
    # tokenized each row of dataframe's txt column
    test_token = word_tokenize(sampledf.loc[i, "txt"])
    # remove stop words
    rem_words = [word for word in test_token if word not in stop_words]
    # coverting list to string
    rem_words1 = " ".join(str(x) for x in rem_words)
    # writting back processed removed stop words to dataframe
    # updating the row values for txt column
    sampledf.loc[i, ["txt"]] = rem_words1
# printing last rows of dataframe showing removed stop words
print(rem_words)
```

```
C:\ProgramData\Anaconda3\lib\site-packages\pandas\core\indexing.py:1637:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  self._setitem_single_block(indexer, value, name)
C:\ProgramData\Anaconda3\lib\site-packages\pandas\core\indexing.py:692:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  iloc._setitem_with_indexer(indexer, value, self.name)
```

```
['disagree', 'think', 'child', 'care', 'actually', 'important', 'issue',
'would', 'implemented', 'company', 'giii', 'actually', 'creates', 'distributes',
'fashion', 'line', 'offers', '0', 'paid', 'parental', 'leave', 'respectfully',
'think', 'youre', 'buying', 'optics', 'exact', 'response', 'designed', 'get']
```

[8]:

```
# print dataframe after updating the removed stop words to txt column
sampledf
```

[8]:

```
      con                                                txt
0       0  well great something beliefs office doubt trum…
1       0                                 right mr president
2       0    given input apart saying wrong argument clearly
3       0  get frustration reason want way foundation com…
4       0  far expert tpp would tend agree lot problems u…
…    ..                                                …
49995   0  gt bad sold soul fox news really cant sympathe…
49996   0                                         globalists
49997   0                                            removed
49998   0  cant disagree machines take many jobs embrace …
49999   0  disagree think child care actually important i…

[50000 rows x 2 columns]
```

D. Apply NLTK's PorterStemmer.

[9]:

```python
# load library
from nltk.stem.porter import PorterStemmer
# create stemmer
porter = PorterStemmer()
```

[10]:

```python
# apply stemmer to each row of dataframe's txt column
for i in range(len(sampledf)) :
    # tokenized each row of dataframe's txt column
    test_token1 = word_tokenize(sampledf.loc[i, "txt"])
    # apply stemmer
    porter_words = [porter.stem(word) for word in test_token1]
    # coverting list to string
    porter_words1 = " ".join(str(x) for x in porter_words)
    # writting back processed removed stop words to dataframe
    # updating the row values for txt column
    sampledf.loc[i, ["txt"]] = porter_words1
# printing last rows of dataframe showing removed stop words
print(porter_words)
```

```
C:\ProgramData\Anaconda3\lib\site-packages\pandas\core\indexing.py:1637:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  self._setitem_single_block(indexer, value, name)
C:\ProgramData\Anaconda3\lib\site-packages\pandas\core\indexing.py:692:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

```
  iloc._setitem_with_indexer(indexer, value, self.name)
```

```
['disagre', 'think', 'child', 'care', 'actual', 'import', 'issu', 'would',
'implement', 'compani', 'giii', 'actual', 'creat', 'distribut', 'fashion',
'line', 'offer', '0', 'paid', 'parent', 'leav', 'respect', 'think', 'your',
'buy', 'optic', 'exact', 'respons', 'design', 'get']
```

[11]:

```python
# print dataframe after updating the applied stemmer to each row of dataframe's␣
 ↪txt column
sampledf
```

[11]:

```
       con                                                txt
0        0  well great someth belief offic doubt trump wou…
1        0                                    right mr presid
2        0       given input apart say wrong argument clearli
3        0  get frustrat reason want way foundat complex p…
4        0  far expert tpp would tend agre lot problem und…
…      ..                                                ...
49995    0  gt bad sold soul fox news realli cant sympathe…
49996    0                                          globalist
49997    0                                              remov
49998    0  cant disagre machin take mani job embrac left …
49999    0  disagre think child care actual import issu wo…

[50000 rows x 2 columns]
```

2. Now that the data is pre-processed, you will apply three different techniques to get it into a usable form for model-building. Apply each of the following steps (individually) to the pre-processed data.

A. Convert each text entry into a word-count vector (see sections 5.3 & 6.8 in the Machine Learning with Python Cookbook).

[12]:

```python
# import libraries
from sklearn.feature_extraction.text import CountVectorizer
```

[13]:

```python
count = CountVectorizer()
bag_of_words = count.fit_transform(sampledf['txt'].to_numpy())
bag_of_words.toarray()
```

[13]:

```
array([[0, 0, 0, …, 0, 0, 0],
       [0, 0, 0, …, 0, 0, 0],
       [0, 0, 0, …, 0, 0, 0],
       …,
       [0, 0, 0, …, 0, 0, 0],
       [0, 0, 0, …, 0, 0, 0],
       [0, 0, 0, …, 0, 0, 0]], dtype=int64)
```

```
[14]: # Show features name
      count.get_feature_names()
```

```
[14]: ['00',
       '000',
       '0000001',
       '000005',
       '00001',
       '00003',
       '00005',
       '000079',
       '00009',
       '0003',
       '000327',
       '000844',
       '001',
       '0017',
       '00189',
       '002',
       '00282',
       '003',
       '0033',
       '00360',
       '005',
       '00598',
       '007',
       '00991',
       '01',
       '010',
       '01187',
       '01188',
       '01340',
       '01561',
       '016',
       '01604',
       '01825',
       '01c',
       '02',
       '02000',
       '02293',
       '02419',
       '025',
       '02561',
       '026',
       '0270',
       '03',
       '0304',
       '03069',
       '0311',
       '03167',
       '0351httpswwwthebalancecomusmcmos0351infantryassaultman3345320',
       '036',
       '03890',
       '04',
       '040',
       '04156',
       '04317',
       '04335fe51cbf',
       '04373',
       '04387',
       '043b605557dd',
       '04865',
       '0495b21268d',
       '04987',
       '05',
       '050',
       '05108',
       '0557',
       '05639',
       '057',
       '05871',
       '05953',
       '05961',
       '06',
       '06051',
       '06297',
       '06487',
       '06660',
       '06710',
       '06793',
       '07',
       '07004',
       '0708',
       '07434',
       '075',
       '077',
       '07766',
       '07787',
       '07b',
       '08',
       '08002000',
       '081',
       '08173',
       '08352',
       '084',
       '08653',
       '08694',
       '087',
       '09',
       '091',
       '0915',
       '092',
       '09220',
       '09474',
       '095',
       '09505',
       '09563',
       '09650',
       '097',
       '09891',
       '099',
       '0a3317d71b32',
       '0ae6aade2758',
       '0ahukewiiyimdkttqahulxomkhv9zaliqfgg9maqampurl',
       '0ahukewij8ps1xdzqahvrsvqkhrweaqqauibigb',
       '0ahukewilokkhudqahxigfqkhbnjbhkqyjcikqampei',
       '0ahukewiznlxndpqahurq1qkhxrgdvwqauiccgbampbiw',
       '0ahukewj8ryek9jqahvm5cykhtw4ccwq6aeisjagv',
       '0ahukewjfcpvm97qahvmdsakhfoed1yq6aeiijabv',
       '0ahukewjj1uj3lodqahvc8wmkhcihcdkqgqmigdaa',
       '0ahukewjt8aby39xqahujrvqkhsv3dzcqmwg8kbuwfqampiact',
       '0ampasvis',
       '0ampelect',
       '0ampfips',
       '0ampn',
       '0ampoffice',
       '0ampsession',
       '0ampyear',
       '0b4056ad8b78',
       '0bama',
       '0bozo',
       '0bwxugnzvhi',
       '0de1d1859c4',
       '0e54ab28c8e9',
       '0kpqvspwjog',
       '0life',
       '0m51',
       '0t7pnrelfdi',
       '10',
       '100',
       '1000',
       '10000',
       '100000',
       '1000000',
       '10000000',
       '1000000000',
       '10000000000i',
       '10000000x',
       '10000karma',
       '10000man',
       '10001100',
       '10002000',
       '1000ish',
       '1000x',
       '1001',
       '10079ampadgroup',
       '100b',
       '100bn',
       '100bnyr',
       '100car',
       '100fold',
       '100ft',
       '100httpwwwelectionreturnspagovenrnew',
       '100k',
       '100k400k',
       '100k8',
       '100lb',
       '100m',
       '100mo',
       '100smo',
       '100th',
       '100x',
       '101',
       '1010',
       '1012',
       '101371journalpone0111629',
       '1013k',
       '1015',
       '1015165',
       '1016',
       '101720420',
       '101httpwwwconstitutionorgjl2ndtr05htm',
       '101st',
       '102',
       '1020',
       '1022',
       '1024',
       '1025',
       '103',
       '1030',
       '104',
       '104121',
       '1043ampcontext',
       '10468',
       '105',
       '1050',
       '105000',
       '1051511',
       '105274',
       '10537155',
       '1056000',
       '106',
       '1069',
       '1069208730',
       '107',
       '1070',
       '10704',
       '108',
       '1081',
       '1081459',
       '108myr',
       '109',
       '109355',
       '10996',
       '109m',
       '109th',
       '10ampisuri',
       '10d',
       '10ft',
       '10hr',
       '10ish',
       '10k',
       '10m',
       '10mm',
       '10red',
       '10s100',
       '10second',
       '10th',
       '10x',
       '10year',
       '10yearold',
       '11',
       '110',
       '1100',
       '11000',
       '110000',
       '11000year',
       '1100mo',
       '1108',
       '11082016ampcountyid',
       '110lb',
       '110m',
       '110th',
       '111',
       '1110',
       '1112',
       '11120',
       '11120312',
       '1115',
       '112',
       '11212015',
       '112th',
       '113',
       '113b',
       '113th',
       '114',
       '114000000',
       '1143',
       '1143611',
       '114m',
       '114th',
       '115',
       '1150',
       '115000',
       '1150hr',
       '115135',
       '11536504',
       '115487579',
       '11570808',
       '116',
       '11615',
       '117',
       '118',
       '118000',
       '118th',
       '119',
       '1199',
       '119day',
       '11am',
       '11ampareatypekeygdp',
       '11b',
       '11billion',
       '11hr',
       '11k',
       '11pm',
       '11t',
       '11th',
       '12',
       '120',
       '1200',
       '12000',
       '120000',
       '120000000',
       '12002c900',
       '1200mo',
       '12017',
       '12060amp7036',
       '121',
       '12116',
       '1212',
       '1215',
       '1216',
       '12167',
       '1218',
       '12184',
       '122',
       '1222016',
       '122577',
       '1228',
       '122m',
       '123',
       '123000',
       '1234',
       '12367541',
       '123960000',
       '124',
       '124301000',
       '124588000',
       '125',
       '125000',
       '1256myr',
       '125b',
       '125m',
       '126',
       '1262016',
       '1269',
       '127',
       '12702379',
       '12740571',
       '1277',
       '12773801',
       '127mm',
       '128',
       '12800000',
       '12830632',
       '12882135',
       '128k',
       '129',
       '12d',
       '12gaug',
       '12hour',
       '12hr',
       '12k',
       '12m',
       '12member',
       '12minut',
       '12mm',
       '12th',
       '12thsmallest',
       '12year',
       '12yearold',
       '13',
       '130',
       '1300',
       '130000',
       '13000word',
       '1300gt800',
       '130k',
       '130m',
       '131',
       '1314',
       '1317m',
       '132',
       '1320',
       '132000',
       '1323459',
       '132797',
       '1328302',
       '132lb',
       '133',
       '13329',
       '1339',
       '134',
       '134142',
       '1346',
       '1347',
       '135',
       '135266',
       '135m',
       '136',
       '136th',
       '137',
       '13700',
       '1375',
       '138',
       '1385000',
       '139',
       '1397ampbih',
       '139bn',
       '13billion',
       '13bn',
       '13d',
       '13hr',
       '13httpcdnphupicomsvemupiupi15114077916352014195Oeb4b89d02f6def7b31f9ae3f5de50f
      auxpasfewlaughingatpresidentsimpromptujokearchivejpg',
       '13httpswwwarchivesgovfederalregisterelectoralcollegekeydateshtml',
       '13k',
       '13m',
       '13ralli',
       '13rd',
       '13th',
       '14',
       '140',
       '1400',
       '14000',
       '140000',
       '140209',
       '1404054',
       '140800',
       '140charact',
       '140db',
       '141',
       '141000',
       '1414',
       '1415',
       '1416',
       '142',
       '14239',
       '142855995',
       '143915',
       '144',
       '145',
       '14500',
       '1459',
       '145x',
       '146',
       '1461',
       '147',
       '148',
       '1480885526ampsr',
       '1484',
       '1488er',
       '149k',
       '14b',
       '14billion',
       '14bn',
       '14m',
       '14th',
       '14yearold',
       '15',
       '150',
       '1500',
       '15000',
       '150000',
       '15001',
       '15080',
       '150k',
       '15168',
       '15171',
       '1517ampbih',
       '1518',
       '151844',
       '152',
       '1520',
       '1520000',
       '1525',
       '152780',
       '1535',
       '153538',
       '154',
       '155',
       '156',
       '156k',
       '157',
       '158',
       '159',
       '1598',
       '15ampp',
       '15ampsmoothing',
       '15b',
       '15bn',
       '15d',
       '15hr',
       '15httpaicucdavisedupublicationsmocamocacurrentmoca09moca09chapter5pdf',
       '15ish',
       '15k',
       '15m',
       '15million',
       '15minut',
       '15second',
       '15th',
       '15year',
       '15yearold',
       '16',
       '160',
       '1600',
       '160000',
       '160000000',
       '1602',
       '160z',
       '1612136',
       '162219png',
       '1627700',
       '163',
       '1630',
       '163943',
       '164',
       '1642',
       '16432',
       '165',
       '1650',
       '16507',
       '1650k',
       '165800',
       '166',
       '167',
       '168',
       '1680',
       '16b',
       '16m',
       '16million',
       '16min',
       '16th',
       '16year',
       '16yearold',
       '17',
       '170',
       '1700',
       '17000',
       '170000',
       '170543',
       '1709',
       '170m',
       '170mm',
       '171',
       '171000',
       '1711',
       '173',
       '173351',
       '174',
       '175',
       '175db',
       '175mil',
       '176',
       '1760',
       '176465',
       '177',
       '1770',
       '1776',
       '17762016',
       '178',
       '1787',
       '1788',
       '179',
       '1790',
       '1792',
       '1796',
       '1799',
       '17990',
       '17d',
       '17mil',
       '17th',
       '17yearold',
       '18',
       '180',
       '1800',
       '18000',
       '180000',
       '1800ampyearend',
       '1801',
       '180180',
       '1803',
       '1804',
       '1804httpfivethirtyeightcomfeatureswarispeacefreedomisslaverytrumpwoninalandsli
```

```
d',
 '1808',
 '180k',
 '180º',
 '1812',
 '1816',
 '1820',
 '1821',
 '1824',
 '1825',
 '1828',
 '183',
 '1830',
 '1832',
 '1835',
 '1836',
 '184',
 '1840',
 '184566',
 '185',
 '1850',
 '1851',
 '1854304',
 '1856',
 '1856httpwwwpresidencyucsbeduwspid',
 '186',
 '1860',
 '1865',
 '1868',
 '1868516',
 '1870',
 '187000',
 '1875',
 '1876',
 '1877',
 '188000',
 '18801310',
 '1882',
 '18896800',
 '1890httpsbooksgooglecombooksid',
 '1890someth',
 '1892lb',
 '1893770',
 '1895',
 '1898',
 '18bn',
 '18th',
 '18yearold',
 '19',
 '190',
 '1900',
 '19000',
 '19021912',
 '1903',
 '1909',
 '190k',
 '1910',
 '1911',
 '19111930',
 '1912',
 '1913',
 '1914',
 '1916',
 '1917',
 '192',
 '1920',
 '192000',
 '19201932',
 '1920ampbih',
 '1920ampved',
 '1920both',
 '1920s30',
 '1923',
 '1924',
 '1927',
 '1929',
 '1929httpsenwikipediaorgwikireapportionmentactof1929',
 '193',
 '1930',
 '1930httpsenwikipediaorgwikiplantpatentactof1930',
 '1932',
 '19322010',
 '1933',
 '1934',
 '1935',
 '19371941',
 '19378102',
 '1938',
 '194',
 '1940',
 '19401941',
 '19401970',
 '1941',
 '194219',
 '1945',
 '194571',
 '1946',
 '1947',
 '1947across',
 '1947in',
 '1948',
 '1949',
 '19491415',
 '195',
 '1950',
 '1951',
 '1952',
 '19521954',
 '1953',
 '195369',
 '1955',
 '19552860',
 '1956',
 '1957',
 '1958',
 '1959',
 '195k',
 '1960',
 '1961',
 '1963',
 '1964',
 '1965',
 '19651127',
 '1966',
 '1967',
 '1968',
 '1969',
 '1969920',
 '1970',
 '197071',
 '1970sbut',
 '1971',
 '1972',
 '1973',
 '1974',
 '1975',
 '1976',
 '1977',
 '19774748',
 '1978',
 '1979',
 '19795791elector',
 '197m',
 '1980',
 '1981',
 '19811992',
 '19812010',
 '1982',
 '1983',
 '1984',
 '1984esqu',
 '1985',
 '1986',
 '1987',
 '198712',
 '1988',
 '1989',
 '199',
 '1990',
 '1990sera',
 '1991',
 '1992',
 '1993',
 '1994',
 '1994amptoyear',
 '1995',
 '199567',
 '1996',
 '1997',
 '1998',
 '1999',
 '19999999',
 '19billion',
 '19th',
 '1a',
 '1actual',
 '1amp',
 '1amp7001',
 '1amp7003',
 '1amp7004',
 '1amp7006',
 '1ampa',
 '1ampacrdn',
 '1ampdrill',
 '1ampil',
 '1ampisuri',
 '1ampmi',
 '1ampnrange',
 '1ampoi',
 '1ampunitofmeasurekeygdp',
 '1ampw',
 '1ampyeargdp',
 '1ampyeargdpend',
 '1b',
 '1bed1bath',
 '1billion',
 '1blue',
 '1bn',
 '1c1chbfenus718us718amptbm',
 '1caasueenus663us664ampq',
 '1call',
 '1child',
 '1clinton',
 '1conform',
 '1d',
 '1gdp',
 '1h08m30',
 '1http911courageorglinkeddocsglobeteaparty02jpg',
 '1httpdailycallercom20140214richliberalwomenpreferconservativemen',
 '1httpsenwikipediaorgwikiredstatesandbluest',
 '1httpsvoatcovpizzag',
 '1httpswwwredditcomrlawcomments57pcw2amygoodmanisfacingprisonforreportingonth',
 '1httpwwwglobalresearchcaamericacreatedalqaedaandtheisisterrorgroup54028812http
wwwglobalresearchcatwentysixthingsabouttheislamicstateisilthatobamadoesnotwantyo
utoknowabout5414735',
 '1httpwwwglobalresearchcabreakingususednukesoniraqafghanistanatomicbombdroppedo
ntoraboraexpert27972',
 '1httpwwwglobalresearchcachemtrailstheconsequencesoftoxicmetalsandchemicalaeros
olsonhumanhealth19047',
 '1httpwwwglobalresearchcacopenhagenandglobalwarmingtenfactsandtenmythsonclimate
change164672httpwwwglobalresearchcaglobalwarmingmediapropaganda5364444',
 '1httpwwwglobalresearchcafluoridekillingussoftly53603972httpwwwglobalresearchca
poisonistreatmentthecampaigntofluoridateamerica315683httpwwwglobalresearchcatheh
ealthimpactsoffluoridatedwatershakyscience5498885',
 '1httpwwwglobalresearchcageneticallymodifiedorganismsgmosplannedsterilizationof
humanity5511206',
 '1httpwwwglobalresearchcapariskillingsmedialiesunansweredquestionswasitafalsefl
ag5424029',
 '1httpwwwpolitifactcomtruthometerstatements2015apr16hillaryclintonhillaryclinto
nflubsfamilysimmigrationhistori',
 '1httpwwwwnycorgstorybrooklynvoterpurgeageclintonsand',
 '1jkchgihptg',
 '1k',
 '1lt',
 '1m',
 '1m02',
 '1m33',
 '1m7',
 '1mday',
 '1million',
 '1mm',
 '1np5nv1bqxozyscr7yjs1ispwqa3d',
 '1nvdh0fmzr0',
 '1productdescriptionfeaturediv',
 '1s',
 '1someth',
 '1st',
 '1state',
 '1the',
 '1to1',
 '1v1',
 '1vote',
 '1wehab7irjaq',
 '1well',
 '1were',
 '1you',
 '1yr',
 '20',
 '200',
 '2000',
 '20000',
 '200000',
 '2000000',
 '20000000',
 '200000httpiimgurcomjjsfcpepng',
 '20002006',
 '20002008ce',
 '2000but',
 '2000lt0',
 '2000mile',
 '2000milelong',
 '2001',
 '200109',
 '2001230',
 '2002',
 '200204',
 '20022006',
 '20023',
 '2003',
 '200300mo',
 '2003caa30903',
 '2004',
 '2005',
 '2006',
 '20062008',
 '20062016',
 '2007',
 '20072012',
 '2008',
 '20082010',
 '20082016',
 '20089',
 '2008ampcorpus',
 '2008ampview',
 '2008httppeoplecomcelebritybarackobamaclaimsvictorybeforehugecrowdinchicago',
 '2008httpwwwsnopescompoliticsobamaairplaneasp',
 '2009',
 '20092011',
 '200amp7035',
 '200ampindustrykey',
 '200b',
 '200ft',
 '200k',
 '200kyear',
 '200m',
 '200month',
 '200x',
 '200year',
 '201',
 '2010',
 '20102016',
 '2010httpwwwtheatlanticcompoliticsarchive201003wouldgenpetraeusrunasademocrat37
595',
 '2011',
 '20112013httpinsidersmorningstarcomtradingexecutivecompensationactiont',
 '2012',
 '20122015',
 '2012httpwwwgallupcompoll184784americansviewsoilgasindustryimprovingaspx',
 '2012httpwwwindependentcouknewsworldamericasdonaldtrumpstafferbrandonhallmichig
anguiltyelectionfrauda7449046html',
 '2013',
 '20130128',
 '20132014',
 '201365',
 '2013httpswwwredditcomrpicscomments1tugnminevertrulyunderstoodhowmuchhealthcare
in',
 '2014',
 '201484',
 '201487',
 '2014httpscdnimages1mediumcommax8001faotyacoypuhjiae3sjofapng',
 '2015',
 '201588',
 '2015amp7093',
 '2015ampstart',
 '2015httpsenwikipediaorgwikilistofusarmssalestotaiwan',
 '2015httpswwwyoutubecomwatchv',
 '2015q2ampyeargdpbegin',
 '2016',
 '20161201010840',
 '20161206',
 '20162020',
 '201623',
 '2016ampelect',
 '2016ampid',
 '2016ampoff',
 '2016hb6097',
 '2016httpbigstoryaporgarticle62acdd911e4d44a5b855acf25122bd22obamaadministratio
nconfirmsdoubledigitpremiumhik',
 '2016httpwwwnytimescominteractive20161028uspoliticsfbiletterhtmlr',
 '2016woohoo',
 '2017',
 '2018',
 '20182020',
 '2018httpsenwikipediaorgwikiunitedstatesgubernatorialelections2018',
 '2018i',
 '2019',
 '202',
 '2020',
 '2020httpiimgurcomnharum0jpg',
 '2020httpreddit5gq710',
 '2021',
 '2022',
 '2024',
 '2025',
 '2026',
 '2028',
 '2030',
 '2030k',
 '2032',
 '2036',
 '204',
 '2040',
 '2048',
 '2050',
 '206',
 '20666',
 '208',
 '20838',
 '2085287',
 '2086',
 '20868067',
 '2087',
 '208877',
 '209',
 '20hour',
 '20hr',
 '20ish',
 '20k',
 '20k30k',
 '20kyear',
 '20m',
 '20m18',
 '20m2',
 '20minut',
 '20someth',
 '20th',
 '20thcenturi',
 '20year',
 '20yearlow',
 '20yearshelflif',
 '21',
 '210',
 '2100',
 '21000',
 '210000',
 '21000year',
 '210khrhttpsenwikipediaorgwikiboeingvc25citenote4',
 '2110',
 '2117',
 '2129569',
 '213',
 '213897',
 '215',
 '215483',
 '216',
 '216739',
 '218',
 '219',
 '21ea67f47b74473525c0bbffaa97108dampoe',
 '21httpinjurypreventionbmjcomcontentearly20150609injuryprev2015041586fullpdfkey
type',
 '21m13',
 '21st',
 '21were',
 '22',
 '220',
 '2200',
 '22000',
 '2200amp7002',
 '221',
 '2226',
 '22300',
 '223000',
 '223553265',
 '224',
 '225',
 '2251',
 '225125831',
 '226',
 '22680',
 '22700',
 '228288',
 '229k',
 '22billstatus2222law2222sponsorship2222sponsored22',
 '22false',
 '22hr',
 '22httpwwwcnncom20161122usveteransstandforstandingrocktrnd',
 '22k',
 '22lr',
 '22mil',
 '22nd',
 '22wyden',
 '22x',
 '23',
 '230',
 '2300',
 …]
```

[ ]:

```

```

B. Convert each text entry into a part-of-speech tag vector (see section 6.7 in the Machine Learning with Python Cookbook).

[15]:

```python
# load libraries
from nltk import pos_tag
from nltk import word_tokenize
# nltk.download('averaged_perceptron_tagger')
```

[16]:

```python
# create list
tagged_discussion = []
# use pre-trained part of speech tagger
```

```
for i in range(len(sampledf)) :
    # tokenized each row of dataframe's txt column
    test_token2 = word_tokenize(sampledf.loc[i, "txt"])
    # apply stemmer
    pos_words = pos_tag(test_token2)
    # print(pos_words)
    tagged_discussion.append([tag for words, tag in pos_words])
    # coverting list to string
    #pos_words1 = " ".join(str(x) for x in pos_words)
    # writting back processed removed stop words to dataframe
    # updating the row values for txt column
    #sampledf.loc[i, ["txt"]] = pos_words1
# printing pos_words of last row in the dataframe
print(pos_words)
# printing the tag list of all rows of data frame
print(tagged_discussion)
```

```
IOPub data rate exceeded.
The notebook server will temporarily stop sending output
to the client in order to avoid crashing it.
To change this limit, set the config variable
`--NotebookApp.iopub_data_rate_limit`.

Current values:
NotebookApp.iopub_data_rate_limit=1000000.0 (bytes/sec)
NotebookApp.rate_limit_window=3.0 (secs)
```

[17]:

```
# import libraries
from sklearn.preprocessing import MultiLabelBinarizer
```

[18]:

```
# Use one-hot encoding to convert the tags into feature
one_hot_multi = MultiLabelBinarizer()
one_hot_multi.fit_transform(tagged_discussion)
```

[18]:

```
array([[0, 0, 0, …, 0, 0, 0],
       [0, 0, 0, …, 0, 0, 0],
       [0, 0, 0, …, 0, 0, 0],
       …,
       [0, 0, 0, …, 0, 0, 0],
       [0, 0, 0, …, 0, 0, 0],
       [0, 0, 0, …, 0, 0, 0]])
```

[19]:

```
# using classes_ we can see that each feature is a part of speech tag:
# show feature name
one_hot_multi.classes_
```

```
[19]: array(['$', "''", 'CC', 'CD', 'DT', 'EX', 'FW', 'IN', 'JJ', 'JJR', 'JJS',
              'LS', 'MD', 'NN', 'NNP', 'NNPS', 'NNS', 'PDT', 'POS', 'PRP',
              'PRP$', 'RB', 'RBR', 'RBS', 'RP', 'SYM', 'TO', 'UH', 'VB', 'VBD',
              'VBG', 'VBN', 'VBP', 'VBZ', 'WDT', 'WP', 'WP$', 'WRB', '``'],
             dtype=object)
```

C. Convert each entry into a term frequency-inverse document frequency (tfidf) vector (see section 6.9 in the Machine Learning with Python Cookbook).

```
[20]: # import libraries
      from sklearn.feature_extraction.text import TfidfVectorizer
      ftidf = TfidfVectorizer()
```

```
[21]: # Create tf-idf feature matrix
      feature_matrix = ftidf.fit_transform(sampledf['txt'].to_numpy())
      # show tf-idf feature matrix as dense matrix
      feature_matrix.toarray()
```

```
[21]: array([[0., 0., 0., …, 0., 0., 0.],
              [0., 0., 0., …, 0., 0., 0.],
              [0., 0., 0., …, 0., 0., 0.],
              …,
              [0., 0., 0., …, 0., 0., 0.],
              [0., 0., 0., …, 0., 0., 0.],
              [0., 0., 0., …, 0., 0., 0.]])
```

```
[22]: # show tf-idf feature names
      ftidf.vocabulary_
```

```
[22]: {'well': 32055,
       'great': 12218,
       'someth': 26970,
       'belief': 4481,
       'offic': 21074,
       'doubt': 9154,
       'trump': 29916,
       'would': 32585,
       'fight': 10820,
       'un': 30498,
       'im': 15969,
       'realli': 24085,
       'happi': 13430,
       'obama': 20959,
       'couldoh': 7449,
       'wait': 31779,
       'right': 24870,
       'mr': 19849,
       'presid': 23075,
       'given': 11882,
       'input': 16405,
       'apart': 3441,
       'say': 25556,
       'wrong': 32641,
       'argument': 3604,
       'clearli': 6590,
       'get': 11801,
       'frustrat': 11434,
       'reason': 24116,
       'want': 31845,
       'way': 31950,
       'foundat': 11249,
       'complex': 7004,
       'problem': 23227,
       'advanc': 2555,
       'grade': 12153,
       'decent': 8152,
       'sat': 25509,
       'type': 30187,
       'test': 28919,
       'math': 18863,
       'dont': 9085,
       'understand': 30637,
       'lot': 18332,
       'mathemat': 18864,
       'answer': 3193,
       'time': 29411,
       'figur': 10823,
       'common': 6945,
       'sens': 25983,
       'work': 32510,
       'around': 3648,
       'question': 23741,
       'ill': 15934,
       'prepar': 23046,
       'take': 28640,
       'colleg': 6830,
       'level': 17908,
       'cours': 7523,
       'despit': 8483,
       'averag': 3977,
       'score': 25699,
       'theyr': 29111,
       'tri': 29798,
       'bust': 5544,
       'kid': 17357,
       'ball': 4164,
       'far': 10556,
       'expert': 10327,
       'tpp': 29667,
       'tend': 28870,
       'agre': 2693,
       'push': 23629,
       'creat': 7617,
       'econom': 9526,
       'bulwark': 5454,
       'china': 6343,
       'pacif': 21695,
       'administr': 2529,
       'recogn': 24188,
       'increas': 16173,
       'strength': 27698,
       'matur': 18888,
       'bellicos': 4490,
       'see': 25833,
       'south': 27085,
       'sea': 25774,
       'us': 31110,
       'alli': 2867,
       'penetr': 22094,
       'mani': 18694,
       'emerg': 9778,
       'market': 18786,
       'otherwis': 21440,
       'natur': 20144,
       'align': 2837,
       'alway': 2954,
       'thought': 29289,
       'curiou': 7828,
       'critiqu': 7676,
       'hardli': 13455,
       'saw': 25553,
       'mention': 19135,
       'track': 29674,
       'record': 24202,
       'someon': 26965,
       'railroad': 23876,
       'worker': 32517,
       'corpor': 7369,
       'interest': 16519,
       'must': 19992,
       'felt': 10731,
       'huge': 15686,
       'import': 16057,
       'use': 31144,
       'much': 19884,
       'polit': 22676,
       'capit': 5819,
       'like': 18032,
       'mayb': 18905,
       'could': 7445,
       'better': 4602,
       'certainli': 6105,
       'seem': 25841,
       'yet': 32884,
       'best': 4581,
       'manag': 18663,
       'vari': 31378,
       'play': 22530,
       'howev': 14215,
       'strateg': 27673,
       'throw': 29338,
       'begin': 4452,
       'signific': 26440,
       'withdraw': 32423,
       'foreign': 11160,
       'definit': 8240,
       'move': 19832,
       'fill': 10834,
       'void': 31623,
       'may': 18903,
       'look': 18279,
       'end': 9847,
       'western': 32123,
       'hegemoni': 13652,
       'elect': 9661,
       'ascend': 3706,
       'nationalist': 20121,
       'movement': 19833,
       'britain': 5294,
       'franc': 11296,
       'elsewher': 9729,
       'what': 32149,
       'sad': 25360,
       'appear': 3486,
       'entir': 9937,
       'selfinflict': 25904,
       'noth': 20766,
       'forc': 11140,
       'retract': 24719,
       'develop': 8530,
       'democraci': 8333,
       'simpli': 26477,
       'result': 24686,
       'myopic': 20020,
       'reaction': 24042,
       'global': 11926,
       'thank': 28950,
       'feel': 10704,
      'nowhttpsthenypostfileswordpresscom201611161109clintoncampelection01jpgquality':
      20824,
       '90ampstrip': 2180,
       'allampw': 2847,
       '642': 1862,
       'lol': 18229,
       'delet': 8287,
       'cant': 5804,
       'racist': 23827,
       'black': 4786,
       'friend': 11381,
       'lololol': 18243,
       'vast': 31388,
       'major': 18604,
       'misogynist': 19497,
       'histori': 13972,
       'werent': 32116,
       'marri': 18803,
       'women': 32470,
       'nope': 20711,
       'your': 32945,
       'smoke': 26768,
       'bad': 4109,
       'lung': 18459,
       'tobacco': 29502,
       'carcinogen': 5847,
       'plain': 22502,
       'old': 21161,
       'heavi': 13625,
       'metal': 19190,
       'mj': 19549,
       'doesnt': 9008,
       'industri': 16246,
       'farm': 10575,
       'ltthat': 18424,
       'exactli': 10214,
       'mean': 18982,
       'especi': 10065,
       'power': 22901,
       'exist': 10283,
       'awaygt': 4001,
       'politician': 22686,
       'one': 21207,
       'clear': 6581,
       'exampl': 10220,
       'democrat': 8339,
       'gay': 11687,
       'marriag': 18804,
       'hillari': 13892,
       'evolv': 10204,
       'respond': 24659,
       'popular': 22769,
       'opinion': 21308,
       'societi': 26884,
       'give': 11879,
       'judg': 17139,
       'show': 26348,
       'mainstream': 18598,
       'republican': 24579,
       'away': 4000,
       'arent': 3592,
       'expand': 10304,
       'youv': 32958,
       'dupe': 9391,
       'propaganda': 23364,
       'believ': 4482,
       'isnt': 16738,
       'canada': 5757,
       'uk': 30422,
       'current': 7832,
       'legal': 17808,
       'mechan': 19013,
       'state': 27437,
       'seced': 25802,
       'save': 25544,
       'success': 28240,
       'revolut': 24778,
       'california': 5689,
       'referendum': 24317,
       'laugh': 17664,
       'feder': 10687,
       'court': 7526,
       'meaningless': 18988,
       'word': 32504,
       'keep': 17278,
       'fire': 10880,
       'contain': 7233,
       'he': 13568,
       'anyth': 3417,
       'peopl': 22119,
       'stop': 27628,
       'pretend': 23121,
       'air': 2745,
       'declar': 8172,
       'dictat': 8613,
       'life': 18006,
       'honestli': 14103,
       'upset': 31056,
       'classic': 6555,
       'case': 5934,
       'govern': 12106,
       'depart': 8402,
       'interior': 16532,
       'rat': 23967,
       'ass': 3740,
       'nativ': 20138,
       'american': 2997,
       'govt': 12128,
       'treati': 29775,
       'commun': 6953,
       'organ': 21375,
       'support': 28385,
       'redistribut': 24260,
       'wealth': 31976,
       'money': 19664,
       'lender': 17851,
       'call': 5699,
       'loser': 18324,
       'cri': 7649,
       'unattract': 30523,
       'good': 12027,
       'invok': 16626,
       'httpswwwredditcomrpoliticscomments5eqcqnstanfordprofessordevelopstrumpconversi
      ondaee57': 14972,
       'gtone': 12850,
       'characterist': 6192,
       'abus': 2376,
       'notic': 20778,
       'disabl': 8724,
       'attitud': 3869,
       'client': 6613,
       'resist': 24640,
       'justif': 17190,
       'feedback': 10699,
       'loop': 18303,
       'establish': 10077,
       'control': 7283,
       'justifi': 17191,
       'acquiesc': 2454,
       'there': 29069,
       'behavior': 4460,
       'wise': 32409,
       'escap': 10054,
       'hold': 14039,
       'true': 29895,
       'relationship': 24417,
       'prison': 23191,
       'polic': 22649,
       'kind': 17380,
       'authoritarian': 3929,
       'regim': 24354,
       'broader': 5305,
       'messag': 19177,
       'behav': 4459,
       'first': 10902,
       'place': 22491,
       'gtrememb': 12912,
       'next': 20377,
       'four': 11257,
       'year': 32818,
       'hear': 13601,
       'trope': 29868,
       'appointe': 3501,
       'that': 28959,
       'boo': 5022,
       'mike': 19314,
       'penc': 22084,
       'protest': 23440,
       'street': 27690,
       'explain': 10330,
       'death': 8108,
       'threat': 29308,
       'got': 12081,
       'think': 29144,
       'poor': 22745,
       'person': 22213,
       'handout': 13393,
       'need': 20217,
       'submiss': 27808,
       'automat': 3947,
       'remov': 24482,
       'either': 9641,
       'link': 18090,
       'shorten': 26315,
       'redirector': 24258,
       'rpolit': 25154,
       'allow': 2881,
       'shortern': 26318,
       'user': 31157,
       'abl': 2310,
       'tell': 28847,
       'go': 11956,
       'click': 6606,
       'encourag': 9843,
       'resubmit': 24684,
       'url': 31105,
       'point': 22623,
       'directli': 8715,
       'content': 7241,
       'submit': 28180,
       'inform': 16311,
       'found': 11248,
      'herehttpswwwredditcomrpoliticswikifiltereddomainswikilinkshorteners2fredirect':
      13743,
       'bot': 5097,
       'action': 2466,
       'perform': 22166,
       'pleas': 22545,
       'contact': 7232,
       'moder': 19599,
       'subredditmessagecomposeto': 28192,
       'concern': 7047,
       'pretti': 23128,
       'lay': 17711,
       'man': 18660,
       'term': 28895,
       'suppli': 28379,
       'side': 26411,
       'discredit': 8771,
       'anybodi': 3403,
       'parti': 21860,
       'fuck': 11449,
       'moron': 19766,
       'thing': 29126,
       'made': 18539,
       'name': 20070,
       'read': 24047,
       'pick': 22372,
       'alreadi': 2918,
       'past': 21922,
       'fall': 10503,
       'goddamn': 11973,
       'funni': 11524,
       'make': 18611,
       'rsolips': 25178,
       'fair': 10476,
       'help': 13693,
       'pay': 21980,
       'crazi': 7608,
       'spend': 27198,
       'plan': 22505,
       'guess': 13120,
       'benefit': 4512,
       'tax': 28739,
       'though': 29286,
       'crucifi': 7721,
       'liber': 17951,
       'univers': 30785,
       'campus': 5753,
       'exact': 10213,
       'hardest': 13452,
       'shut': 26385,
       'conserv': 7172,
       'speech': 27181,
       'idea': 15851,
       'honest': 14101,
       'extrem': 10382,
       'are': 3575,
       'hav': 13527,
       'ideologu': 15870,
       'care': 5852,
       'team': 28788,
       'real': 24071,
       'issu': 16754,
       'oh': 21122,
       'yeah': 32805,
       'still': 27590,
       'schedul': 25620,
       'raid': 23873,
       'dispensari': 8846,
       'violat': 31563,
       'law': 17682,
       'consid': 7188,
       'libertarian': 17968,
       'cut': 7860,
       'militari': 19336,
       'dismantl': 8831,
       'dea': 8066,
       'big': 4655,
       'ever': 10158,
       'theyll': 29110,
       'leav': 17767,
       'high': 13836,
       'dri': 9257,
       'ask': 3727,
       'snowden': 26841,
       'brain': 5167,
       'rtheredpil': 25204,
       'guy': 13194,
       'didnt': 8633,
       'antitrump': 3371,
       'come': 6874,
       'said': 25391,
       'stuff': 27759,
       'suppos': 28396,
       'intro': 16585,
       'phone': 22334,
       'knowingli': 17460,
       'harbor': 13440,
       'bin': 4706,
       'laden': 17560,
       'rescind': 24617,
       'aid': 2734,
       'suck': 28250,
       'fat': 10612,
       'dick': 8599,
       'ugli': 30356,
       'also': 2927,
       'die': 8636,
       'conway': 7310,
       'queen': 23734,
       'long': 18255,
       'game': 11613,
       'complet': 7001,
       'elimin': 9710,
       'public': 23552,
       'educ': 9578,
       'despis': 8482,
       'fact': 10442,
       'dollar': 9036,
       'lower': 18362,
       'class': 6549,
       'turn': 30081,
       'compet': 6989,
       'privileg': 23204,
       'posit': 22806,
       'truli': 29909,
       'incred': 16177,
       'mental': 19133,
       'gymnast': 13206,
       'done': 9067,
       'weird': 32041,
       'stock': 27608,
       'run': 25259,
       'integr': 16491,
       'refus': 24341,
       'know': 17453,
       'ahead': 2716,
       'debat': 8114,
       'leak': 17749,
       'camp': 5737,
       'wouldv': 32593,
       'possibl': 22815,
       'reduc': 24273,
       'everyon': 10175,
       'cheer': 6258,
       'settl': 26064,
       'gt': 12342,
       'hell': 13674,
       'she': 26156,
       'less': 17879,
       'corrupt': 7400,
       'almost': 2904,
       'zero': 33010,
       'substanc': 28211,
       '100': 136,
       'fear': 10666,
       'monger': 19675,
       'rhetor': 24822,
       'jesu': 16981,
       'notch': 20759,
       'wikileak': 32313,
       'undeni': 30595,
       'proof': 23357,
       'cabinet': 5642,
       'assembl': 3757,
       'citigroup': 6501,
       'execut': 10259,
       'even': 10150,
       '2008': 837,
       'safe': 25375,
       'space': 27116,
       'ff': 10779,
       'topic': 29584,
       'statement': 27453,
       'gtto': 13022,
       'explicitli': 10336,
       'focu': 11076,
       'follow': 11093,
       'servic': 26051,
       'policymak': 22669,
       'privat': 23198,
       'stori': 27636,
       'demonstr': 8363,
       'lobbi': 18187,
       'candidaci': 5770,
       'fund': 11508,
       'group': 12305,
       'donor': 9081,
       'gtthi': 13011,
       'includ': 16140,
       'nonpolit': 20638,
       'ex': 10210,
       'barack': 4243,
       'paint': 21728,
       'pictur': 22381,
       'rel': 24413,
       'associ': 3780,
       'dian': 8585,
       'feinstein': 10716,
       'father': 10621,
       'predat': 22981,
       'attempt': 3859,
       'murder': 19955,
       'arnold': 3643,
       'schwarzenegg': 25670,
       'intern': 16540,
       'unless': 30800,
       'discuss': 8779,
       'focus': 11077,
       'implic': 16051,
       'tension': 28882,
       'greec': 12229,
       'itali': 16775,
       'rise': 24935,
       'cost': 7418,
       'feta': 10761,
       'chees': 6261,
       'media': 19019,
       'explicit': 10335,
       'connot': 7152,
       'cnn': 6701,
       'wolf': 32455,
       'blitzer': 4860,
       'simpl': 26471,
       'straight': 27651,
       'lie': 17997,
       'era': 10014,
       'order': 21362,
       'behind': 4465,
       'never': 20314,
       'id': 15847,
       'day': 8037,
       'account': 2419,
       'massiv': 18841,
       'growth': 12327,
       'came': 5724,
       'tenur': 28889,
       'gtpolit': 12881,
       'correct': 7388,
       'invent': 16604,
       'phantom': 22293,
       'enemi': 9871,
       'let': 17898,
       'trigger': 29830,
       'incid': 16132,
       'havent': 13534,
       'happen': 13424,
       'john': 17076,
       'stuart': 27740,
       'mill': 19350,
       'talk': 28660,
       'back': 4055,
       '19th': 732,
       'centuri': 6092,
       'pakistan': 21733,
       'compliment': 7011,
       'liter': 18128,
       'everi': 10165,
       'sentenc': 26001,
       'headlin': 13577,
       'joke': 17090,
       'wors': 32561,
       'thedonald': 29001,
       'meme': 19103,
       'conspiraci': 7200,
       'theori': 29054,
       'pure': 23609,
       'nonsens': 20666,
       '247': 1036,
       'tic': 29377,
       'tac': 28606,
       'skittl': 26594,
       'recount': 24207,
       'michigan': 19248,
       'pointless': 22627,
       'yah': 32773,
       'actual': 2482,
       'outsourc': 21527,
       'job': 17046,
       '1000': 137,
       'send': 25974,
       '1100': 230,
       'countri': 7495,
       'annoy': 3175,
       'sinc': 26488,
       'neg': 20229,
       'might': 19305,
       'selfidentifi': 25898,
       'sjw': 26559,
       'internet': 16545,
       'full': 11489,
       'bullshit': 5449,
       'youtub': 32955,
       'enjoy': 9898,
       'clown': 6676,
       'fiesta': 10812,
       'shit': 26233,
       'awkwardli': 4008,
       'view': 31529,
       'social': 26872,
       'eg': 9608,
       'gun': 13155,
       'appeal': 3483,
       'voter': 31662,
       'hint': 13945,
       'learn': 17757,
       'love': 18351,
       'revolv': 24783,
       'breachload': 5206,
       'rifl': 24865,
       'shotgun': 26334,
       'walmart': 31828,
       'situat': 26545,
       'locat': 18197,
       'town': 29651,
       'infrastructur': 16326,
       'build': 5418,
       'road': 25000,
       'sure': 28424,
       'store': 27634,
       'connect': 7147,
       'water': 31930,
       'sewer': 26080,
       'electr': 9684,
       'sale': 25413,
       'revenu': 24755,
       'enough': 9904,
       'bond': 5003,
       'took': 29565,
       'httpwwwstrongtownsorgthegrowthponzischem': 15570,
       'miss': 19510,
       'second': 25805,
       'welfar': 32052,
       'taxpay': 28759,
       'put': 23643,
       'disson': 8893,
       'shown': 26354,
       'theoret': 29053,
       'scenario': 25610,
       'white': 32228,
       'pride': 23155,
       'asshol': 3770,
       'okay': 21150,
       'adopt': 2541,
       'fail': 10470,
       'substanti': 28215,
       'differ': 8649,
       'two': 30158,
       'stupid': 27771,
       'broken': 5315,
       'famili': 10527,
       'low': 18360,
       'drug': 9291,
       'agenc': 2659,
       'statist': 27494,
       'uwhyyouareverywrong': 31295,
       'apropo': 3525,
       'smackdown': 26719,
       'udamean1': 30286,
       'clinton': 6631,
       'daughter': 8025,
       'mother': 19802,
       'jone': 17099,
       'writer': 32636,
       'rag': 23866,
       'accus': 2431,
       'news': 20338,
       'stretch': 27703,
       'david': 8030,
       'corn': 7355,
       'pussi': 23635,
       'access': 2397,
       '17': 511,
       'report': 24543,
       'employe': 9813,
       'bill': 4686,
       'countless': 7493,
       'sexual': 26086,
       'improprieti': 16080,
       'retort': 24717,
       'girl': 11868,
       'dad': 7920,
       'probabl': 23216,
       'told': 29529,
       'write': 32632,
       '15': 435,
       'total': 29617,
       'agricultur': 2703,
       'product': 23263,
       'ship': 26221,
       'rural': 25280,
       'area': 3577,
       'farmer': 10577,
       'vote': 31650,
       'maga': 18556,
       'copout': 7335,
       'scientif': 25679,
       'evid': 10188,
       'precursor': 22980,
       'medicin': 19047,
       'bigli': 4668,
       'dumb': 9358,
       'nah': 20057,
       'ha': 13223,
       'goodman': 12036,
       'upvot': 31073,
       'uniron': 30779,
       'unqualifi': 30868,
       'investig': 16610,
       'treason': 29768,
       'johnson': 17079,
       'option': 21341,
       'jill': 17019,
       'stein': 27533,
       'respect': 24653,
       'candid': 5769,
       'deserv': 8467,
       'els': 9726,
       'withdrawn': 32424,
       'left': 17785,
       'ideal': 15854,
       'third': 29159,
       'world': 32546,
       'counti': 7487,
       'easi': 9481,
       'fix': 10927,
       'beyond': 4618,
       'remind': 24473,
       'subreddit': 28191,
       'civil': 6511,
       'discussionhttpswwwredditcomrpoliticswikirulesandregswikipleasebecivil': 8781,
       'troll': 29858,
       'children': 6327,
       'clever': 6602,
       'attack': 3851,
       'whether': 32196,
       'implicit': 16052,
       'permittedhttpswwwredditcomrpoliticswikirulesandregswikinopersonalattack':
      22194,
       'shill': 26209,
       'proper': 23375,
       'conduct': 7084,
       'modmail': 19614,
       'gener': 11728,
       'jerk': 16968,
       'bait': 4144,
       'hate': 13514,
       'etc': 10090,
       'downvot': 9196,
       'comment': 6900,
       'disagre': 8729,
       'will': 32335,
       'qualiti': 23705,
       'held': 13670,
       'incivil': 16137,
       'escal': 10053,
       'ban': 4186,
       'uncivil': 30557,
       'repli': 24536,
       'sourc': 27049,
       'eh': 9625,
       'spell': 27191,
       'layman': 17714,
       'mislead': 19491,
       'caricatur': 5874,
       'reflect': 24325,
       'realiti': 24083,
       'supplysid': 28383,
       'policyhttpswwwyoutubecomwatchv': 22667,
       'rcbelgaowu': 24016,
       'intent': 16504,
       'part': 21855,
       'httpsenwikipediaorgwikidemocratpartyepithet': 14608,
       'hahahahaha': 13270,
       'crack': 7571,
       'whole': 32252,
       'bigot': 4672,
       'equat': 9998,
       'along': 2908,
       'express': 10351,
       'empathi': 9802,
       'evil': 10195,
       'anyon': 3415,
       'caught': 5996,
       'rape': 23942,
       'child': 6315,
       'strangl': 27667,
       'puppi': 23604,
       'moon': 19726,
       'walk': 31810,
       'immigr': 16008,
       'isi': 16714,
       'idiot': 15880,
       'patriot': 21957,
       'muh': 19905,
       'fascism': 10592,
       'lib': 17947,
       'huh': 15692,
       'repeal': 24520,
       'small': 26721,
       'token': 29525,
       'remain': 24459,
       'continu': 7260,
       'disappoint': 8734,
       'afraid': 2623,
       'special': 27159,
       'damn': 7959,
       'shame': 26121,
       'wall': 31816,
       'vs': 31715,
       'without': 32431,
       'ad': 2491,
       'number': 20885,
       'comfort': 6887,
       'similar': 26465,
       'janki': 16905,
       'written': 32639,
       '26': 1075,
       '20': 806,
       '10': 135,
       '30': 1212,
       '13': 348,
       '43': 1505,
       'new': 20322,
       'subtract': 28227,
       'break': 5212,
       '1s': 793,
       'togeth': 29520,
       'combin': 6869,
       'add': 2500,
       'harder': 13450,
       'least': 17764,
       'head': 13569,
       'assum': 3786,
       'admit': 2534,
       'feloni': 10729,
       'gave': 11681,
       'classifi': 6558,
       'materi': 18860,
       'biograph': 4719,
       'qualifi': 23702,
       'handl': 13389,
       'drop': 9280,
       'tall': 28663,
       'gtif': 12705,
       'pot': 22872,
       'set': 26061,
       'decad': 8139,
       'marijuana': 18771,
       'littl': 18141,
       'bit': 4752,
       'exagger': 10216,
       'hmm': 14000,
       'last': 17639,
       'rememb': 24468,
       'video': 31517,
       'quit': 23769,
       'idk': 15882,
       'sound': 27042,
       'nearli': 20196,
       'deal': 8087,
       'describ': 8456,
       'hard': 13442,
       'alcohol': 2805,
       'danger': 7973,
       'pack': 21698,
       'stan': 27394,
       'fabul': 10420,
       'gold': 11999,
       'furnitur': 11537,
       'whitehous': 32233,
       'makeshttpsenmwikipediaorgwikiinternationallaw': 18617,
       'watch': 31922,
       'everyth': 10179,
       'constitu': 7213,
       'everywher': 10183,
       'incub': 16182,
       'cmon': 6695,
       'cheat': 6238,
       'incas': 16117,
       'hire': 13961,
       'kindheart': 17384,
       'intellig': 16497,
       'woman': 32465,
       'wrangl': 32611,
       'whackitud': 32141,
       'lost': 18329,
       'lose': 18322,
       'wasnt': 31907,
       'worthi': 32574,
       'talent': 28656,
       'shitton': 26275,
       'unfortun': 30721,
       'tangl': 28692,
       'spin': 27224,
       'longer': 18259,
       'tie': 29386,
       'gotten': 12091,
       'defens': 8226,
       'vocal': 31614,
       'nice': 20391,
       'anymor': 3411,
       'levelhead': 17909,
       'victim': 31509,
       'stay': 27509,
       'inner': 16389,
       'circl': 6474,
       'reput': 24603,
       'suffer': 28266,
       'sadli': 25369,
       'uh': 30362,
       'fyi': 11561,
       'brought': 5333,
       'race': 23805,
       'guilti': 13137,
       'live': 18147,
       'echo': 9514,
       'chamber': 6142,
       'somewhat': 26978,
       'scienc': 25674,
       'standard': 27398,
       'practic': 22923,
       'texa': 28932,
       'biolog': 4722,
       'york': 32918,
       'wiggl': 32308,
       'room': 25089,
       'cram': 7581,
       'curriculum': 7835,
       'needless': 20222,
       'confus': 7115,
       'error': 10045,
       'find': 10854,
       'effect': 9597,
       …}
```

Follow-Up Question

For the three techniques in problem (2) above, give an example where each would be useful.

[23]:
```
#### Tagging Parts of Speech
#### POS tagging is very key in text-to-speech systems, information extraction,␣
 ↪machine translation, and word sense disambiguation.
#### It is useful in labeling named entities like people or places.
#### example,
#### let’s say we have a language model that understands the English language
#### How can our model tell the difference between the word "address" used in␣
 ↪different contexts?
#### "I would like to address the public on this issue"
#### "We need your shipping address"
#### "address" in the first sentence is a Verb
#### whereas "address" in the second sentence is a Noun
#### Identifying the part of speech of the various words in a sentence can help␣
 ↪in defining its meanings.
#### In the example above, if the word "address" in the first sentence was a␣
 ↪Noun, the sentence would have an entirely different meaning. Its part of␣
 ↪speech is dependent on the context
```

[24]:
```
#### Encoding Text as BAG of Words
#### Bag of words (BOW) is a technique to extract features from the text for␣
 ↪Natural Language Processing.
#### It’s an algorithm that transforms the text into fixed-length vectors. This␣
 ↪is possible by counting the number of times the word is present in a␣
 ↪document in a document.
#### The word occurrences allow to compare different documents and evaluate␣
 ↪their similarities for applications, such as search, document␣
 ↪classification, and topic modeling..
#### example,
#### We could be interested in analyzing the reviews about Game of Thrones:
#### Review 1: Game of Thrones is an amazing tv series!
#### Review 2: Game of Thrones is the best tv series!
#### Review 3: Game of Thrones is so great
#### we only considered only unigram (single words) or bigrams(couples of␣
 ↪words), but also trigrams can be taken into account to extract features.␣
 ↪Stop words can be removed too as we saw, but there are still some␣
 ↪disadvantages.
#### The order and the meaning of the words are lost using this method.
#### For this reason, other approaches are preferred to extract features from␣
 ↪the text, like TF-IDF
```

[25]:
```
#### Weighted Word Importance
#### here we are comparing the frequence of words in a document (a tweet,␣
 ↪moview review speech transcripyt)
```

```
#### with the frequency of words in all other documents using term frequency-inverse document frequency
#### example
#### TF*IDF is used by search engines to better understand the content that is undervalued. For example, when you search for "Coke" on Google,
#### Google may use TF*IDF to figure out if a page titled "COKE" is about:
#### a) Coca-Cola. b) Cocaine. c) A solid, carbon-rich residue derived from the distillation of crude oil.d) A county in Texas.
```