1. Analysis of State of the Union Addresses dataset: Description

State of Union Messages to the Congress are mandated by Article II, Section 3 of the United States Constitute. He shall from time to time give to the Congress information of the state of the union and recommend to their consideration such measures as he shall judge necessary and expedient" George Washington established the precedent that clarifies the phrase "From time to time" it indicates since 1790 with occasional exceptions, State of the union messages have been delivered once annually.

The mission of Project Gutenberg is to encourage the creation and distribution of ebooks. Complete State of the Union addresses from 1790 to 2016, contains 41 US Presidential State of the Union Addresses are small subsets of the Project Gutenberg eBook corpus. The State of Union Addresses dataset is a collection of annual speeches delivered by the presidents of the United States, From George Washington to Barack Obama, to a joint session of the United States Congress for the span of 1790-2016. The Dataset is divided in to 2 sets. It contains 3 files,

- The first one addresses delivered between 1790 and 1860, It contains the collection of around 70 documents,each document is the address by the President of the United States to its Fellow Citizens of the Senate and House of Representatives From George Washington to James Buchanan.
- The second one addresses delivered between 1946 and 2016, It contains the collection of around 70 documents,each document is the address by the President of the United States to its Fellow Citizens of the Senate and House of Representatives From Harry Truman to Barack Obama.
- The third one is a policy description of the Project Gutenberg eBook("State Union Policy).

This eBook was produced by James Linden. These eBooks were prepared by thousands of volunteers. The eBooks are readable by both Humans and by Computers, Since 1971

2. Analysis of State of the Union Addresses dataset: Part1

A) The analysis tasks areas follows

- list the top 50 words by frequency (normalized by the length of the document)
- list the top 50 bigrams by frequencies, and
- list the top 50 bigrams by their Mutual Information scores (using min frequency 5)

Note- I divided the word count by total length of document into normalize word frequency distribution.

Description of Task 2:

As part of pre-processing I did following steps:

- Getting the Content and go throw the documents to know what it's about
- **Tokenization and Changing to lower case(Normalization)**

  Words are tokenized by the text:-
  --

```
#and converted all the characters to lowercase
SUAtokens = nltk.word_tokenize(raw)
SUAwords = [w.lower() for w in SUAtokens]

#print first 200 words which is tokenized and are in lowercase
print("Tokenized and lowercase")
print(SUAwords[:200])
```

This approach produced some unnecessary words. In order to remove unnecessary words like non-alphanumeric words and stoplist, I used following code:

```
def alpha_filter(w):
  # pattern to match word of non-alphabetical characters
  pattern = re.compile('^[^a-z]+$')
  if (pattern.match(w)):
    return True
  else:
    return False
```

```
#to get the stopwords list
stopwords =nltk.corpus.stopwords.words('english') +
list(string.punctuation)
print("List of stopwords")
print(stopwords)
```

Note-I normalized the text to lowercase so that the distinction between The and the is ignored. I converted all tokens into lower case as many functions of nltk are case-sensitive.

After getting the output, I updated stop word list to -

```
morestopword=['upon','would','say','u','shall','\'s','could','must',’us’,’also’]
stopwords += morestopwords
```

- **Stemming and lemmatization**

I tried three stemmers-Lancaster, Porter and Snowball stemmer. I also examined document using WordNet lemmatizer

```
porter = nltk.PorterStemmer()
lancaster = nltk.LancasterStemmer()
wnl=nltk.WordNetLemmatizer()
from nltk.stem import SnowballStemmer
snowball_stemmer = SnowballStemmer('english')

SUAPstem = [porter.stem(t) for t in stoppedSUAwords]
print('Porter\n', SUAPstem[:200])

SUALstem = [lancaster.stem(t) for t in stoppedSUAwords]
print('Lancaster\n', SUALstem[:200])

SUALemma = [wnl.lemmatize(t) for t in stoppedSUAwords]
print('WordNet Lemmatizer\n', SUALemma[:200])

SUASnstem = [snowball_stemmer.stem(t) for t in stoppedSUAwords]
print('Snowball Stemmer\n', SUAPstem[:200])
```

I found that Lancaster stemmer was severe on some words like december resulted into ev and union into ‘un’ whereas Snowball stemmer hardly changed any word compared to other two stemmers.

The WordNet lemmatizer only removes affixes if the resulting word is in its dictionary like lying remains same instead of changing to lie. So, I decided to use either of WordNet lemmatization or Porter stemming.

In summary of all the above steps:-

- Briefly state why you chose the processing options that you did.

Word-tokenizer-> Normalized text(lower case)->alpha filter->stopwords->Porter stemmer->WordNet Lemmatizer-> generated frequency distribution-> updated stopwords->generated frequency distribution

- Here I did tokenizer from NLTK and Lemmatizer both to check the result from both of them.
- After seeing the results the NLTK tokenizer produced the results for bigrams and PMI better than the Lemmatizer does.
- So I stick to the tokenizer from NLTK.

3. Analysis of State of the Union Addresses dataset: Part2

A) The analysis tasks areas follows

- list the top 50 words by frequency (normalized by the length of the document)
- list the top 50 bigrams by frequencies, and
- list the top 50 bigrams by their Mutual Information scores (using min frequency 5)

Description of Task 3:

As part of pre-processing I did following steps:

- Getting the Content and go throw the documents to know what it's about
- **Tokenization and Changing to lower case(Normalization)**

  Words are tokenized by the text:-
  --

```
#and converted all the characters to lowercase
SUBtokens = nltk.word_tokenize(raw)
SUBwords = [w.lower() for w in SUAtokens]

#print first 200 words which is tokenized and are in lowercase
print("Tokenized and lowercase")
print(SUBwords[:200])
```

This approach produced some unnecessary words. In order to remove unnecessary words like non-alphanumeric words and stoplist, I used following code:

```
def alpha_filter(w):
  # pattern to match word of non-alphabetical characters
  pattern = re.compile('^[^a-z]+$')
  if (pattern.match(w)):
    return True
  else:
    return False

#to get the stopwords list
stopwords =nltk.corpus.stopwords.words('english') +
list(string.punctuation)
print("List of stopwords")
print(stopwords)
```

Note-I normalized the text to lowercase so that the distinction between The and the is ignored. I converted all tokens into lower case as many functions of nltk are case-sensitive.

After getting the output, I updated stop word list to -
morestopword=['upon','would','say','u','shall','\'s','could','must',’us’,’also’]

- **Stemming and lemmatization**

I tried three stemmers-Lancaster, Porter and Snowball stemmer. I also examined document using WordNet lemmatizer

```
porter = nltk.PorterStemmer()
lancaster = nltk.LancasterStemmer()
wnl=nltk.WordNetLemmatizer()
from nltk.stem import SnowballStemmer
snowball_stemmer = SnowballStemmer('english')

SUBPstem = [porter.stem(t) for t in stoppedSUBwords]
print('Porter\n', SUBPstem[:200])

SUBLstem = [lancaster.stem(t) for t in stoppedSUBwords]
print('Lancaster\n', SUBLstem[:200])
SUBLemma = [wnl.lemmatize(t) for t in stoppedSUBwords]
print('WordNet Lemmatizer\n', SUBLemma[:200])

SUBSnstem = [snowball_stemmer.stem(t) for t in stoppedSUBwords]
print('Snowball Stemmer\n', SUBPstem[:200])
```

I found the same result affected in this document too by the Lancaster and snowball.

The WordNet lemmatizer only removes affixes if the resulting word is in its dictionary like lying remains same instead of changing to lie. So, I decided to use either of WordNet lemmatization or Porter stemming.

In summary of all the above steps:-

- Briefly state why you chose the processing options that you did.

Word-tokenizer-> Normalized text(lower case)->alpha filter->stopwords->Porter stemmer->WordNet Lemmatizer-> generated frequency distribution-> updated stopwords->generated frequency distribution

- **As you can see I did the same what I did for the task 2 . But there is one little change in the process for part2 Data**.

- Here I did tokenizer from NLTK and Lemmatizer to check the result of both of them.

- After seeing the results the NLTK tokenizer produced the results for bigrams are same as Porter Lemmatizer so I kept the results of NLTK tokenizer. But I used the results of Porter Lemmatizer results for PMI because it turns to give more meaningful data than the tokenizer does.

- So I used both NLTK tokenizer and Porter Lemmatizer.

4. Comparison (30%)

. **A) How are state_union_part1 and state_union_part2 similar or different in the use of the language, based on your results? Why?**

- From the Bigrams of state_union_part1 we can observe that the speakers are more oriented to the worldwide nations and all about their stuff. The friendly relations and about the Spanish, British and maxican government as we can see in the bigrams. (For Example:- ('mexican', 'government') , ('federal', 'government'), ('british', 'government'),etc.)
- In state_union_part1 we can also see talk about something about related time as ('september', 'last'), ('ensuing', 'year') , ('year', 'ending'), ('last', 'session'),etc. It's kind of redundancy but it indicates about the time period in their respective speeches.
- In state_union_part2 its all about what had happened in america and what can be done in america. By seeing the bigrams we can get the idea that it's very much different from the part1 document. It described about the health issues and other things which are major in USA.
- The Leaders talked about specific personalities like ('bin', 'laden'), ('saddam', 'hussein'), ('barack', 'obama'),etc.
- So In conclusion to this question, the part1 is all about worldwide nation talk and the part2 is about what's happening in United States of America in focus.

B) Are there any problems with the word or bigram lists that you found? Could you get a better list of bigrams?

- Some of the bigrams (raw frequency) are generally (did not render any useful information) associated to my document topic like (('***', 'state'), 0.00012904636528366005), (('taken', 'place'), 7.706935704440808e-05), (('ca', "n't"), 0.00023711666139740752),etc. **(Note :- Here the Asterisk used to divided the speech in the Doc so I didn't remove it).**
- In some I found some kind of redundancy in this bigram list like (('last', 'year'),('past', 'year'), ('present', 'year'), (('ensuing', 'year')
- I tried to update more stopwords and to get more meaningful data but this are the only efficient bigrams I got through the analysis.

C) How are the top 50 bigrams by frequency different from the top 50 bigrams scored by Mutual Information?

- The top 50 bigrams from both of the documents are quite meaningful and helpful to get the information about what the data actual is about in the document.
- In compare to bigrams the PMI are less meaningful in form of two words frequency. Most of them In document part2 are names of personalities in the real world. But the PMI words focused on the topic that were discussed in detail in the whole document.
- The same thing happen in Part1 for the PMI words as the words are divided in to two parts for example :- ('bona', 'fide'), (('andrew', 'jackson'), as person's name.
- In mutual Information most of the bigrams are meaningful but the only problem was there are lots of single words were divided into two words like ('costa', 'rica'), ('bona', 'fide'), ('per', 'cent'), etc.
- In conclusion Bigrams by Frequency are more meaningful than PMI bigrams. And PMI focus on the topics that are covered in The whole document than the biagrams.

Conclusion:-

- To sum-up State Union part1 is all about addressing people about worldwide nations, their relations with each other and about their government.
- State Union Part2 is all about what has happened in United states in past years and what can be happened in future. Most of the talk was about the personalities who affected the United States the most. Like Health issues and major factors that are affecting USA.

Appendix:-

Output:-

state_union_part1 : - (Task 2)

List the top 50 words by frequency (normalized by the length of the document):

Note- I divided the word count by total length of document into normalize word frequency distribution.

states	0.0048840464638607445
government	0.003978929596246184
united	0.0033408670123436434
may	0.002799589202403847
congress	0.002688465943409584
upon	0.002607811965107297
would	0.002475180978565757
public	0.002464427114792119
country	0.0020844572614568974
great	0.0019231493048523225
made	0.001901641577305046
state	0.0018729646072420103
last	0.0016327949829640874
war	0.0014947870645357288
present	0.001455356230699055
time	0.001448186988183296
people	0.0014087561543466222
year	0.0014069638437176823
power	0.0013334791079311539
citizens	0.0012958405847234196
subject	0.001274332857176143
without	0.0011883019469870362
union	0.0011524557344082418
act	0.0011237787643452062
treaty	0.001118401832458387
one	0.0011112325899426281
part	0.0011076479686847487
mexico	0.0010843479305085323
general	0.0010771786879927735
every	0.0010574632710744364
treasury	0.0010574632710744364
necessary	0.0010305786116403407
constitution	0.0009983170203194257
new	0.000982186224658968
duty	0.0009481323227091134
foreign	0.0009302092164197161

two	0.0009140784207592586
commerce	0.0009069091782434997
nations	0.0008997399357277408
peace	0.0008979476250988011
system	0.000885401450696223
laws	0.0008818168294383436
duties	0.0008746475869225847
within	0.0008585167912621272
law	0.0008549321700042477
interests	0.000808332093651815
interest	0.000795785919249237
amount	0.0007939936086202972
territory	0.0008214249479048786
important	0.0008136204353357111

List the top 50 bigrams by frequencies:

(('united', 'states'), 0.003265589965928175)
(('great', 'britain'), 0.000491093112329484)
(('last', 'session'), 0.0004337391722034129)
(('public', 'debt'), 0.00032082360258021037)
(('fiscal', 'year'), 0.0002580927305673201)
(('union', 'address'), 0.0002580927305673201)
(('public', 'lands'), 0.00023300038176216395)
(('two', 'countries'), 0.00021866189673064619)
(('present', 'year'), 0.00018998492666761062)
(('fellow', 'citizens'), 0.0001738541310071531)
(('general', 'government'), 0.00016668488849139423)
(('british', 'government'), 0.0001648925778624545)
(('two', 'governments'), 0.00015951564597563533)
(('federal', 'government'), 0.00015234640345987644)
(('annual', 'message'), 0.00014517716094411754)
(('public', 'service'), 0.00014338485031517782)
(('year', 'ending'), 0.00013980022905729837)
(('last', 'annual'), 0.00013442329717047922)
(('***', 'state'), 0.00012904636528366005)
(('public', 'money'), 0.00012366943339684087)
(('indian', 'tribes'), 0.0001182925015100217)
(('mexican', 'government'), 0.00011650019088108198)
(('treasury', 'notes'), 0.00011650019088108198)
(('commercial', 'intercourse'), 0.00011291556962320254)
(('several', 'states'), 0.0001021617058495642)

(('new', 'mexico'), 0.00010036939522062447)
(('favorable', 'consideration'), 9.857708459168476e-05)
(('naval', 'force'), 9.857708459168476e-05)
(('central', 'america'), 9.140784207592586e-05)
(('present', 'session'), 9.140784207592586e-05)
(('french', 'government'), 8.961553144698614e-05)
(('new', 'york'), 8.782322081804641e-05)
(('friendly', 'relations'), 8.603091018910669e-05)
(('existing', 'laws'), 8.065397830228753e-05)
(('good', 'faith'), 8.065397830228753e-05)
(('american', 'citizens'), 7.706935704440808e-05)
(('foreign', 'nations'), 7.706935704440808e-05)
(('taken', 'place'), 7.706935704440808e-05)
(('last', 'year'), 7.527704641546836e-05)
(('past', 'year'), 7.527704641546836e-05)
(('september', 'last'), 7.527704641546836e-05)
(('american', 'people'), 7.348473578652863e-05)
(('june', '30'), 7.348473578652863e-05)
(('military', 'force'), 7.348473578652863e-05)
(('ensuing', 'year'), 7.169242515758891e-05)
(('minister', 'plenipotentiary'), 7.169242515758891e-05)
(('slave', 'trade'), 6.990011452864918e-05)
(('spanish', 'government'), 6.990011452864918e-05)
(('charge', "d'affaires"), 6.631549327076975e-05)
(('present', 'fiscal'), 6.631549327076975e-05)

List the top 50 bigrams by their Mutual Information scores (using min frequency 5):

(('bona', 'fide'), 16.767819779008715)
(('posse', 'comitatus'), 16.767819779008715)
(('punta', 'arenas'), 16.767819779008715)
(('ballot', 'box'), 16.50478537317492)
(('del', 'norte'), 16.50478537317492)
(('millard', 'fillmore'), 16.50478537317492)
(('guadalupe', 'hidalgo'), 15.919822872453764)
(('porto', 'rico'), 15.919822872453764)
(('franklin', 'pierce'), 15.767819779008715)
(('la', 'plata'), 15.630316255258778)
(('vera', 'cruz'), 15.504785373174922)
(('entangling', 'alliances'), 15.43439604528352)
(('gun', 'boats'), 15.112467950396159)
(('costa', 'rica'), 15.089747873896076)
(('nucleus', 'around'), 15.089747873896076)
(('santa', 'anna'), 15.002285032645737)
(('santa', 'fe'), 15.002285032645737)
(('van', 'buren'), 15.002285032645737)

(('project', 'gutenberg'), 15.002285032645736)
(('sublime', 'porte'), 14.96046485695111)
(('martin', 'van'), 14.832360031203425)
(('ad', 'valorem'), 14.76781977900871)
(('quincy', 'adams'), 14.63031625525878)
(('water', 'witch'), 14.630316255258778)
(('statute', 'book'), 14.566185917839064)
(('buenos', 'ayres'), 14.50478537317492)
(('de', 'facto'), 14.356393533282247)
(('franking', 'privilege'), 14.334860371732606)
(('rocky', 'mountains'), 14.282392951838471)
(('andrew', 'jackson'), 14.199930791646498)
(('retired', 'list'), 14.144889428088536)
(('circulating', 'medium'), 14.045353754537622)
(('john', 'quincy'), 14.002285032645739)
(('precious', 'metals'), 13.94339134359217)
(('thomas', 'jefferson'), 13.914822191395396)
(('lake', 'erie'), 13.860929183400197)
(('almighty', 'god'), 13.832360031203425)
(('john', 'tyler'), 13.832360031203425)
(('san', 'francisco'), 13.80434565503383)
(('san', 'jacinto'), 13.804345655033828)
(('san', 'juan'), 13.804345655033828)
(('per', 'cent'), 13.732195869277994)
(('rio', 'grande'), 13.697430451117315)
(('inferior', 'quality'), 13.623773409392006)
(('grateful', 'acknowledgments'), 13.597894777566399)
(('hudsons', 'bay'), 13.535159022218439)
(('4.5', '%'), 13.41026777439063)
(('cumberland', 'road'), 13.373540839896668)
(('st.', 'marys'), 13.36182741933288)
(('st.', 'croix'), 13.361827419332878)

state_union_part2 : - (Task 3)

• list the top 50 words by frequency (normalized by the length of the document)

people	0.0035009577653381933
world	0.003463762994922914
new	0.00334985401052612
america	0.0029546595748637743
year	0.0029407115359580444
congress	0.00285934797567462
government	0.002582711870710978
years	0.002582711870710978
american	0.002208439493407227
nation	0.002001543582972234
one	0.0018690372133678005
every	0.0018132450577448811
make	0.001808595711442971
work	0.0017528035558200517
federal	0.001729556824310502
time	0.001722582804857637
states	0.0016528426103289877
americans	0.0015993751278570232
help	0.0015947257815551133
security	0.0015924011084041583
war	0.0015668297037436537
economic	0.0015598556842907887
peace	0.0015528816648379237
united	0.0015133622212716892
nations	0.0014994141823659593
program	0.0014831414703092745
country	0.0014645440851016346
national	0.0014157259489315803
economy	0.0013669078127615257
great	0.0013552844470067508
last	0.001329713042346246
many	0.0013087909839876513
free	0.0012971676182328765
need	0.0012878689256290566
first	0.0012855442524781016
let	0.0012762455598742818
state	0.0012088300384965874
tax	0.0011948819995908574
know	0.0011786092875341726
million	0.0011786092875341726
freedom	0.0011693105949303528
budget	0.001164661248628443
health	0.001136765170816983
n't	0.0011135184393074333
future	0.0011042197467036135

system	0.0010763236688921538
programs	0.0010739989957411987
tonight	0.001071674322590244
union	0.0010693496494392888
jobs	0.0010298302058730543

- list the top 50 bigrams by frequencies,

(('united', 'states'), 0.0010739989957411987)
(('american', 'people'), 0.0005555968830782392)
(('last', 'year'), 0.0005230514589648695)
(('fiscal', 'year'), 0.0004323892060776255)
(('federal', 'government'), 0.0004277398597757155)
(('social', 'security'), 0.0004207658403228506)
(('health', 'care'), 0.0004137918208699857)
(('years', 'ago'), 0.00037659705045470606)
(('union', 'address'), 0.00032080489483178663)
(('united', 'nations'), 0.0003138308753789217)
(('billion', 'dollars'), 0.00030220750962414684)
(('million', 'dollars'), 0.0002952334901712819)
(('soviet', 'union'), 0.00029058414386937196)
(('free', 'world'), 0.00025106470030313737)
(('ca', "n't"), 0.00023711666139740752)
(('every', 'american'), 0.00023014264194454 26)
(('economic', 'growth'), 0.00021851927618976772)
(('middle', 'east'), 0.00021154525673690278)
(('make', 'sure'), 0.00020457123728403787)
(('free', 'nations'), 0.00019592725446802 1797)
(('first', 'time'), 0.00018829852522735303)
(('four', 'years'), 0.00018829852522735303)
(('armed', 'forces'), 0.00017435048632162317)
(('world', 'war'), 0.0001720258131706682)
(('21st', 'century'), 0.00016970114001971324)
(('work', 'together'), 0.00016737646686875826)
(('foreign', 'policy'), 0.0001627271205668483)
(('mr.', 'speaker'), 0.0001627271205668483)
(('new', 'jobs'), 0.0001627271205668483)
(('two', 'years'), 0.00015575310111398338)
(('vice', 'president'), 0.00015575310111398338)
(('next', 'years'), 0.0001534284279630284)
(('national', 'security'), 0.0001441297353592085)
(('address', 'january'), 0.00013948038905729854)
(('human', 'rights'), 0.00013715571590634356)
(('health', 'insurance'), 0.00013483104275538858)
(('fellow', 'americans'), 0.00013018169645347865)
(('fellow', 'citizens'), 0.00013018169645347865)

(('past', 'year'), 0.00013018169645347865)
(('civil', 'rights'), 0.00012553235015156869)
(('young', 'people'), 0.00012553235015156869)
(('past', 'years'), 0.00012088300384965874)
(('private', 'sector'), 0.00012088300384965874)
(('god', 'bless'), 0.00011855833069870376)
(('local', 'governments'), 0.00011855833069870376)
(('nuclear', 'weapons'), 0.00011855833069870376)
(('interest', 'rates'), 0.00011390898439679381)
(('next', 'year'), 0.00011390898439679381)
(('balanced', 'budget'), 0.00011158431124583883)
(('high', 'school'), 0.00010925963809488386)

• list the top 50 bigrams by their Mutual Information scores (using min frequency 5)

Special Porter PMI words

(('el', 'salvador'), 15.161459969548133)
(('bin', 'laden'), 14.939067548211682)
(('saudi', 'arabia'), 14.939067548211682)
(('sam', 'rayburn'), 14.746422470269287)
(('gerald', 'r.'), 14.52403004893284)
(('jimmi', 'carter'), 14.28699085163199)
(('vol', 'p.'), 14.161459969548133)
(('northern', 'ireland'), 14.16145996954813)
(("o'neil", 'jr.'), 14.091070641656732)
(('r.', 'ford'), 14.064598430295543)
(('lyndon', 'b.'), 14.045982752128197)
(('william', 'j.'), 13.839531874660768)
(('thoma', 'jefferson'), 13.7829483462944)
(('red', 'tape'), 13.74642247026929)
(('iron', 'curtain'), 13.746422470269287)
(('200th', 'anniversari'), 13.702028350910837)
(('jill', 'biden'), 13.67603314237789)
(('b.', 'johnson'), 13.65895962901895)
(('barack', 'obama'), 13.658959629018947)
(('teen', 'pregnanc'), 13.52403004893284)
(('abraham', 'lincoln'), 13.489034627576636)
(('j.', 'clinton'), 13.424494375381924)
(('p.', "o'neil"), 13.31346306299318)
(('ronald', 'reagan'), 13.28699085163199)
(('mom', 'dad'), 13.260995643099045)
(('greec', 'turkey'), 13.202101954045478)
(('elementari', 'secondari'), 13.119983333571971)
(('endow', 'creator'), 13.091070641656732)
(('harri', 's.'), 13.083457457546857)
(('small-busi', 'owner'), 13.076571071961618)

(('old-ag', 'survivor'), 13.051035979854479)
(('dwight', 'd.'), 13.023956445798195)
(('intercontinent', 'ballist'), 13.000468092875828)
(('h.w', 'bush'), 12.991534968105821)
(('w.', 'bush'), 12.991534968105821)
(('ladi', 'gentlemen'), 12.939067548211685)
(('empower', 'zone'), 12.839531874660771)
(('nationwid', 'radio'), 12.79888989016342)
(('spoke', 'p.m.'), 12.79222615988241)
(('thoma', 'p.'), 12.7829483462944)
(('radio', 'televis'), 12.746422470269287)
(('statu', 'quo'), 12.702028350910833)
(('floor', 'appear'), 12.65895962901895)
(('f.', 'kennedi'), 12.658959629018945)
(('al', 'qaeda'), 12.617139453324322)
(('al', 'qaida'), 12.61713945332432)
(('richard', 'nixon'), 12.600065939965381)
(('georg', 'h.w'), 12.576497468826975)
(('georg', 'w.'), 12.576497468826975)
(('saddam', 'hussein'), 12.576497468826975)

. 3) Python processing screenshot (included in an appendix)

. Document 1:

```
Rishis-MacBook-Pro:Assignment_1 rishi$ python3 doc1.py
Tokenized and lowercase
['the', 'project', 'gutenberg', 'ebook', 'of', 'complete', 'state', 'of', 'the', 'union', 'addresses', ',', 'from', '1790', 'to', 'the', 'present', '(', '#', '41', 'in', 'our', 's
eries', 'of', 'us', 'presidential', 'state', 'of', 'the', 'union', 'addresses', ')', 'copyright', 'laws', 'are', 'changing', 'all', 'over', 'the', 'world', '.', 'be', 'sure', 'to'
, 'check', 'the', 'copyright', 'laws', 'for', 'your', 'country', 'before', 'downloading', 'or', 'redistributing', 'this', 'or', 'any', 'other', 'project', 'gutenberg', 'ebook', '.
', 'this', 'header', 'should', 'be', 'the', 'first', 'thing', 'seen', 'when', 'viewing', 'this', 'project', 'gutenberg', 'file', '.', 'please', 'do', 'not', 'remove', 'it', '.', '
do', 'not', 'change', 'or', 'edit', 'the', 'header', 'without', 'written', 'permission', '.', 'please', 'read', 'the', '``', 'legal', 'small', 'print', ',', "''", 'and', 'other',
'information', 'about', 'the', 'ebook', 'and', 'project', 'gutenberg', 'at', 'the', 'bottom', 'of', 'this', 'file', '.', 'included', 'is', 'important', 'information', 'about', 'yo
ur', 'specific', 'rights', 'and', 'restrictions', 'in', 'how', 'the', 'file', 'may', 'be', 'used', '.', 'you', 'can', 'also', 'find', 'out', 'about', 'how', 'to', 'make', 'a', 'do
nation', 'to', 'project', 'gutenberg', ',', 'and', 'how', 'to', 'get', 'involved', '.', '**welcome', 'to', 'the', 'world', 'of', 'free', 'plain', 'vanilla', 'electronic', 'texts**
', '**ebooks', 'readable', 'by', 'both', 'humans', 'and', 'by', 'computers', ',', 'since', '1971**', '*****these', 'ebooks', 'were', 'prepared', 'by', 'thousands', 'of', 'voluntee
rs', '!', '*****', 'title', ':', 'complete', 'state', 'of', 'the', 'union', 'addresses', 'from', '1790']
List of stopwords
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself
', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'
ll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if',
 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', '
from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', '
few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "
should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", '
haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "weren'
t", 'won', "won't", 'wouldn', "wouldn't", '!', '"', '#', '$', '%', '&', "'", '(', ')', '*', '+', ',', '-', '.', '/', ':', ';', '<', '=', '>', '?', '@', '[', '\\', ']', '^', '_', '
`', '{', '|', '}', '~']
Porter
 ['project', 'gutenberg', 'ebook', 'complet', 'state', 'union', 'address', 'present', 'seri', 'presidenti', 'state', 'union', 'address', 'copyright', 'law', 'chang', 'world', 'sur
e', 'check', 'copyright', 'law', 'countri', 'download', 'redistribut', 'project', 'gutenberg', 'ebook', 'header', 'first', 'thing', 'seen', 'view', 'project', 'gutenberg', 'file',
 'pleas', 'remov', 'chang', 'edit', 'header', 'without', 'written', 'permiss', 'pleas', 'read', 'legal', 'small', 'print', 'inform', 'ebook', 'project', 'gutenberg', 'bottom', 'fi
le', 'includ', 'import', 'inform', 'specif', 'right', 'restrict', 'file', 'may', 'use', 'find', 'make', 'donat', 'project', 'gutenberg', 'get', 'involv', '**welcom', 'world', 'fre
e', 'plain', 'vanilla', 'electron', 'texts**', '**ebook', 'readabl', 'human', 'comput', 'sinc', '*****these', 'ebook', 'prepar', 'thousand', 'volunt', 'titl', 'complet', 'state',
'union', 'address', 'author', 'variou', 'releas', 'date', 'februari', 'ebook', 'date', 'last', 'updat', 'june', 'edit', 'languag', 'english', 'start', 'project', 'gutenberg', 'ebo
ok', 'complet', 'address', 'ebook', 'produc', 'jame', 'linden', 'address', 'separ', 'three', 'asterisk', 'complet', 'state', 'union', 'address', 'content', 'georg', 'washington',
'state', 'union', 'address', 'januari', 'georg', 'washington', 'state', 'union', 'address', 'decemb', 'georg', 'washington', 'state', 'union', 'address', 'octob', 'georg', 'washin
gton', 'state', 'union', 'address', 'novemb', 'georg', 'washington', 'state', 'union', 'address', 'decemb', 'georg', 'washington', 'state', 'union', 'address', 'novemb', 'georg',
'washington', 'state', 'union', 'address', 'decemb', 'georg', 'washington', 'state', 'union', 'address', 'decemb', 'john', 'adam', 'state', 'union', 'address', 'novemb', 'john', '
adam', 'state', 'union', 'address', 'decemb', 'john', 'adam', 'state', 'union', 'address', 'decemb', 'john', 'adam', 'state', 'union', 'address', 'novemb', 'thoma', 'jefferson', '
state', 'union']
Lancaster
 ['project', 'gutenberg', 'ebook', 'complet', 'stat', 'un', 'address', 'pres', 'sery', 'presid', 'stat', 'un', 'address', 'copyright', 'law', 'chang', 'world', 'sur', 'check', 'co
pyright', 'law', 'country', 'download', 'redistribut', 'project', 'gutenberg', 'ebook', 'head', 'first', 'thing', 'seen', 'view', 'project', 'gutenberg', 'fil', 'pleas', 'remov',
'chang', 'edit', 'head', 'without', 'writ', 'permit', 'pleas', 'read', 'leg', 'smal', 'print', 'inform', 'ebook', 'project', 'gutenberg', 'bottom', 'fil', 'includ', 'import', 'inf
orm', 'spec', 'right', 'restrict', 'fil', 'may', 'us', 'find', 'mak', 'don', 'project', 'gutenberg', 'get', 'involv', '**welcome', 'world', 'fre', 'plain', 'vanill', 'electron', '
texts**', '**ebooks', 'read', 'hum', 'comput', 'sint', '*****these', 'ebook', 'prep', 'thousand', 'volunt', 'titl', 'complet', 'stat', 'un', 'address', 'auth', 'vary', 'releas', '
dat', 'febru', 'ebook', 'dat', 'last', 'upd', 'jun', 'edit', 'langu', 'engl', 'start', 'project', 'gutenberg', 'ebook', 'complet', 'address', 'ebook', 'produc', 'jam', 'lind', 'ad
dress', 'sep', 'three', 'asterisk', 'complet', 'stat', 'un', 'address', 'cont', 'georg', 'washington', 'stat', 'un', 'address', 'janu', 'georg', 'washington', 'stat', 'un', 'addre
ss', 'decemb', 'georg', 'washington', 'stat', 'un', 'address', 'octob', 'georg', 'washington', 'stat', 'un', 'address', 'novemb', 'georg', 'washington', 'stat', 'un', 'address', '
decemb', 'georg', 'washington', 'stat', 'un', 'address', 'novemb', 'georg', 'washington', 'stat', 'un', 'address', 'decemb', 'georg', 'washington', 'stat', 'un', 'address', 'decem
b', 'john', 'adam', 'stat', 'un', 'address', 'novemb', 'john', 'adam', 'stat', 'un', 'address', 'decemb', 'john', 'adam', 'stat', 'un', 'address', 'decemb', 'john', 'adam', 'stat'
, 'un', 'address', 'novemb', 'thoma', 'jefferson', 'stat', 'un']
```

WordNet Lemmatizer

['project', 'gutenberg', 'ebook', 'complete', 'state', 'union', 'address', 'present', 'series', 'presidential', 'state', 'union', 'address', 'copyright', 'law', 'changing', 'world', 'sure', 'check', 'copyright', 'law', 'country', 'downloading', 'redistributing', 'project', 'gutenberg', 'ebook', 'header', 'first', 'thing', 'seen', 'viewing', 'project', 'gutenberg', 'file', 'please', 'remove', 'change', 'edit', 'header', 'without', 'written', 'permission', 'please', 'read', 'legal', 'small', 'print', 'information', 'ebook', 'project', 'gutenberg', 'bottom', 'file', 'included', 'important', 'information', 'specific', 'right', 'restriction', 'file', 'may', 'used', 'find', 'make', 'donation', 'project', 'gutenberg', 'get', 'involved', '**welcome', 'world', 'free', 'plain', 'vanilla', 'electronic', 'texts**', '**ebooks', 'readable', 'human', 'computer', 'since', '*****these', 'ebooks', 'prepared', 'thousand', 'volunteer', 'title', 'complete', 'state', 'union', 'address', 'author', 'various', 'release', 'date', 'february', 'ebook', 'date', 'last', 'updated', 'june', 'edition', 'language', 'english', 'start', 'project', 'gutenberg', 'ebook', 'complete', 'address', 'ebook', 'produced', 'james', 'linden', 'address', 'separated', 'three', 'asterisk', 'complete', 'state', 'union', 'address', 'content', 'george', 'washington', 'state', 'union', 'address', 'january', 'george', 'washington', 'state', 'union', 'address', 'december', 'george', 'washington', 'state', 'union', 'address', 'october', 'george', 'washington', 'state', 'union', 'address', 'november', 'george', 'washington', 'state', 'union', 'address', 'december', 'george', 'washington', 'state', 'union', 'address', 'november', 'george', 'washington', 'state', 'union', 'address', 'december', 'george', 'washington', 'state', 'union', 'address', 'december', 'john', 'adam', 'state', 'union', 'address', 'november', 'john', 'adam', 'state', 'union', 'address', 'december', 'john', 'adam', 'state', 'union', 'address', 'december', 'john', 'adam', 'state', 'union', 'address', 'november', 'thomas', 'jefferson', 'state', 'union']

Snowball Stemmer

['project', 'gutenberg', 'ebook', 'complet', 'state', 'union', 'address', 'present', 'seri', 'presidenti', 'state', 'union', 'address', 'copyright', 'law', 'chang', 'world', 'sure', 'check', 'copyright', 'law', 'countri', 'download', 'redistribut', 'project', 'gutenberg', 'ebook', 'header', 'first', 'thing', 'seen', 'view', 'project', 'gutenberg', 'file', 'pleas', 'remov', 'chang', 'edit', 'header', 'without', 'written', 'permiss', 'pleas', 'read', 'legal', 'small', 'print', 'inform', 'ebook', 'project', 'gutenberg', 'bottom', 'file', 'includ', 'import', 'inform', 'specif', 'right', 'restrict', 'file', 'may', 'use', 'find', 'make', 'donat', 'project', 'gutenberg', 'get', 'involv', '**welcom', 'world', 'free', 'plain', 'vanilla', 'electron', 'texts**', '**ebook', 'readabl', 'human', 'comput', 'sinc', '*****these', 'ebook', 'prepar', 'thousand', 'volunt', 'titl', 'complet', 'state', 'union', 'address', 'author', 'various', 'releas', 'date', 'februari', 'ebook', 'date', 'last', 'updat', 'june', 'edit', 'languag', 'english', 'start', 'project', 'gutenberg', 'ebook', 'complet', 'address', 'ebook', 'produc', 'jame', 'linden', 'address', 'separ', 'three', 'asterisk', 'complet', 'state', 'union', 'address', 'content', 'georg', 'washington', 'state', 'union', 'address', 'januari', 'georg', 'washington', 'state', 'union', 'address', 'decemb', 'georg', 'washington', 'state', 'union', 'address', 'octob', 'georg', 'washington', 'state', 'union', 'address', 'novemb', 'georg', 'washington', 'state', 'union', 'address', 'decemb', 'georg', 'washington', 'state', 'union', 'address', 'novemb', 'georg', 'washington', 'state', 'union', 'address', 'decemb', 'georg', 'washington', 'state', 'union', 'address', 'decemb', 'john', 'adam', 'state', 'union', 'address', 'novemb', 'john', 'adam', 'state', 'union', 'address', 'decemb', 'john', 'adam', 'state', 'union', 'address', 'decemb', 'john', 'adam', 'state', 'union', 'address', 'novemb', 'thoma', 'jefferson', 'state', 'union']

tokenized, lowercase list without stopwords

['project', 'gutenberg', 'ebook', 'complete', 'state', 'union', 'addresses', 'present', 'series', 'presidential', 'state', 'union', 'addresses', 'copyright', 'laws', 'changing', 'world', 'sure', 'check', 'copyright', 'laws', 'country', 'downloading', 'redistributing', 'project', 'gutenberg', 'ebook', 'header', 'first', 'thing', 'seen', 'viewing', 'project', 'gutenberg', 'file', 'please', 'remove', 'change', 'edit', 'header', 'without', 'written', 'permission', 'please', 'read', 'legal', 'small', 'print', 'information', 'ebook', 'project', 'gutenberg', 'bottom', 'file', 'included', 'important', 'information', 'specific', 'rights', 'restrictions', 'file', 'may', 'used', 'find', 'make', 'donation', 'project', 'gutenberg', 'get', 'involved', '**welcome', 'world', 'free', 'plain', 'vanilla', 'electronic', 'texts**', '**ebooks', 'readable', 'humans', 'computers', 'since', '*****these', 'ebooks', 'prepared', 'thousands', 'volunteers', 'title', 'complete', 'state', 'union', 'addresses', 'author', 'various', 'release', 'date', 'february', 'ebook', 'date', 'last', 'updated', 'june', 'edition', 'language', 'english', 'start', 'project', 'gutenberg', 'ebook', 'complete', 'addresses', 'ebook', 'produced', 'james', 'linden', 'addresses', 'separated', 'three', 'asterisks', 'complete', 'state', 'union', 'addresses', 'contents', 'george', 'washington', 'state', 'union', 'address', 'january', 'george', 'washington', 'state', 'union', 'address', 'december', 'george', 'washington', 'state', 'union', 'address', 'october', 'george', 'washington', 'state', 'union', 'address', 'november', 'george', 'washington', 'state', 'union', 'address', 'december', 'george', 'washington', 'state', 'union', 'address', 'november', 'george', 'washington', 'state', 'union', 'address', 'december', 'george', 'washington', 'state', 'union', 'address', 'december', 'john', 'adams', 'state', 'union', 'address', 'november', 'john', 'adams', 'state', 'union', 'address', 'december', 'john', 'adams', 'state', 'union', 'address', 'december', 'john', 'adams', 'state', 'union', 'address', 'november', 'thomas', 'jefferson', 'state', 'union']

```
top 50 words by frequency
states   0.005316824187745354
government       0.004331504475887958
united   0.0036369028572320514
may      0.003047662158259906
congress         0.00292669221343781
public   0.0026828011956513257
country          0.0022691620294854483
great    0.00209356049667918
made     0.0020701469589716774
state    0.0020389289086950074
last     0.0017774777376278964
war      0.0016272408706714222
present          0.001584316051541001
time     0.0015765115389718336
people   0.0015335867198414124
year     0.0015316355916991204
power    0.0014516393378651536
citizens         0.0014106656468770244
subject          0.0013872521091695218
without          0.001293597958339512
union    0.0012545753954936744
act      0.0012233573452170045
treaty   0.0012175039607901289
one      0.0012096994482209613
part     0.0012057971919363776
mexico   0.0011804325260865833
general          0.0011726280135174157
every    0.001151165603952205
treasury         0.001151165603952205
necessary        0.001121898681817827
constitution     0.0010867783752565733
new      0.0010692182219759466
duty     0.001032146787272401
foreign          0.001012635505849482
two      0.0009950753525688554
commerce         0.0009872708399996878
nations          0.0009794663274305204
peace    0.0009775151992882285
system   0.0009638573022921854
laws     0.0009599550460076016
duties   0.0009521505334384341
within   0.0009345903801578073
law      0.0009306881238732235
interests        0.0008799587921736348
interest         0.0008663008951775916
amount   0.0008643497670352998
well     0.000844838485612381
territory        0.0008214249479048786
important        0.0008136204353357111
session          0.0008116693071934192
```

```
Sample Bigrams
[('the', 'project'), ('project', 'gutenberg'), ('gutenberg', 'ebook'), ('ebook', 'of'), ('of', 'complete'), ('complete', 'state'), ('state', 'of'), ('of', 'the'), ('the', 'union')
, ('union', 'addresses'), ('addresses', 'from'), ('from', 'to'), ('to', 'the'), ('the', 'present'), ('present', 'in'), ('in', 'our'), ('our', 'series'), ('series', 'of'), ('of', '
us'), ('us', 'presidential'), ('presidential', 'state'), ('state', 'of'), ('of', 'the'), ('the', 'union'), ('union', 'addresses'), ('addresses', 'copyright'), ('copyright', 'laws'
), ('laws', 'are'), ('are', 'changing'), ('changing', 'all'), ('all', 'over'), ('over', 'the'), ('the', 'world'), ('world', 'be'), ('be', 'sure'), ('sure', 'to'), ('to', 'check'),
 ('check', 'the'), ('the', 'copyright'), ('copyright', 'laws'), ('laws', 'for'), ('for', 'your'), ('your', 'country'), ('country', 'before'), ('before', 'downloading'), ('download
ing', 'or'), ('or', 'redistributing'), ('redistributing', 'this'), ('this', 'or'), ('or', 'any')]
without any filter
(('of', 'the'), 0.023036969976040145)
(('to', 'the'), 0.007751832109325612)
(('in', 'the'), 0.005986061140551467)
(('by', 'the'), 0.003980301410275422)
(('for', 'the'), 0.003664218651224138)
(('united', 'states'), 0.003554955475255793)
(('the', 'united'), 0.0035295908094059984)
(('and', 'the'), 0.0032896020479040980)
(('on', 'the'), 0.0032037524096432558)
(('of', 'our'), 0.0030769290880394284)
(('it', 'is'), 0.0028525493440307186)
(('to', 'be'), 0.0028505982158884267)
(('have', 'been'), 0.0026184139669556940)
(('with', 'the'), 0.0024954928939913057)
(('that', 'the'), 0.0024408613060071330)
(('has', 'been'), 0.0024096432557304632)
(('from', 'the'), 0.0021306319313827253)
(('of', 'a'), 0.0019023499387345764)
(('the', 'public'), 0.0017872333783393558)
(('will', 'be'), 0.0016740679460864270)
(('the', 'government'), 0.0016389476395251735)
(('at', 'the'), 0.0014984664132801585)
(('may', 'be'), 0.0013014024709086794)
(('of', 'congress'), 0.0012623799080620420)
(('and', 'to'), 0.0012233573452170045)
(('upon', 'the'), 0.0011160452973909515)
(('of', 'this'), 0.0011004362722526165)
(('of', 'their'), 0.0010965340159680328)
(('the', 'same'), 0.0010965340159680328)
(('the', 'present'), 0.0010848272471142815)
```

```
removed low frequency words
(('of', 'the'), 0.023036969976040145)
(('to', 'the'), 0.007751832109325612)
(('in', 'the'), 0.005986061140551467)
(('by', 'the'), 0.003980301410275422)
(('for', 'the'), 0.003664218651224138)
(('united', 'states'), 0.003554955475255793)
(('the', 'united'), 0.0035295908094059984)
(('and', 'the'), 0.003289602047904098)
(('on', 'the'), 0.0032037524096432558)
(('of', 'our'), 0.003076929080394284)
(('it', 'is'), 0.0028525493440307186)
(('to', 'be'), 0.0028505982158884267)
(('have', 'been'), 0.002618413966955694)
(('with', 'the'), 0.0024954928939913057)
(('that', 'the'), 0.002440861306007133)
(('has', 'been'), 0.0024096432557304632)
(('from', 'the'), 0.0021306319313827253)
(('of', 'a'), 0.0019023499387345764)
(('the', 'public'), 0.0017872333783393558)
(('will', 'be'), 0.001674067946086427)
(('the', 'government'), 0.0016389476395251735)
(('at', 'the'), 0.0014984664132801585)
(('may', 'be'), 0.0013014024709086794)
(('of', 'congress'), 0.001262379908062842)
(('and', 'to'), 0.0012233573452170045)
(('upon', 'the'), 0.0011160452973909515)
(('of', 'this'), 0.0011004362722526165)
(('of', 'their'), 0.0010965340159680328)
(('the', 'same'), 0.0010965340159680328)
(('the', 'present'), 0.0010848272471142815)
Bigrams after removing stopwords
(('united', 'states'), 0.003554955475255793)
(('great', 'britain'), 0.0005346091109879733)
(('last', 'session'), 0.0004721730104346333)
(('public', 'debt'), 0.0003492519374702453)
(('fiscal', 'year'), 0.0002809624524900297)
(('union', 'address'), 0.0002809624524900297)
(('public', 'lands'), 0.0002536466584979435)
(('two', 'countries'), 0.00023803763335960852)
(('present', 'year'), 0.00020681958308293855)
(('fellow', 'citizens'), 0.00018925942980231117)
(('general', 'government'), 0.00018145491723314420)
(('year', 'ending'), 0.00018145491723314420)
(('british', 'government'), 0.00017950378909085233)
(('two', 'governments'), 0.00017365040466397672)
(('federal', 'government'), 0.00016584589209480920)
(('annual', 'message'), 0.00015804137952564172)
(('public', 'service'), 0.00015609025138334985)
(('last', 'annual'), 0.00014633461067189050)
(('public', 'money'), 0.00013462784181813924)
(('indian', 'tribes'), 0.00012877445739126363)
(('mexican', 'government'), 0.00012682332924897176)
(('treasury', 'notes'), 0.00012682332924897176)
(('commercial', 'intercourse'), 0.00012292107296438801)
(('several', 'states'), 0.00011121430411063677)
```

```
Bigrams after removing stopwords
(('united', 'states'), 0.003554955475255793)
(('great', 'britain'), 0.0005346091109879733)
(('last', 'session'), 0.0004721730104346333)
(('public', 'debt'), 0.0003492519374702453)
(('fiscal', 'year'), 0.0002809624524900297)
(('union', 'address'), 0.0002809624524900297)
(('public', 'lands'), 0.0002536466584979435)
(('two', 'countries'), 0.00023803763335960852)
(('present', 'year'), 0.00020681958308293855)
(('fellow', 'citizens'), 0.0001892594298023117)
(('general', 'government'), 0.0001814549172331442)
(('year', 'ending'), 0.0001814549172331442)
(('british', 'government'), 0.00017950378909085233)
(('two', 'governments'), 0.00017365040466397672)
(('federal', 'government'), 0.0001658458920948092)
(('annual', 'message'), 0.00015804137952564172)
(('public', 'service'), 0.00015609025138334985)
(('last', 'annual'), 0.0001463346106718905)
(('public', 'money'), 0.00013462784181813924)
(('indian', 'tribes'), 0.00012877445739126363)
(('mexican', 'government'), 0.00012682332924897176)
(('treasury', 'notes'), 0.00012682332924897176)
(('commercial', 'intercourse'), 0.00012292107296438801)
(('several', 'states'), 0.00011121430411063677)
(('address', 'december'), 0.0001092631759683449)
(('new', 'mexico'), 0.0001092631759683449)
(('favorable', 'consideration'), 0.00010731204782605302)
(('naval', 'force'), 0.00010731204782605302)
(('central', 'america'), 9.950753525688553e-05)
(('present', 'session'), 9.950753525688553e-05)
(('french', 'government'), 9.755640711459366e-05)
(('new', 'york'), 9.560527897230178e-05)
(('friendly', 'relations'), 9.365415083000991e-05)
(('existing', 'laws'), 8.78007664031343e-05)
(('good', 'faith'), 8.78007664031343e-05)
(('american', 'citizens'), 8.389851011855055e-05)
(('foreign', 'nations'), 8.389851011855055e-05)
(('taken', 'place'), 8.389851011855055e-05)
(('last', 'year'), 8.194738197625867e-05)
(('past', 'year'), 8.194738197625867e-05)
(('september', 'last'), 8.194738197625867e-05)
(('american', 'people'), 7.99962538339668e-05)
(('military', 'force'), 7.99962538339668e-05)
(('ensuing', 'year'), 7.804512569167493e-05)
(('minister', 'plenipotentiary'), 7.804512569167493e-05)
(('slave', 'trade'), 7.609399754938305e-05)
(('spanish', 'government'), 7.609399754938305e-05)
(('charge', "d'affaires"), 7.219174126479931e-05)
(('present', 'fiscal'), 7.219174126479931e-05)
(('two', 'nations'), 7.219174126479931e-05)
```

```
pmi data with minimum frequency
(('bona', 'fide'), 16.645331942938817)
(('posse', 'comitatus'), 16.645331942938817)
(('punta', 'arenas'), 16.645331942938817)
(('ballot', 'box'), 16.38229753710502)
(('del', 'norte'), 16.38229753710502)
(('millard', 'fillmore'), 16.38229753710502)
(('guadalupe', 'hidalgo'), 15.797335036383867)
(('porto', 'rico'), 15.797335036383867)
(('franklin', 'pierce'), 15.645331942938817)
(('la', 'plata'), 15.50782841918888)
(('vera', 'cruz'), 15.382297537105025)
(('entangling', 'alliances'), 15.311908209213623)
(('gun', 'boats'), 14.989980114326261)
(('nucleus', 'around'), 14.967260037826179)
(('costa', 'rica'), 14.967260037826176)
(('santa', 'anna'), 14.87979719657584)
(('santa', 'fe'), 14.87979719657584)
(('van', 'buren'), 14.87979719657584)
(('project', 'gutenberg'), 14.879797196575838)
(('sublime', 'porte'), 14.837977020881212)
(('martin', 'van'), 14.709872195133528)
(('ad', 'valorem'), 14.645331942938812)
(('quincy', 'adams'), 14.507828419188883)
(('beacons', 'buoys'), 14.50782841918888)
(('water', 'witch'), 14.50782841918888)
(('statute', 'book'), 14.443698081769167)
(('buenos', 'ayres'), 14.382297537105023)
(('de', 'facto'), 14.23390569721235)
(('franking', 'privilege'), 14.212372535662709)
(('rocky', 'mountains'), 14.159905115768574)
(('andrew', 'jackson'), 14.0774429555766)
(('retired', 'list'), 14.022401592018639)
(('circulating', 'medium'), 13.922865918467725)
(('john', 'quincy'), 13.879797196575842)
(('th', 'jefferson'), 13.87979719657584)
(('precious', 'metals'), 13.820903507522273)
(('thomas', 'jefferson'), 13.7923343553255)
(('lake', 'erie'), 13.7384413473303)
(('almighty', 'god'), 13.709872195133528)
(('john', 'tyler'), 13.709872195133528)
(('san', 'francisco'), 13.681857818963932)
(('san', 'jacinto'), 13.68185781896393)
(('san', 'juan'), 13.68185781896393)
(('per', 'cent'), 13.609708033208097)
(('rio', 'grande'), 13.574942615047418)
(('inferior', 'quality'), 13.501285573322109)
(('grateful', 'acknowledgments'), 13.475406941496502)
(('hudsons', 'bay'), 13.412671186148541)
(('cumberland', 'road'), 13.25105300382677)
(('st.', 'marys'), 13.239339583262982)
```

Document 2:-

```
Rishis-MacBook-Pro:Assignment_1 rishi$ python3 doc2.py
Tokenized and lowercase
['the', 'project', 'gutenberg', 'ebook', 'of', 'complete', 'state', 'of', 'the', 'union', 'addresses', ',', 'from', '1946', 'to', 'the', 'present', '(', '#', '41', 'in', 'our', 's
eries', 'of', 'us', 'presidential', 'state', 'of', 'the', 'union', 'addresses', ')', 'copyright', 'laws', 'are', 'changing', 'all', 'over', 'the', 'world', '.', 'be', 'sure', 'to'
, 'check', 'the', 'copyright', 'laws', 'for', 'your', 'country', 'before', 'downloading', 'or', 'redistributing', 'this', 'or', 'any', 'other', 'project', 'gutenberg', 'ebook', '.
', 'this', 'header', 'should', 'be', 'the', 'first', 'thing', 'seen', 'when', 'viewing', 'this', 'project', 'gutenberg', 'file', '.', 'please', 'do', 'not', 'remove', 'it', '.', '
do', 'not', 'change', 'or', 'edit', 'the', 'header', 'without', 'written', 'permission', '.', 'please', 'read', 'the', '*``', 'legal', 'small', 'print', ',', "'", "'", 'and', 'other',
'information', 'about', 'the', 'ebook', 'and', 'project', 'gutenberg', 'at', 'the', 'bottom', 'of', 'this', 'file', '.', 'included', 'is', 'important', 'information', 'about', 'yo
ur', 'specific', 'rights', 'and', 'restrictions', 'in', 'how', 'the', 'file', 'may', 'be', 'used', '.', 'you', 'can', 'also', 'find', 'out', 'about', 'how', 'to', 'make', 'a', 'do
nation', 'to', 'project', 'gutenberg', ',', 'and', 'how', 'to', 'get', 'involved', '.', '**welcome', 'to', 'the', 'world', 'of', 'free', 'plain', 'vanilla', 'electronic', 'texts**
', '**ebooks', 'readable', 'by', 'both', 'humans', 'and', 'by', 'computers', ',', 'since', '1971**', '*****these', 'ebooks', 'were', 'prepared', 'by', 'thousands', 'of', 'voluntee
rs', '!', '*****', 'title', ':', 'complete', 'state', 'of', 'the', 'union', 'addresses', 'from', '1946']
List of stopwords
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself
', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'
ll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if',
 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', '
from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', '
few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "
should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", '
haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "weren'
t", 'won', "won't", 'wouldn', "wouldn't", '!', '"', '#', '$', '%', '&', "'", '(', ')', '*', '+', ',', '-', '.', '/', ':', ';', '<', '=', '>', '?', '@', '[', '\\', ']', '^', '_', '
`', '{', '|', '}', '~']
tokenized, lowercase list without stopwords
['project', 'gutenberg', 'ebook', 'complete', 'state', 'union', 'addresses', 'present', 'series', 'presidential', 'state', 'union', 'addresses', 'copyright', 'laws', 'changing', '
world', 'sure', 'check', 'copyright', 'laws', 'country', 'downloading', 'redistributing', 'project', 'gutenberg', 'ebook', 'header', 'first', 'thing', 'seen', 'viewing', 'project'
, 'gutenberg', 'file', 'please', 'remove', 'change', 'edit', 'header', 'without', 'written', 'permission', 'please', 'read', 'legal', 'small', 'print', 'information', 'ebook', 'pr
oject', 'gutenberg', 'bottom', 'file', 'included', 'important', 'information', 'specific', 'rights', 'restrictions', 'file', 'may', 'used', 'find', 'make', 'donation', 'project',
'gutenberg', 'get', 'involved', '**welcome', 'world', 'free', 'plain', 'vanilla', 'electronic', 'texts**', '**ebooks', 'readable', 'humans', 'computers', 'since', '*****these', 'e
books', 'prepared', 'thousands', 'volunteers', 'title', 'complete', 'state', 'union', 'addresses', 'present', 'author', 'various', 'edition', 'language', 'english', 'start', 'proj
ect', 'gutenberg', 'ebook', 'complete', 'addresses', 'ebook', 'produced', 'james', 'linden', 'addresses', 'separated', 'three', 'asterisks', 'complete', 'state', 'union', 'address
es', 'contents', 'harry', 's.', 'truman', 'state', 'union', 'address', 'january', 'harry', 's.', 'truman', 'state', 'union', 'address', 'january', 'harry', 's.', 'truman', 'state
', 'union', 'address', 'january', 'harry', 's.', 'truman', 'state', 'union', 'address', 'january', 'harry', 's.', 'truman', 'state', 'union', 'address', 'january', 'harry', 's.', '
truman', 'state', 'union', 'address', 'january', 'harry', 's.', 'truman', 'state', 'union', 'address', 'january', 'harry', 's.', 'truman', 'state', 'union', 'address', 'january',
'dwight', 'd.', 'eisenhower', 'state', 'union', 'address', 'february', 'dwight', 'd.', 'eisenhower', 'state', 'union', 'address', 'january', 'dwight', 'd.', 'eisenhower', 'state',
 'union', 'address', 'january', 'dwight', 'd.', 'eisenhower', 'state', 'union', 'address']
Porter
 ['project', 'gutenberg', 'ebook', 'complet', 'state', 'union', 'address', 'present', 'seri', 'presidenti', 'state', 'union', 'address', 'copyright', 'law', 'chang', 'world', 'sur
e', 'check', 'copyright', 'law', 'countri', 'download', 'redistribut', 'project', 'gutenberg', 'ebook', 'header', 'first', 'thing', 'seen', 'view', 'project', 'gutenberg', 'file',
 'pleas', 'remov', 'chang', 'edit', 'header', 'without', 'written', 'permiss', 'pleas', 'read', 'legal', 'small', 'print', 'inform', 'ebook', 'project', 'gutenberg', 'bottom', 'fi
le', 'includ', 'import', 'inform', 'specif', 'right', 'restrict', 'file', 'may', 'use', 'find', 'make', 'donat', 'project', 'gutenberg', 'get', 'involv', '**welcom', 'world', 'fre
e', 'plain', 'vanilla', 'electron', 'texts**', '**ebook', 'readabl', 'human', 'comput', 'sinc', '*****these', 'ebook', 'prepar', 'thousand', 'volunt', 'titl', 'complet', 'state',
'union', 'address', 'present', 'author', 'variou', 'edit', 'languag', 'english', 'start', 'project', 'gutenberg', 'ebook', 'complet', 'address', 'ebook', 'produc', 'jame', 'linden
', 'address', 'separ', 'three', 'asterisk', 'complet', 'state', 'union', 'address', 'content', 'harri', 's.', 'truman', 'state', 'union', 'address', 'januari', 'harri', 's.', 'tru
man', 'state', 'union', 'address', 'januari', 'harri', 's.', 'truman', 'state', 'union', 'address', 'januari', 'harri', 's.', 'truman', 'state', 'union', 'address', 'januari', 'ha
rri', 's.', 'truman', 'state', 'union', 'address', 'januari', 'harri', 's.', 'truman', 'state', 'union', 'address', 'januari', 'harri', 's.', 'truman', 'state', 'union', 'address'
, 'januari', 'harri', 's.', 'truman', 'state', 'union', 'address', 'januari', 'dwight', 'd.', 'eisenhow', 'state', 'union', 'address', 'februari', 'dwight', 'd.', 'eisenhow', 'sta
te', 'union', 'address', 'januari', 'dwight', 'd.', 'eisenhow', 'state', 'union', 'address', 'januari', 'dwight', 'd.', 'eisenhow', 'state', 'union', 'address']
```

```
top 50 words by frequency
people   0.0035009577653381933
world    0.003463762994922914
new      0.00334985401052612
america          0.0029546595748637743
year     0.0029407115359580444
congress         0.00285934797567462
government       0.002582711870710978
years    0.002582711870710978
american         0.002208439493407227
nation   0.0020015435829722 34
one      0.0018690372133678005
every    0.0018132450577448811
make     0.001808595711442971
work     0.0017528035558200517
federal          0.001729556824310502
time     0.001722582804857637
states   0.0016528426103289877
americans        0.0015993751278570232
help     0.0015947257815551133
security         0.0015924011084041583
war      0.0015668297037436537
economic         0.0015598556842907887
peace    0.0015528816648379237
united   0.0015133622212716892
nations          0.0014994141823659593
program          0.0014831414703092745
country          0.0014645440851016346
national         0.0014157259489315803
economy          0.0013669078127615257
great    0.0013552844470067508
last     0.001329713042346246
many     0.0013087909839876513
free     0.0012971676182328765
need     0.0012878689256290566
first    0.0012855442524781016
let      0.0012762455598742818
state    0.0012088300384965874
tax      0.0011948819995908574
know     0.0011786092875341726
million          0.0011786092875341726
freedom          0.0011693105949303528
budget   0.001164661248628443
health   0.001136765170816983
n't      0.0011135184393074333
future   0.0011042197467036135
system   0.0010763236688921538
programs         0.0010739989957411987
tonight          0.001071674322590244
union    0.0010693496494392888
jobs     0.0010298302058730543
```

```
Sample Bigrams
[('project', 'gutenberg'), ('gutenberg', 'ebook'), ('ebook', 'complete'), ('complete', 'state'), ('state', 'union'), ('union', 'addresses'), ('addresses', 'present'), ('present',
'series'), ('series', 'presidential'), ('presidential', 'state'), ('state', 'union'), ('union', 'addresses'), ('addresses', 'copyright'), ('copyright', 'laws'), ('laws', 'changing
'), ('changing', 'world'), ('world', 'sure'), ('sure', 'check'), ('check', 'copyright'), ('copyright', 'laws'), ('laws', 'country'), ('country', 'downloading'), ('downloading', 'r
edistributing'), ('redistributing', 'project'), ('project', 'gutenberg'), ('gutenberg', 'ebook'), ('ebook', 'header'), ('header', 'first'), ('first', 'thing'), ('thing', 'seen'),
('seen', 'viewing'), ('viewing', 'project'), ('project', 'gutenberg'), ('gutenberg', 'file'), ('file', 'please'), ('please', 'remove'), ('remove', 'change'), ('change', 'edit'), (
'edit', 'header'), ('header', 'without'), ('without', 'written'), ('written', 'permission'), ('permission', 'please'), ('please', 'read'), ('read', 'legal'), ('legal', 'small'), (
'small', 'print'), ('print', 'information'), ('information', 'ebook'), ('ebook', 'project')]
without any filter
(('of', 'the'), 0.0075598370869055815)
(('in', 'the'), 0.005793085492179799)
(('to', 'the'), 0.0032847631622993805)
(('of', 'our'), 0.0031336594074873074)
(('and', 'the'), 0.0025129716761823286)
(('for', 'the'), 0.0023688419408231204)
(('we', 'have'), 0.00229445239992561)
(('we', 'must'), 0.0021479979914823975)
(('the', 'world'), 0.0019364527347454949)
(('the', 'congress'), 0.0015273102601774191)
(('will', 'be'), 0.0014691934314035448)
(('we', 'are'), 0.0014362202674304)
(('it', 'is'), 0.0014250246415354)
(('we', 'can'), 0.0014180506220825351)
(('on', 'the'), 0.0014157259489315803)
(('the', 'united'), 0.0013390011734950066)
(('we', 'will'), 0.0013157650034405163)
(('with', 'the'), 0.0013064663108366963)
(('and', 'to'), 0.0012599728478175968)
(('in', 'our'), 0.001257648174666642)
(('by', 'the'), 0.0011925573264399026)
(('that', 'the'), 0.0011669859217793978)
(('that', 'we'), 0.0011251418050622082)
(('united', 'states'), 0.0010739989957411987)
(('and', 'we'), 0.001053076937382604)
(('in', 'this'), 0.0010484275910806942)
(('to', 'be'), 0.0010391288984768741)
(('more', 'than'), 0.0009717133770991798)
(('in', 'a'), 0.000955440665042495)
(('of', 'a'), 0.0009531159918915401)
(('and', 'i'), 0.0009484666455896301)
(('has', 'been'), 0.0009414926261367652)
(('is', 'the'), 0.0008973238362686207)
(('have', 'been'), 0.0008833757973628908)
(('of', 'this'), 0.0008694277584571609)
(('the', 'american'), 0.0008531550464004761)
(('a', 'new'), 0.0008275836417399714)
(('i', 'have'), 0.0007880641981737367)
(('must', 'be'), 0.0007624927935132 32)
(('is', 'a'), 0.0007601681203622771)
(('to', 'make'), 0.0007578434472113221)
(('from', 'the'), 0.0007508694277584572)
(('as', 'a'), 0.0007345967157017724)
(('to', 'help'), 0.0007345967157017724)
(('state', 'of'), 0.0007299473693998624)
(('at', 'the'), 0.0007090253110412676)
(('the', 'people'), 0.0006997266184374476)
```

```
removed low frequency words
(('of', 'the'), 0.0075598370869055815)
(('in', 'the'), 0.005793085492179799)
(('to', 'the'), 0.0032847631622993805)
(('of', 'our'), 0.0031336594074873074)
(('and', 'the'), 0.0025129716761823286)
(('for', 'the'), 0.0023688419408231204)
(('we', 'have'), 0.002294452399992561)
(('we', 'must'), 0.0021479979914823975)
(('the', 'world'), 0.0019364527347454949)
(('the', 'congress'), 0.0015273102601774191)
(('will', 'be'), 0.0014691934314035448)
(('we', 'are'), 0.00144362202674304)
(('it', 'is'), 0.0014250246415354)
(('we', 'can'), 0.0014180506220825351)
(('on', 'the'), 0.0014157259489315803)
(('the', 'united'), 0.001339011734950066)
(('we', 'will'), 0.0013157650034405163)
(('with', 'the'), 0.0013064663108366963)
(('and', 'to'), 0.0012599728478175968)
(('in', 'our'), 0.001257648174666642)
(('by', 'the'), 0.0011925573264399026)
(('that', 'the'), 0.0011669859217793978)
(('that', 'we'), 0.0011251418050622082)
(('united', 'states'), 0.0010739989957411987)
(('and', 'we'), 0.001053076937382604)
(('in', 'this'), 0.0010484275910806942)
(('to', 'be'), 0.0010391288984768741)
(('more', 'than'), 0.0009717133770991798)
(('in', 'a'), 0.000955440665042495)
(('of', 'a'), 0.0009531159918915401)
```

```
Bigrams after removing stopwords
(('united', 'states'), 0.0010739989957411987)
(('american', 'people'), 0.0005555968830782392)
(('last', 'year'), 0.0005230514589648695)
(('fiscal', 'year'), 0.0004323892060776255)
(('federal', 'government'), 0.0004277398597757155)
(('social', 'security'), 0.0004207658403228506)
(('health', 'care'), 0.0004137918208699857)
(('years', 'ago'), 0.0003765970505470606)
(('union', 'address'), 0.00032080489483178663)
(('united', 'nations'), 0.0003138308753789217)
(('billion', 'dollars'), 0.00030220750962414684)
(('million', 'dollars'), 0.0002952334901712819)
(('soviet', 'union'), 0.00029058414386937196)
(('free', 'world'), 0.00025106470030313737)
(('ca', "n't"), 0.00023711666139740752)
(('every', 'american'), 0.00023014264194454 26)
(('economic', 'growth'), 0.00021851927618976772)
(('middle', 'east'), 0.00021154525673690278)
(('make', 'sure'), 0.00020457123728403787)
(('free', 'nations'), 0.00019527254468021797)
(('first', 'time'), 0.00018829852522735303)
(('four', 'years'), 0.00018829852522735303)
(('armed', 'forces'), 0.00017435048632162317)
(('world', 'war'), 0.0001720258131706682)
(('21st', 'century'), 0.00016970114001971324)
(('work', 'together'), 0.00016737646686875826)
(('foreign', 'policy'), 0.0001627271205668483)
(('mr.', 'speaker'), 0.0001627271205668483)
(('new', 'jobs'), 0.0001627271205668483)
(('two', 'years'), 0.00015575310111398338)
(('vice', 'president'), 0.00015575310111398338)
(('next', 'years'), 0.0001534284279630284)
(('national', 'security'), 0.0001441297353592085)
(('address', 'january'), 0.00013948038905729854)
(('human', 'rights'), 0.00013715571590634356)
(('health', 'insurance'), 0.00013483104275538858)
(('fellow', 'americans'), 0.00013018169645347865)
(('fellow', 'citizens'), 0.00013018169645347865)
(('past', 'year'), 0.00013018169645347865)
(('civil', 'rights'), 0.00012553235015156869)
(('young', 'people'), 0.00012553235015156869)
(('past', 'years'), 0.00012088300384965874)
(('private', 'sector'), 0.00012088300384965874)
(('god', 'bless'), 0.00011855833069870376)
(('local', 'governments'), 0.00011855833069870376)
(('nuclear', 'weapons'), 0.00011855833069870376)
(('interest', 'rates'), 0.00011390898439679381)
(('next', 'year'), 0.00011390898439679381)
(('balanced', 'budget'), 0.00011158431124583883)
(('high', 'school'), 0.00010925963809488386)
```

```
pmi on filtered data
(('el', 'salvador'), 16.12957818104658)
(('bin', 'laden'), 15.907185759710131)
(('saudi', 'arabia'), 15.907185759710128)
(('sam', 'rayburn'), 15.714540681767733)
(('gerald', 'r.'), 15.492148260431286)
(('jimmy', 'carter'), 15.392612586880372)
(('vol', 'p.'), 15.255109063130437)
(('northern', 'ireland'), 15.129578181046579)
(("o'neill", 'jr.'), 15.059188853155177)
(('r.', 'ford'), 15.03271664179399)
(('lyndon', 'b.'), 15.014100963626642)
(('floor', 'appears'), 14.97757508760153)
(('iron', 'curtain'), 14.907185759710131)
(('grass', 'roots'), 14.866543775212783)
(('200th', 'anniversary'), 14.807650086159217)
(('william', 'j.'), 14.807650086159214)
(('thomas', 'jefferson'), 14.751066557792846)
(('red', 'tape'), 14.714540681767735)
(('jill', 'biden'), 14.644151353876335)
(('b.', 'johnson'), 14.627077840517396)
(('barack', 'obama'), 14.627077840517392)
(('teen', 'pregnancy'), 14.544615680325421)
(('abraham', 'lincoln'), 14.457152839075082)
(('p.', "o'neill"), 14.407112156575486)
(('j.', 'clinton'), 14.392612586880037)
(('empowerment', 'zones'), 14.322223258988974)
(('ronald', 'reagan'), 14.255109063130435)
(('synthetic', 'fuels'), 14.240609493435322)
(('small-business', 'owner'), 14.229113854597491)
(('harry', 's.'), 14.051575669045302)
(('dwight', 'd.'), 13.99207465729664)
(('intercontinental', 'ballistic'), 13.968586304374274)
(('h.w', 'bush'), 13.959653179604267)
(('w.', 'bush'), 13.959653179604267)
(('small-business', 'owners'), 13.907185759710131)
(('thomas', 'p.'), 13.876597439876706)
(('river', 'basins'), 13.856559686640162)
(('status', 'quo'), 13.85655968664016)
(('prime', 'minister'), 13.807650086159214)
(('nationwide', 'radio'), 13.767008101661869)
(('project', 'gutenberg'), 13.627077840517396)
(('f.', 'kennedy'), 13.62707784051739)
(('al', 'qaeda'), 13.585257664822768)
(('al', 'qaida'), 13.585257664822766)
(('richard', 'nixon'), 13.568184151463827)
(('george', 'h.w'), 13.544615680325421)
(('george', 'w.'), 13.544615680325421)
(('saddam', 'hussein'), 13.544615680325421)
(('d.', 'eisenhower'), 13.532643038659344)
(('line-item', 'veto'), 13.505087316138782)
```

```
Special Porter frequency words
(('unit', 'state'), 0.0021010505252626313)
(('last', 'year'), 0.0012460775842466688)
(('state', 'union'), 0.0012324343990176908)
(('american', 'peopl'), 0.0011005502751375688)
(('fiscal', 'year'), 0.0008822593114739187)
(('year', 'ago'), 0.0008686161262449406)
(('feder', 'govern'), 0.0008458774841966438)
(('social', 'secur'), 0.0008276865705580062)
(('health', 'care'), 0.0008094956569193688)
(('billion', 'dollar'), 0.0006958024466778844)
(('union', 'address'), 0.0006457774341716313)
(('unit', 'nation'), 0.000618491063713675)
(('million', 'dollar'), 0.0006003001500750375)
(('soviet', 'union'), 0.0005684660512074219)
(('next', 'year'), 0.0005593705943881032)
(('work', 'togeth'), 0.0005184410387011688)
(('men', 'women'), 0.0005138933102915094)
(('state', 'local'), 0.0005138933102915094)
(('past', 'year'), 0.00050934558188185)
(('free', 'world'), 0.0005002501250625312)
(('make', 'sure'), 0.00047751148301423437)
(('ca', "n't"), 0.0004638682977852563)
(('everi', 'american'), 0.00045022511255627816)
(('ask', 'congress'), 0.00043203419891764064)
(('econom', 'growth'), 0.00042748647050798124)
(('million', 'american'), 0.0004229387420983219)
(('middl', 'east'), 0.0004138432852790031)
(('free', 'nation'), 0.0004047478284596844)
(('four', 'year'), 0.0003820091864113875)
(('member', 'congress'), 0.0003729137295920688)
(('first', 'time'), 0.0003683660011824094)
(('small', 'busi'), 0.0003683660011824094)
(('nation', 'secur'), 0.0003592705443630906 5)
(('world', 'war'), 0.0003592705443630906 5)
(('arm', 'forc'), 0.00035017508754377186)
(('balanc', 'budget'), 0.0003456273591341125)
(('tax', 'cut'), 0.0003456273591341125)
(('foreign', 'polici'), 0.0003365319023147938)
(('new', 'job'), 0.0003365319023147938)
(('21st', 'centuri'), 0.0003319841739051344)
(('mr.', 'speaker'), 0.00031834098867615626)
(('two', 'year'), 0.00030469780344717813)
(('vice', 'presid'), 0.00030469780344717813)
(('local', 'govern'), 0.00030015007503751874)
(('around', 'world'), 0.0002910546182182)
(('address', 'januari'), 0.00027286370457956253)
(('human', 'right'), 0.00027286370457956253)
(('health', 'insur'), 0.00026376824776024374)
(('nation', 'world'), 0.00026376824776024374)
(('civil', 'right'), 0.0002592205193505844)
```

```
Special Porter PMI words
(('el', 'salvador'), 15.161459969548133)
(('bin', 'laden'), 14.939067548211682)
(('saudi', 'arabia'), 14.939067548211682)
(('sam', 'rayburn'), 14.746422470269287)
(('gerald', 'r.'), 14.52403004893284)
(('jimmi', 'carter'), 14.28699085163199)
(('vol', 'p.'), 14.161459969548133)
(('northern', 'ireland'), 14.16145996954813)
(("o'neil", 'jr.'), 14.091070641656732)
(('r.', 'ford'), 14.064598430295543)
(('lyndon', 'b.'), 14.045982752128197)
(('william', 'j.'), 13.839531874660768)
(('thoma', 'jefferson'), 13.7829483462944)
(('red', 'tape'), 13.74642247026929)
(('iron', 'curtain'), 13.746422470269287)
(('200th', 'anniversari'), 13.702028350910837)
(('jill', 'biden'), 13.67603314237789)
(('b.', 'johnson'), 13.65895962901895)
(('barack', 'obama'), 13.658959629018947)
(('teen', 'pregnanc'), 13.52403004893284)
(('abraham', 'lincoln'), 13.489034627576636)
(('j.', 'clinton'), 13.424494375381924)
(('p.', "o'neil"), 13.31346306299318)
(('ronald', 'reagan'), 13.28699085163199)
(('mom', 'dad'), 13.260995643099045)
(('greec', 'turkey'), 13.202101954045478)
(('elementari', 'secondari'), 13.119983333571971)
(('endow', 'creator'), 13.091070641656732)
(('harri', 's.'), 13.083457457546857)
(('small-busi', 'owner'), 13.076571071961618)
(('old-ag', 'survivor'), 13.051035979854479)
(('dwight', 'd.'), 13.023956445798195)
(('intercontinent', 'ballist'), 13.000468092875828)
(('h.w', 'bush'), 12.991534968105821)
(('w.', 'bush'), 12.991534968105821)
(('ladi', 'gentlemen'), 12.939067548211685)
(('empower', 'zone'), 12.839531874660771)
(('nationwid', 'radio'), 12.79888989016342)
(('spoke', 'p.m.'), 12.79222615988241)
(('thoma', 'p.'), 12.7829483462944)
(('radio', 'televis'), 12.746422470269287)
(('statu', 'quo'), 12.702028350910833)
(('floor', 'appear'), 12.65895962901895)
(('f.', 'kennedi'), 12.658959629018945)
(('al', 'qaeda'), 12.617139453324322)
(('al', 'qaida'), 12.61713945332432)
(('richard', 'nixon'), 12.600065939965381)
(('georg', 'h.w'), 12.576497468826975)
(('georg', 'w.'), 12.576497468826975)
(('saddam', 'hussein'), 12.576497468826975)
```