IST 664 Natural Language Processing Homework 1

Harper He
xhe128@syr.edu

Dataset description

The State of the Union Addresses dataset is a collection of annual speeches delivered by the presidents of the United States to a joint session of the United States Congress for the span of 1790-2016. This dataset contains the two texts combined, which are 1) Complete State of the Union Addresses from 1790 to 1860, 2) Complete State of the Union Addresses from 1946 to the Present.

These datasets are small subsets of the Project Gutenberg Ebook corpus. Complete State of the Union Addresses from 1790 to 1860 was release in February, 2004 and was last updated on June 24 2007 while the other dataset has no exact release date.

These documents in the corpus are named in the format as "President, State of the Union Address, Date.txt", for example, "George Washington, State of the Union Address, January 8, 1790. txt". So that we can get information about when and who delivered these addresses as well as the format of the file is text.

Both the addresses delivered between 1790 and 1860 and addresses delivered between 1946 and 2016 have 70 addresses.

There is also a policy description of the Project Gutenberg Ebook called 'state_union_policy'. This document provides terms of use and redistributing Project Gutenberg-tm electronic works.

Regarding the related policies, since these texts are the Project Gutenberg EBook, they are for the use of anyone anywhere in the United States and most other parts of the world at no cost and with almost no restrictions whatsoever. We can copy and distribute it in the United States without permission and without paying copyright royalties. This and all associated files of various formats can be found in: http://www.gutenberg.org/5/0/9/5/50950/.

Analysis 1: Analysis of State of the Union Addresses dataset: Part1

The following tasks are analyzed with "state_union_part1.txt" file.

To decide how to process the words, i.e. decide on tokenization and whether to use

all lower case, use or modify the stop word list, or lemmatization, the analyses were performed based on 4 circumstances to find out the best processing methods. The 4 circumstances were:

- Lowercase + Punctuation Removed
- Lowercase + Punctuation Removed + Stop Words Removed
- Lowercase + Punctuation Removed + Stop Words Removed + Lemmatization
- Lowercase + Punctuation Removed + NLTK & Self-defined Stop Words Removed + Lemmatization

A) list the top 50 words by frequency

The following steps were performed via python coding to list the top 50 words by frequency.

a) For Lowercase + Punctuation Removed circumstance:

Step 1: Load the state union_part1.txt using NLTK package

Step 2: Perform tokenization to separate words from the text file

Step 3: Convert tokenized words to lower case and remove punctuation marks

Step 4: Calculate the frequency distribution of the tokenized words

Step 5: Find the top 50 words by frequency based using most_common function

b) For Lowercase + Punctuation Removed + Stop Words Removed circumstance:

Step1-3 are the same as circumstance a)

Step 4: Get a list of stop words from NLTK and remove these stop words from text

Step 5-6 are the same as circumstance a) step 4-5

c) For Lowercase + Punctuation Removed + Stop Words Removed + Lemmatization circumstance:

Step1-4 are the same as circumstance b)

Step 5: Use NLTK lemmatizer to obtain the root words

Step 6-7 are the same as circumstance a) step 4-5

d) For Lowercase + Punctuation Removed + NLTK & Self-defined Stop Words Removed + Lemmatization circumstance:

Step1-4 are the same as circumstance b)

Step 5: Define a list of stop words based on previous outputs and remove both NLTK and self-defined stop words from text

Step 6-8 are the same as circumstance c) step 5-7

B) list the top 50 bigrams by frequencies

The following steps were performed via python coding to list the top 50 bigrams by frequencies. The 4 circumstances were also applied in this analysis and would be shown in the output.

Step 1: import collocation finder package from NLTK to calculate bigram measure

Step 2: Apply collocation finder to the initial list of tokenized words from step 3 of previous task (task A)

Step 3: Use score_ngrams to score the bigrams by frequency and print the top 50 bigrams.

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

The following steps were performed via python coding to list the top 50 bigrams by frequencies. The 4 circumstances were also applied in this analysis and would be shown in the output.

Step 1: import collocation finder package from NLTK to calculate bigram measure

Step 2: Use apply_freq_filter to identify the bigrams with min frequency as 5

Step 3: Apply filtered finder to the initial list of tokenized words from step 3 of previous task (task A)

Step 4: Print the top 50 bigrams with their pmi scores

Output of Analysis 1

The following table is the output of the three tasks of analysis 1

Text 1 (Complete State of the Union Addresses from 1790 to 1860)		
Lowercase + Punctuation Removed		
Top 50 words by frequency	Top 50 bigrams by frequencies	Top 50 bigrams by PMI
('the', 49523),	(('of', 'the'), 0.023139252704378124),	(('bona', 'fide'), 16.63979571898761),
('of', 32404),	(('to', 'the'), 0.007783594908345917),	(('posse', 'comitatus'), 16.63979571898761),
('to', 19129),	(('in', 'the'), 0.006090076290590155),	(('punta', 'arenas'), 16.63979571898761),
('and', 16905),	(('by', 'the'), 0.00399560483468185),	(('ballot', 'box'), 16.376761313153814),
('in', 10464),	(('for', 'the'), 0.0036783068036924087),	(('del', 'norte'), 16.376761313153814),
('a', 7839),	(('united', 'states'), 0.0035686235337207503),	(('millard', 'fillmore'), 16.376761313153814),
('that', 6658),	(('the', 'united'), 0.003543161346048758),	(('guadalupe', 'hidalgo'), 15.79179881243266),
('be', 6354),	(('and', 'the'), 0.0033061671377171385),	(('porto', 'rico'), 15.79179881243266),
('it', 5492),	(('on', 'the'), 0.0032160701659547045),	(('franklin', 'pierce'), 15.63979571898761),
('by', 5451),	(('of', 'our'), 0.0030887592275947438),	(('la', 'plata'), 15.502292195237674),
('which', 5243),	(('it', 'is'), 0.0028635167981886592),	(('vera', 'cruz'), 15.376761313153814),
('for', 4746),	(('to', 'be'), 0.0028615581683677366),	(('entangling', 'alliances'), 15.306371985262416),
('our', 4318),	(('have', 'been'), 0.0026284812196779622),	(('gun', 'boats'), 14.984443890375054),
('with', 4144),	(('with', 'the'), 0.002505087540959846),	(('costa', 'rica'), 14.961723813874972),
('is', 4000),	(('that', 'the'), 0.002450245905974017),	(('nucleus', 'around'), 14.961723813874972),
('have', 3786),	(('has', 'been'), 0.0024189078288392573),	(('santa', 'anna'), 14.874260972624633),
('as', 3734),	(('from', 'the'), 0.0021388237644473433),	(('santa', 'fe'), 14.874260972624633),
('been', 3692),	(('of', 'a'), 0.0019096640753994136),	(('van', 'buren'), 14.874260972624633),
('on', 3370),	(('the', 'public'), 0.00179606354578591),	(('project', 'gutenberg'), 14.874260972624631),
('this', 3151),	(('will', 'be'), 0.001680504386351484),	(('sublime', 'porte'), 14.832440796930005),
('has', 2788),	(('the', 'government'), 0.0016456490495748795),	(('martin', 'van'), 14.70433597118232),
('their', 2775),	(('at', 'the'), 0.0015042277024684611),	(('ad', 'valorem'), 14.639795718987608),
('from', 2758),	(('may', 'be'), 0.001306406090555291),	(('beacons', 'buoys'), 14.502292195237674),
('not', 2750),	(('of', 'congress'), 0.0012672334941368417),	(('water', 'witch'), 14.502292195237674),
('states', 2725),	(('and', 'to'), 0.0012300195275393147),	(('quincy', 'adams'), 14.502292195237672),
('will', 2546),	(('upon', 'the'), 0.001120336257567656),	(('statute', 'book'), 14.43816185781796),
('was', 2231),	(('of', 'this'), 0.0011046672190002762),	(('buenos', 'ayres'), 14.376761313153816),
('government', 2220),	(('of', 'their'), 0.0011007499593584313),	(('de', 'facto'), 14.228369473261143),
('at', 2198),	(('the', 'same'), 0.0011007499593584313),	(('franking', 'privilege'), 14.206836311711502),
('its', 2041),	(('the', 'present'), 0.0010889981804328964),	(('rocky', 'mountains'), 14.15436889181737),
('all', 1949),	(('the', 'people'), 0.0010086943577750748),	(('andrew', 'jackson'), 14.071906731625393),
('i', 1947),	(('between', 'the'), 0.0009949839490286175),	(('retired', 'list'), 14.016865368067432),
('an', 1940),	(('all', 'the'), 0.0009832321701030828),	(('circulating', 'medium'), 13.917329694516518),
('are', 1895),	(('in', 'a'), 0.0009812735402821601),	(('th', 'jefferson'), 13.874260972624633),
('united', 1864),	(('state', 'of'), 0.0009773562806403152),	(('john', 'quincy'), 13.874260972624631),
('or', 1683),	(('under', 'the'), 0.0009695217613566254),	(('precious', 'metals'), 13.815367283571065),
('they', 1575),	(('the', 'country'), 0.000959728612252013),	(('thomas', 'jefferson'), 13.786798131374292),
('may', 1562),	(('which', 'the'), 0.00095581352610168),	(('lake', 'erie'), 13.732905123379092),
('congress', 1500),	(('of', 'that'), 0.0009538527227892456),	(('almighty', 'god'), 13.70433597118232),
('them', 1475),	(('the', 'union'), 0.0009479768333264781),	(('john', 'tyler'), 13.70433597118232),
('but', 1462),	(('should', 'be'), 0.0009342664245800208),	(('san', 'jacinto'), 13.676321595012723),
('upon', 1455),	(('it', 'was'), 0.0009283905351172534),	(('san', 'juan'), 13.676321595012723),
('would', 1381),	(('the', 'treasury'), 0.0009264319052963309),	(('san', 'francisco'), 13.676321595012721),
('public', 1375),	(('would', 'be'), 0.0008539626019221994),	(('per', 'cent'), 13.60417180925689),
('such', 1340),	(('the', 'last'), 0.0008480867124594319),	(('rio', 'grande'), 13.569406391096214),
('other', 1269),	(('part', 'of'), 0.0008382935633548195),	(('inferior', 'quality'), 13.495749349370902),
('these', 1208),	(('the', 'constitution'), 0.000836334933533897),	(('grateful', 'acknowledgments'), 13.469870717545295),
('any', 1171),	(('can', 'not'), 0.0008147900055037498),	(('hudsons', 'bay'), 13.407134962197334),
('country', 1163),	(('the', 'most'), 0.0008128313756828274),	(('cumberland', 'road'), 13.245516779875564),
('should', 1159)	(('and', 'of'), 0.0007971623371154476)	(('cut', 'off'), 13.180364100350312)

Text 1 (Complete State of the Union Addresses from 1790 to 1860)		
Lowercase + Punctuation Removed + Stop Words Removed		
Top 50 words by frequency	Top 50 bigrams by frequencies	Top 50 bigrams by PMI
('states', 2725),	(('united', 'states'), 0.007648905979748451),	(('bona', 'fide'), 15.539910019165042),
('government', 2220),	(('great', 'britain'), 0.0011502745545834663),	(('posse', 'comitatus'), 15.539910019165042),
('united', 1864),	(('last', 'session'), 0.0010159359204715286),	(('punta', 'arenas'), 15.539910019165042),
('may', 1562),	(('public', 'debt'), 0.0007514567345636513),	(('ballot', 'box'), 15.276875613331246),
('congress', 1500),	(('state', 'union'), 0.0007262682406676631),	(('del', 'norte'), 15.276875613331246),
('upon', 1455),	(('house', 'representatives'), 0.0006213161827677117),	(('millard', 'fillmore'), 15.276875613331246),
('would', 1381),	(('fiscal', 'year'), 0.0006045238535037195),	(('clayton', 'bulwer'), 14.861838114052404),
('public', 1375),	(('union', 'address'), 0.0006045238535037195),	(('guadalupe', 'hidalgo'), 14.691913112610091),
('country', 1163),	(('report', 'secretary'), 0.0005835334419237293),	(('porto', 'rico'), 14.691913112610091),
('great', 1073),	(('public', 'lands'), 0.0005457507010797468),	(('writ', 'mandamus'), 14.598803708218608),
('made', 1061),	(('two', 'countries'), 0.0005121660425517623),	(('franklin', 'pierce'), 14.539910019165042),
('state', 1045),	(('present', 'year'), 0.0004449967254957935),	(('la', 'plata'), 14.402406495415105),
('last', 911),	(('within', 'limits'), 0.00041980823159980523),	(('vera', 'cruz'), 14.276875613331246),
('war', 834),	(('secretary', 'treasury'), 0.00041561014928380717),	(('entangling', 'alliances'), 14.206486285439848),
('present', 812),	(('fellow', 'citizens'), 0.00040721398465181106),	(('seminaries', 'learning'), 14.013841207497453),
('time', 808),	(('session', 'congress'), 0.00040721398465181106),	(('gun', 'boats'), 13.884558190552486),
('people', 786),	(('act', 'congress'), 0.0003946197377038169),	(('nucleus', 'around'), 13.861838114052404),
('year', 785),	(('general', 'government'), 0.00039042165538781884),	(('ruler', 'universe'), 13.861838114052404),
('power', 744),	(('year', 'ending'), 0.0003904216553878188 4),	(('costa', 'rica'), 13.8618381140524),
('citizens', 723),	(('british', 'government'), 0.0003862235730718208),	(('santa', 'anna'), 13.774375272802065),
('subject', 711),	(('two', 'governments'), 0.0003736293261238266),	(('santa', 'fe'), 13.774375272802065),
('shall', 694),	(('citizens', 'united'), 0.0003610350791758325),	(('van', 'buren'), 13.774375272802065),
('without', 663),	(('federal', 'government'), 0.00035683699685983445),	(('project', 'gutenberg'), 13.774375272802063),
('union', 643),	(('secretary', 'war'), 0.0003526389145438364),	(('sublime', 'porte'), 13.732555097107436),
('act', 627),	(('annual', 'message'), 0.0003400446675958422),	(('tea', 'coffee'), 13.613910600608818),
('treaty', 624),	(('public', 'service'), 0.00033584658527984417),	(('martin', 'van'), 13.604450271359752),
('one', 620),	(('senate', 'house'), 0.00033584658527984417),	(('ad', 'valorem'), 13.53991001916504),
('part', 618),	(('consideration', 'congress'), 0.00032325233833185),	(('beacons', 'buoys'), 13.402406495415105),
('mexico', 605),	(('ending', 'june'), 0.0003148561736998539),	(('water', 'witch'), 13.402406495415105),
('general', 601),	(('last', 'annual'), 0.0003148561736998539),	(('quincy', 'adams'), 13.402406495415104),
('every', 590),	(('attention', 'congress'), 0.00031065809138385584),	(('statute', 'book'), 13.338276157995391),
('treasury', 590),	(('government', 'united'), 0.0003064600090678578),	(('buenos', 'ayres'), 13.276875613331244),
('necessary', 575),	(('public', 'money'), 0.0002896676798038656),	(('indiana', 'illinois'), 13.139372089581311),
('constitution', 557),	(('indian', 'tribes'), 0.00027707343285587145),	(('de', 'facto'), 13.128483773438575),
('new', 548),	(('mexican', 'government'), 0.0002728753505398734),	(('franking', 'privilege'), 13.106950611888934),
('duty', 529),	(('part', 'united'), 0.0002728753505398734),	(('rocky', 'mountains'), 13.054483191994798),
('foreign', 519),	(('treasury', 'notes'), 0.0002728753505398734),	(('andrew', 'jackson'), 12.972021031802825),
('two', 510),	(('upon', 'subject'), 0.00026867726822387534),	(('retired', 'list'), 12.916979668244863),
('commerce', 506),	(('commercial', 'intercourse'), 0.0002644791859078773),	(('sooner', 'later'), 12.876945006442611),
('nations', 502),	(('several', 'states'), 0.0002644791859078773),	(('circulating', 'medium'), 12.81744399469395),
('peace', 501),	(('secretary', 'state'), 0.0002602811035918792),	(('intent', 'meaning'), 12.798828316526604),
('system', 494),	(('provision', 'made'), 0.00024768685664388506),	(('th', 'jefferson'), 12.774375272802065),
('laws', 492),	(('article', 'treaty'), 0.00023929069201188898),	(('john', 'quincy'), 12.774375272802063),
('duties', 488),	(('claims', 'citizens'), 0.00023929069201188898),	(('precious', 'metals'), 12.715481583748494),
('within', 479),	(('address', 'december'), 0.00023509260969589092),	(('thomas', 'jefferson'), 12.686912431551724),
('law', 477),	(('new', 'mexico'), 0.00023509260969589092),	(('lake', 'erie'), 12.633019423556524),
('us', 463),	(('favorable', 'consideration'), 0.00023089452737989286),	(('almighty', 'god'), 12.604450271359752),
('interests', 451),	(('naval', 'force'), 0.0002308945273798 9286),	(('john', 'tyler'), 12.604450271359752),
('interest', 444),	(('bank', 'united'), 0.0002266964450638948),	(('san', 'jacinto'), 12.576435895190155),
('amount', 443)	(('people', 'united'), 0.0002266964450638948)	(('san', 'juan'), 12.576435895190155)

Text 1 (Complete State of the Union Addresses from 1790 to 1860)		
Lowercase + Punctuation Removed + Stop Words Removed + Lemmatization		
Top 50 words by frequency	Top 50 bigrams by frequencies	Top 50 bigrams by PMI
('state', 3770),	(('united', 'states'), 0.007648905979748451),	(('bona', 'fide'), 15.539910019165042),
('government', 2561),	(('great', 'britain'), 0.0011502745545834663),	(('del', 'norte'), 15.276875613331246),
('united', 1864),	(('last', 'session'), 0.0010159359204715286),	(('millard', 'fillmore'), 15.276875613331246),
('may', 1562),	(('public', 'debt'), 0.0007514567345636513),	(('punta', 'arena'), 15.276875613331246),
('congress', 1501),	(('state', 'union'), 0.0007262682406676631),	(('ballot', 'box'), 15.054483191994798),
('upon', 1455),	(('house', 'representatives'), 0.0006213161827677117),	(('clayton', 'bulwer'), 14.861838114052404),
('country', 1427),	(('fiscal', 'year'), 0.0006045238535037195),	(('guadalupe', 'hidalgo'), 14.691913112610091),
('would', 1381),	(('union', 'address'), 0.0006045238535037195),	(('porto', 'rico'), 14.691913112610091),
('public', 1375),	(('report', 'secretary'), 0.0005835334419237293),	(('writ', 'mandamus'), 14.598803708218608),
('power', 1159),	(('public', 'lands'), 0.0005457507010797468),	(('franklin', 'pierce'), 14.539910019165042),
('year', 1145),	(('two', 'countries'), 0.0005121660425517623),	(('la', 'plata'), 14.402406495415105),
('great', 1073),	(('present', 'year'), 0.0004449967254957935),	(('vera', 'cruz'), 14.276875613331246),
('made', 1061),	(('within', 'limits'), 0.00041980823159980523),	(('entangling', 'alliance'), 14.054483191994798),
('duty', 1017),	(('secretary', 'treasury'), 0.00041561014928380717),	(('seminary', 'learning'), 14.0398364160304),
('law', 969),	(('fellow', 'citizens'), 0.00040721398465181106),	(('nucleus', 'around'), 13.861838114052404),
('time', 914),	(('session', 'congress'), 0.00040721398465181106),	(('costa', 'rica'), 13.8618381140524),
('last', 911),	(('act', 'congress'), 0.0003946197377038169),	(('santa', 'anna'), 13.774375272802065),
('war', 898),	(('general', 'government'), 0.00039042165538781884),	(('santa', 'fe'), 13.774375272802065),
('interest', 895),	(('year', 'ending'), 0.0003904216553878188 4),	(('van', 'buren'), 13.774375272802065),
('subject', 888),	(('british', 'government'), 0.0003862235730718208),	(('sublime', 'porte'), 13.732555097107436),
('present', 852),	(('two', 'governments'), 0.0003736293261238266),	(('tea', 'coffee'), 13.613910600608818),
('nation', 830),	(('citizens', 'united'), 0.0003610350791758325),	(('martin', 'van'), 13.604450271359752),
('act', 807),	(('federal', 'government'), 0.00035683699685983445),	(('ad', 'valorem'), 13.53991001916504),
('citizen', 801),	(('secretary', 'war'), 0.0003526389145438364),	(('quincy', 'adam'), 13.402406495415104),
('people', 786),	(('annual', 'message'), 0.0003400446675958422),	(('buenos', 'ayres'), 13.276875613331244),
('treaty', 748),	(('public', 'service'), 0.00033584658527984417),	(('beacon', 'buoy'), 13.264902971665169),
('part', 741),	(('senate', 'house'), 0.00033584658527984417),	(('ruler', 'universe'), 13.198873101329973),
('shall', 694),	(('consideration', 'congress'), 0.00032325233833185),	(('indiana', 'illinois'), 13.139372089581311),
('without', 663),	(('ending', 'june'), 0.0003148561736998539),	(('de', 'facto'), 13.128483773438575),
('union', 643),	(('last', 'annual'), 0.0003148561736998539),	(('project', 'gutenberg'), 13.106950611888934),
('right', 636),	(('attention', 'congress'), 0.00031065809138385584),	(('gun', 'boat'), 12.991473394468999),
('one', 634),	(('government', 'united'), 0.0003064600090678578),	(('andrew', 'jackson'), 12.972021031802825),
('general', 605),	(('public', 'money'), 0.0002896676798038656),	(('sooner', 'later'), 12.876945006442611),
('mexico', 605),	(('indian', 'tribes'), 0.00027707343285587145),	(('retired', 'list'), 12.852849330825148),
('treasury', 592),	(('mexican', 'government'), 0.0002728753505398734),	(('circulating', 'medium'), 12.81744399469395),
('every', 590),	(('part', 'united'), 0.0002728753505398734),	(('rocky', 'mountain'), 12.81744399469395),
('necessary', 585),	(('treasury', 'notes'), 0.0002728753505398734),	(('intent', 'meaning'), 12.798828316526604),
('constitution', 570),	(('upon', 'subject'), 0.00026867726822387534),	(('th', 'jefferson'), 12.774375272802065),
('territory', 562),	(('commercial', 'intercourse'), 0.0002644791859078773),	(('john', 'quincy'), 12.774375272802063),
('new', 548),	(('several', 'states'), 0.0002644791859078773),	(('thomas', 'jefferson'), 12.686912431551724),
('object', 532),	(('secretary', 'state'), 0.0002602811035918792),	(('precious', 'metal'), 12.616833995815583),
('foreign', 519),	(('provision', 'made'), 0.00024768685664388506),	(('almighty', 'god'), 12.604450271359752),
('measure', 512),	(('article', 'treaty'), 0.00023929069201188898),	(('john', 'tyler'), 12.604450271359752),
('two', 510),	(('claims', 'citizens'), 0.00023929069201188898),	(('san', 'jacinto'), 12.576435895190155),
('system', 509),	(('address', 'december'), 0.00023509260969589092),	(('san', 'juan'), 12.576435895190155),
('commerce', 506),	(('new', 'mexico'), 0.00023509260969589092),	(('san', 'francisco'), 12.576435895190153),
('peace', 501),	(('favorable', 'consideration'), 0.00023089452737989286),	(('seizure', 'confiscation'), 12.539910019165044),
('consideration', 493),	(('naval', 'force'), 0.00023089452737989286),	(('rio', 'grande'), 12.435573359350302),
('within', 479),	(('bank', 'united'), 0.00022669644506389 48),	(('effusion', 'blood'), 12.38935034258966),
('service', 479)	(('people', 'united'), 0.00022669644506389 48)	(('inferior', 'quality'), 12.091009068019913)

Text 1 (Complete State of the Union Addresses from 1790 to 1860)		
Lowercase + Punctuation Removed + NLTK & Self-defined Stop Words Removed + Lemmatization		
Top 50 words by frequency	Top 50 bigrams by frequencies	Top 50 bigrams by PMI
('state', 3770),	(('united', 'state'), 0.007768223152117406),	(('bona', 'fide'), 15.518370391206467),
('government', 2561),	(('great', 'britain'), 0.0011675771495777121),	(('del', 'norte'), 15.255335985372675),
('united', 1864),	(('last', 'session'), 0.0010397402353903714),	(('millard', 'fillmore'), 15.255335985372675),
('may', 1562),	(('state', 'union'), 0.0008991196297842965),	(('punta', 'arena'), 15.255335985372671),
('congress', 1501),	(('public', 'debt'), 0.0007712827155969558),	(('ballot', 'box'), 15.032943564036223),
('country', 1427),	(('house', 'representative'), 0.0006306621099908809),	(('clayton', 'bulwer'), 14.840298486093833),
('public', 1375),	(('union', 'address'), 0.0006306621099908809),	(('guadalupe', 'hidalgo'), 14.670373484651517),
('power', 1159),	(('report', 'secretary'), 0.0006264008795179696),	(('porto', 'rico'), 14.670373484651517),
('year', 1145),	(('fiscal', 'year'), 0.0006178784185721469),	(('writ', 'mandamus'), 14.577264080260033),
('great', 1073),	(('public', 'land'), 0.0005710048833701219),	(('franklin', 'pierce'), 14.51837039120647),
('made', 1061),	(('act', 'congress'), 0.0005198701176951857),	(('la', 'plata'), 14.380866867456534),
('duty', 1017),	(('two', 'country'), 0.0005198701176951857),	(('vera', 'cruz'), 14.255335985372675),
('law', 969),	(('present', 'year'), 0.0004559516606015153),	(('entangling', 'alliance'), 14.032943564036227),
('time', 914),	(('within', 'limit'), 0.0004474291996556926),	(('seminary', 'learning'), 14.018296788071826),
('last', 911),	(('public', 'money'), 0.0004431679691827812 4),	(('costa', 'rica'), 13.840298486093829),
('war', 898),	(('session', 'congress'), 0.0004389067387098699),	(('nucleus', 'around'), 13.840298486093829),
('interest', 895),	(('secretary', 'treasury'), 0.0004218618168182244 3),	(('santa', 'anna'), 13.75283564484349),
('subject', 888),	(('fellow', 'citizen'), 0.0004133393558724017 3),	(('santa', 'fe'), 13.75283564484349),
('present', 852),	(('general', 'government'), 0.0004090781253994904),	(('van', 'buren'), 13.75283564484349),
('nation', 830),	(('british', 'government'), 0.0003962944339807563),	(('sublime', 'porte'), 13.711015469148865),
('act', 807),	(('secretary', 'war'), 0.0003962944339807563),	(('tea', 'coffee'), 13.592370972650247),
('citizen', 801),	(('year', 'ending'), 0.0003962944339807563),	(('martin', 'van'), 13.582910643401181),
('people', 786),	(('citizen', 'united'), 0.0003835107425620222),	(('ad', 'valorem'), 13.518370391206469),
('treaty', 748),	(('two', 'government'), 0.0003792495120891109),	(('quincy', 'adam'), 13.380866867456533),
('part', 741),	(('federal', 'government'), 0.0003707270511432881),	(('buenos', 'ayres'), 13.255335985372673),
('without', 663),	(('annual', 'message'), 0.0003622045901974654),	(('beacon', 'buoy'), 13.243363343706594),
('union', 643),	(('public', 'service'), 0.0003408984378329086),	(('ruler', 'universe'), 13.177333473371398),
('right', 636),	(('senate', 'house'), 0.0003408984378329086),	(('indiana', 'illinois'), 13.117832461622736),
('one', 634),	(('consideration', 'congress'), 0.00033663720735999727),	(('de', 'facto'), 13.10694414548),
('general', 605),	(('ending', 'june'), 0.00031959228546835186),	(('project', 'gutenberg'), 13.085410983930363),
('mexico', 605),	(('last', 'annual'), 0.00031959228546835186),	(('gun', 'boat'), 12.969933766510428),
('treasury', 592),	(('attention', 'congress'), 0.00031533105499544046),	(('andrew', 'jackson'), 12.95048140384425),
('every', 590),	(('government', 'united'), 0.0003110698245225291),	(('sooner', 'later'), 12.85540537848404),
('necessary', 585),	(('part', 'united'), 0.0002940249026308837),	(('retired', 'list'), 12.831309702866573),
('constitution', 570),	(('indian', 'tribe'), 0.0002812412112121496),	(('circulating', 'medium'), 12.795904366735378),
('territory', 562),	(('mexican', 'government'), 0.00028124121121214 96),	(('rocky', 'mountain'), 12.795904366735378),
('new', 548),	(('naval', 'force'), 0.0002812412112121496),	(('intent', 'meaning'), 12.77728868856803),
('object', 532),	(('several', 'state'), 0.00027697998073923825),	(('john', 'quincy'), 12.752835644843492),
('foreign', 519),	(('treasury', 'note'), 0.00027697998073923825),	(('th', 'jefferson'), 12.75283564484349),
('measure', 512),	(('article', 'treaty'), 0.00026845751979341555),	(('thomas', 'jefferson'), 12.665372803593149),
('two', 510),	(('commercial', 'intercourse'), 0.00026845751979341555),	(('precious', 'metal'), 12.595294367857012),
('system', 509),	(('provision', 'made'), 0.0002684575197934155 5),	(('almighty', 'god'), 12.582910643401181),
('commerce', 506),	(('secretary', 'state'), 0.0002641962893205042),	(('john', 'tyler'), 12.582910643401181),
('peace', 501),	(('state', 'government'), 0.00025567382837468145),	(('san', 'francisco'), 12.554896267231582),
('consideration', 493),	(('american', 'citizen'), 0.0002514125979017701),	(('san', 'jacinto'), 12.55489626723158),
('within', 479),	(('claim', 'citizen'), 0.0002514125979017701),	(('san', 'juan'), 12.55489626723158),
('service', 479),	(('public', 'interest'), 0.0002514125979017701),	(('seizure', 'confiscation'), 12.518370391206469),
('condition', 476),	(('foreign', 'power'), 0.0002428901369559474),	(('rio', 'grande'), 12.414033731391731),
('relation', 476),	(('address', 'december'), 0.00023862890648303605),	(('effusion', 'blood'), 12.367810714631089),
('effect', 469)	(('new', 'mexico'), 0.00023862890648303605)	(('inferior', 'quality'), 12.069469440061338)

Analysis 2: Analysis of State of the Union Addresses dataset: Part2

The following tasks are analyzed with "state_union_part2.txt" file. Since four circumstance of processing the words have been tested in part 1, the analysis of part2 would apply the most meaningful method which is the "Lowercase + Punctuation Removed + Stop Words Removed + Lemmatization" circumstance.

A) list the top 50 words by frequency

The following steps were performed via python coding to list the top 50 words by frequency.

Step 1: Load the state union_part1.txt using NLTK package

Step 2: Perform tokenization to separate words from the text file

Step 3: Convert tokenized words to lower case and remove punctuation marks

Step 4: Get a list of stop words from NLTK and remove these stop words from text

Step 5: Use NLTK lemmatizer to obtain the root words

Step 6 Calculate the frequency distribution of the lemmatized words

Step 7: Find the top 50 words by frequency based using most_common function

B) list the top 50 bigrams by frequencies

The following steps were performed via python coding to list the top 50 bigrams by frequencies.

Step 1: import collocation finder package from NLTK to calculate bigram measure

Step 2: Apply collocation finder to the initial list of tokenized words from step 5 of previous task (task A)

Step 3: Use score_ngrams to score the bigrams by frequency and print the top 50 bigrams.

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

The following steps were performed via python coding to list the top 50 bigrams by frequencies.

Step 1: import collocation finder package from NLTK to calculate bigram measure

Step 2: Use apply_freq_filter to identify the bigrams with min frequency as 5

Step 3: Apply filtered finder to the initial list of tokenized words from step 5 of previous task (task A)

Step 4: Print the top 50 bigrams with their pmi scores

Output of Analysis 2

The following table is the output of the three tasks of analysis 2

Text 2 (Complete State of the Union Addresses from 1946 to 2016)		
Lowercase + Punctuation Removed + Stop Words Removed + Lemmatization		
Top 50 words by frequency	Top 50 bigrams by frequencies	Top 50 bigrams by PMI
('year', 2376),	(('united', 'state'), 0.002096959408857157),	(('el', 'salvador'), 15.164271890106203),
('american', 1638),	(('last', 'year'), 0.0012481901243197363),	(('bin', 'laden'), 14.94187946876976),
('must', 1628),	(('state', 'union'), 0.0012300346316023584),	(('saudi', 'arabia'), 14.941879468769756),
('people', 1597),	(('american', 'people'), 0.001098407309401368),	(('sam', 'rayburn'), 14.749234390827358),
('nation', 1506),	(('fiscal', 'year'), 0.0008805413967928322),	(('endowed', 'creator'), 14.316274983551253),
('world', 1490),	(('year', 'ago'), 0.0008669247772547988),	(('jimmy', 'carter'), 14.289802772190063),
('new', 1441),	(('federal', 'government'), 0.0008442304113580763),	(('northern', 'ireland'), 14.164271890106207),
('america', 1288),	(('social', 'security'), 0.0008260749186406982),	(('gerald', 'ford'), 14.097157694247668),
('congress', 1236),	(('health', 'care'), 0.0008079194259233203),	(('iron', 'curtain'), 13.842343795218842),
('state', 1231),	(('let', 'u'), 0.0008079194259233203),	(('floor', 'appears'), 13.74923439082736),
('government', 1219),	(('billion', 'dollar'), 0.0006944475964397079),	(('red', 'tape'), 13.74923439082736),
('u', 1216),	(('union', 'address'), 0.0006445199914669184),	(('jill', 'biden'), 13.678845062935963),
('program', 1100),	(('united', 'nation'), 0.000612747879211507),	(('thomas', 'jefferson'), 13.678845062935963),
('time', 871),	(('million', 'dollar'), 0.0005991312596734735),	(('barack', 'obama'), 13.66177154957702),
('country', 855),	(('soviet', 'union'), 0.0005673591474180619),	(('lyndon', 'johnson'), 13.66177154957702),
('make', 853),	(('next', 'year'), 0.000558281401059373),	(('teen', 'pregnancy'), 13.526841969490913),
('one', 840),	(('men', 'woman'), 0.0005128926692659281),	(('abraham', 'lincoln'), 13.49184654813471),
('work', 834),	(('past', 'year'), 0.0005083537960865835),	(('william', 'clinton'), 13.327770622389085),
('need', 822),	(('free', 'world'), 0.0004947371765485501),	(('ronald', 'reagan'), 13.289802772190063),
('every', 780),	(('member', 'congress'), 0.00045842619111379407),	(('mom', 'dad'), 13.263807563657119),
('federal', 744),	(('every', 'american'), 0.0004493484447551051),	(('greece', 'turkey'), 13.204913874603552),
('help', 720),	(('million', 'american'), 0.0004357318252170716),	('elementary', 'secondary'), 13.122795254130045
('war', 702),	(('economic', 'growth'), 0.0004266540788583826),	intercontinental', 'ballistic'), 13.00328001343389
('million', 694),	(('middle', 'east'), 0.0004130374593203491),	(('feeding', 'hungry'), 12.9678746773027),
('security', 689),	(('state', 'local'), 0.00040849858614100463),	(('grass', 'root'), 12.962638028936553),
('tax', 685),	(('free', 'nation'), 0.00040395971296166015),	(('lady', 'gentleman'), 12.941879468769756),
('job', 678),	(('make', 'sure'), 0.00040395971296166015),	(('status', 'quo'), 12.891253395699788),
('economic', 671),	(('four', 'year'), 0.00038126534706493765),	(('empowerment', 'zone'), 12.842343795218842)
('peace', 668),	(('first', 'time'), 0.00036764872752690416),	(('nationwide', 'radio'), 12.801701810721497),
('united', 651),	(('small', 'business'), 0.00036764872752690416),	(('radio', 'television'), 12.749234390827361),
('also', 639),	(('ask', 'congress'), 0.0003585709811682152),	(('dwight', 'eisenhower'), 12.643439726804763),
('economy', 622),	(('world', 'war'), 0.0003585709811682152),	(('al', 'qaeda'), 12.619951373882396),
('right', 619),	(('armed', 'force'), 0.0003449543616301817),	(('al', 'qaida'), 12.619951373882394),
('national', 610),	(('tax', 'cut'), 0.0003449543616301817),	(('richard', 'nixon'), 12.602877860523451),
('child', 609),	(('foreign', 'policy'), 0.0003358766152714927),	(('saddam', 'hussein'), 12.579309389385045),
('great', 583),	(('must', 'continue'), 0.0003358766152714927),	(('introduced', 'thomas'), 12.563367845516026),
('last', 574),	(('new', 'job'), 0.0003358766152714927),	(('harry', 'truman'), 12.539781025198408),
('many', 563),	(('work', 'together'), 0.0003358766152714927),	(('prime', 'minister'), 12.501306877383774),
('free', 559),	(('two', 'year'), 0.00030410450301608124),	(('reported', 'floor'), 12.501306877383772),
('let', 554),	(('vice', 'president'), 0.00030410450301608124),	(('persian', 'gulf'), 12.437176539964058),
('first', 553),	(('local', 'government'), 0.00029956562983673676),	(('carbon', 'pollution'), 12.427306295940001),
('would', 548),	(('around', 'world'), 0.00029048788347804774),	(('capitol', 'introduced'), 12.396717976106578),
('effort', 547),	(('national', 'security'), 0.00028594901029870327),	(('george', 'bush'), 12.283853505858877),
('know', 536),	(('must', 'also'), 0.00028141013711935873),	(('synthetic', 'fuel'), 12.25417886245934),
('budget', 531),	(('address', 'january'), 0.00027723323907606698),	(('panama', 'canal'), 12.176344722406778),
('system', 531),	(('human', 'right'), 0.00027723323907606698),	(('per', 'caput'), 12.105378201052636),
('life', 526),	(('health', 'insurance'), 0.00026325464440198076),	(('franklin', 'roosevelt'), 12.076809048855866),
('family', 525),	(('nation', 'world'), 0.00026325464440198076),	(('chemical', 'biological'), 12.012268796661154),
('force', 518),	(('civil', 'right'), 0.0002587157712226363),	(('baby', 'boom'), 11.983699644464382),
('freedom', 515)	(('fellow', 'citizen'), 0.0002587157712226363)	(('steam', 'coal'), 11.90123748427241)

Comparison on Analysis 1 and Analysis 2

The following comparison was based on the analysis results from question 2 and question 3.

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

1) Comparing the results of top 50 words from analysis 1 and 2

Out of the most frequent 50 words that were generated from text 1 and text 2, There were 20 words in common.

The common words from the top 50 words are: state, government, united, congress, country, would, year, great, made, time, last, war, nation, people, right, one, every, new, system, peace. From these words, we can see that these addresses focus on some everlasting political topics like the relationship between government and people, war and peace.

2) Comparing the results of top 50 bigrams by frequencies from analysis 1 and 2

Out of the most frequent 50 bigrams that were generated from text 1 and text 2, There were 8 bigrams in common.

The common bigrams from the top 50 bigrams by frequencies are: united states, state union, fiscal year, union address, fellow citizens, federal government. These bigrams indicate that no matter when the addresses were delivered, they were about the government work and economy and facing to the citizens.

3) Comparing the results of top 50 bigrams by PMI from analysis 1 and 2

Out of the top 50 bigrams by PMI that were generated from text 1 and text 2, There was only one bigram in common, which was thomas jefferson.

Interpretation

We can assume the similarity between these 2 texts by measuring how many top 50 words/bigrams in common, but these assumptions need more solid supports.

The similarities between these texts reflected some topics were always in the president addresses because they are basic political issues.

However, the difference between these analyses are more interesting. In text 1, foreign countries like Britain, Mexico were mentioned a lot, which indicated that for the US presidents, foreign affairs with some countries were their main concerns. While in the text 2, soviet union were more mentioned, which reflected the ‘cold war’ background. And the word “world” were more frequent in text 2, which indicated that the international status of America has changed. Another remarkable difference between these texts was that in the text 1, words like “act”, “senate” and “secretary” are frequent, which means that in the early days of America, laying the foundation of

this country like forming a government, establishing acts is the most important political issues. While in the text 2, the addresses emphasis more on individual's wellbeing, so the words/phrases "health care/insurance", "tax", "job" and "small business" are more frequent.

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

There are several problems with the words or bigrams in the analyses.

First, the definition of "stop word". Words like 'upon' can be regarded as "not that meaningful/important" when it appears as the most frequent words, so I tried to customized stop words list to remove it. But this word turned out to be meaningful in the bigram "upon subject". This problem informed me how important the context is when analyzing words.

Second, polysemous words. For example, the word "right" can refer to the citizen rights or the evaluation of the government, it has to be understood in a bigram/phrase. Another example is the word "act", which can be either verb or noun. I think this is a drawback of lemmatization. If the word "act" appears as "acted" or "acts", we might understand its meaning better. Though lemmatization can eliminate some duplicate words/bigrams, it may confuse us when there are polysemous words.

To get a better list of bigrams, we can:

1). Strategically execute stemming and lemmatization to reduce duplicates but keep the meaning of the words;

2). Strategically define stop words to remove auxiliary words.

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

After analyzing the top 50 bigrams by frequency and scored by PMI in both text1 and text 2, no commonality was found.

One possible reason is that the frequency of bigram was set as 5 in the analysis.

Besides frequency, PMI also considers the collocations and associations between words, which could provide more meaningful results.

Appendix: Python code & output

In [ ]:

```
# NLP Homework 1
# Due: Sunday, September 22, 11:59 pm.
# Harper He
# xhe128@syr.edu
```

In [1]:

```
import nltk
```

In [105]:

```
#### Analysis of State of the Union Addresses dataset: Part1
# Read file
sup1=open('state_union_part1.txt')
raw1=sup1.read()
```

In [106]:

```
# check the type and length of the file
print(type(raw1))
print(len(raw1))
```

```
<class 'str'>
3087306
```

In [107]:

```
# show some of the words
raw1[:300]
```

Out[107]:

```
'The Project Gutenberg EBook of Complete State of the Union Addresses,\nfrom 1790 to the
Present\n(#41 in our series of US Presidential State of the Union Addresses)\n\nCopyright laws are
changing all over the world. Be sure to check the\ncopyright laws for your country before
downloading or redistributin'
```

In [111]:

```
# convert string into tokens and then text
from nltk.tokenize import *
tokens1=word_tokenize(raw1)
text1=nltk.Text(tokens1)
```

In [112]:

```
# remove punctuations
words1 = [w.lower() for w in text1 if w.isalpha()]
```

In [113]:

```
print(type(words1))
words1[:30]
```

```
<class 'list'>
```

Out[113]:

```
['the',
 'project',
 'gutenberg',
 'ebook',
 'of',
```

```
 'complete',
 'state',
 'of',
 'the',
 'union',
 'addresses',
 'from',
 'to',
 'the',
 'present',
 'in',
 'our',
 'series',
 'of',
 'us',
 'presidential',
 'state',
 'of',
 'the',
 'union',
 'addresses',
 'copyright',
 'laws',
 'are',
 'changing']
```

In [88]:

```
### use all lower case
```

In [89]:

```
## list the top 50 words by frequency
```

In [114]:

```
# Creating a frequency distribution of words
fdist1 = nltk.FreqDist(words1)
# print the top 50 tokens by frequency
fdist1.most_common(50)
```

Out[114]:

```
[('the', 49523),
 ('of', 32404),
 ('to', 19129),
 ('and', 16905),
 ('in', 10464),
 ('a', 7839),
 ('that', 6658),
 ('be', 6354),
 ('it', 5492),
 ('by', 5451),
 ('which', 5243),
 ('for', 4746),
 ('our', 4318),
 ('with', 4144),
 ('is', 4000),
 ('have', 3786),
 ('as', 3734),
 ('been', 3692),
 ('on', 3370),
 ('this', 3151),
 ('has', 2788),
 ('their', 2775),
 ('from', 2758),
 ('not', 2750),
 ('states', 2725),
 ('will', 2546),
 ('was', 2231),
 ('government', 2220),
 ('at', 2198),
 ('its', 2041),
 ('all', 1949),
```

```
 ('i', 1947),
 ('an', 1940),
 ('are', 1895),
 ('united', 1864),
 ('or', 1683),
 ('they', 1575),
 ('may', 1562),
 ('congress', 1500),
 ('them', 1475),
 ('but', 1462),
 ('upon', 1455),
 ('would', 1381),
 ('public', 1375),
 ('such', 1340),
 ('other', 1269),
 ('these', 1208),
 ('any', 1171),
 ('country', 1163),
 ('should', 1159)]
```

In [115]:

```
## list the top 50 bigrams by frequencies
bigrams1=list(nltk.bigrams(words1))
```

In [116]:

```
# setup for bigrams and bigram measures
# from nltk.collocations import *
bigram_measures = nltk.collocations.BigramAssocMeasures()
```

In [119]:

```
# create the bigram finder and score the bigrams by frequency
# scored is a list of bigram pairs with their score
finder1 = BigramCollocationFinder.from_words(words1)
scored1 = finder1.score_ngrams(bigram_measures.raw_freq)
scored1[:50]
```

Out[119]:

```
[(('of', 'the'), 0.023139252704378124),
 (('to', 'the'), 0.007783594908345917),
 (('in', 'the'), 0.006009076290590155),
 (('by', 'the'), 0.00399560483468185),
 (('for', 'the'), 0.0036783068036924087),
 (('united', 'states'), 0.0035686235337207503),
 (('the', 'united'), 0.003543161346048758),
 (('and', 'the'), 0.0033061671377171385),
 (('on', 'the'), 0.0032160701659547045),
 (('of', 'our'), 0.0030887592275947438),
 (('it', 'is'), 0.0028635167981886592),
 (('to', 'be'), 0.0028615581683677366),
 (('have', 'been'), 0.0026284812196779622),
 (('with', 'the'), 0.002505087540959846),
 (('that', 'the'), 0.002450245905974017),
 (('has', 'been'), 0.0024189078288392573),
 (('from', 'the'), 0.0021388237644473433),
 (('of', 'a'), 0.0019096640753994136),
 (('the', 'public'), 0.00179606354578591),
 (('will', 'be'), 0.001680504386351484),
 (('the', 'government'), 0.0016452490495748795),
 (('at', 'the'), 0.0015042277024684611),
 (('may', 'be'), 0.001306406090555291),
 (('of', 'congress'), 0.0012672334941368417),
 (('and', 'to'), 0.0012300195275393147),
 (('upon', 'the'), 0.001120336257567656),
 (('of', 'this'), 0.0011046672190002762),
 (('of', 'their'), 0.0011007499593584313),
 (('the', 'same'), 0.0011007499593584313),
 (('the', 'present'), 0.0010889981804328964),
 (('the', 'people'), 0.0010086943577750748),
 (('between', 'the'), 0.0009949839490286175),
 (('all', 'the'), 0.0009832321701030828),
```

(('in', 'a'), 0.0009812735402821601),
(('state', 'of'), 0.0009773562806403152),
(('under', 'the'), 0.0009695217613566254),
(('the', 'country'), 0.000959728612252013),
(('which', 'the'), 0.000955811352610168),
(('of', 'that'), 0.0009538527227892456),
(('the', 'union'), 0.0009479768333264781),
(('should', 'be'), 0.0009342664245800208),
(('it', 'was'), 0.0009283905351172534),
(('the', 'treasury'), 0.0009264319052963309),
(('would', 'be'), 0.0008539626019221994),
(('the', 'last'), 0.0008480867124594319),
(('part', 'of'), 0.0008382935633548195),
(('the', 'constitution'), 0.000836334933533897),
(('can', 'not'), 0.0008147900055037498),
(('the', 'most'), 0.0008128313756828274),
(('and', 'of'), 0.0007971623371154476)]

In [120]:

```
## list the top 50 bigrams by their Mutual Information scores (using min frequency 5)
finder11 = BigramCollocationFinder.from_words(words1)
finder11.apply_freq_filter(5)
scored11 = finder11.score_ngrams(bigram_measures.pmi)
scored11[:50]
```

Out[120]:

[(('bona', 'fide'), 16.63979571898761),
 (('posse', 'comitatus'), 16.63979571898761),
 (('punta', 'arenas'), 16.63979571898761),
 (('ballot', 'box'), 16.376761313153814),
 (('del', 'norte'), 16.376761313153814),
 (('millard', 'fillmore'), 16.376761313153814),
 (('guadalupe', 'hidalgo'), 15.79179881243266),
 (('porto', 'rico'), 15.79179881243266),
 (('franklin', 'pierce'), 15.63979571898761),
 (('la', 'plata'), 15.502292195237674),
 (('vera', 'cruz'), 15.376761313153814),
 (('entangling', 'alliances'), 15.306371985262416),
 (('gun', 'boats'), 14.984443890375054),
 (('costa', 'rica'), 14.961723813874972),
 (('nucleus', 'around'), 14.961723813874972),
 (('santa', 'anna'), 14.874260972624633),
 (('santa', 'fe'), 14.874260972624633),
 (('van', 'buren'), 14.874260972624633),
 (('project', 'gutenberg'), 14.874260972624631),
 (('sublime', 'porte'), 14.832440796930005),
 (('martin', 'van'), 14.70433597118232),
 (('ad', 'valorem'), 14.639795718987608),
 (('beacons', 'buoys'), 14.502292195237674),
 (('water', 'witch'), 14.502292195237674),
 (('quincy', 'adams'), 14.502292195237672),
 (('statute', 'book'), 14.43816185781796),
 (('buenos', 'ayres'), 14.376761313153816),
 (('de', 'facto'), 14.228369473261143),
 (('franking', 'privilege'), 14.206836311711502),
 (('rocky', 'mountains'), 14.15436889181737),
 (('andrew', 'jackson'), 14.071906731625393),
 (('retired', 'list'), 14.016865368067432),
 (('circulating', 'medium'), 13.917329694516518),
 (('th', 'jefferson'), 13.874260972624633),
 (('john', 'quincy'), 13.874260972624631),
 (('precious', 'metals'), 13.815367283571065),
 (('thomas', 'jefferson'), 13.786798131374292),
 (('lake', 'erie'), 13.732905123379092),
 (('almighty', 'god'), 13.70433597118232),
 (('john', 'tyler'), 13.70433597118232),
 (('san', 'jacinto'), 13.676321595012723),
 (('san', 'juan'), 13.676321595012723),
 (('san', 'francisco'), 13.676321595012721),
 (('per', 'cent'), 13.60417180925689),
 (('rio', 'grande'), 13.569406391096214),
 (('inferior', 'quality'), 13.495749349370902),
 (('grateful', 'acknowledgments'), 13.469870717545295),
 (('hudsons', 'bay'), 13.407134962197334)

```
(('hudsons', 'bay'), 13.407134962197334),
(('cumberland', 'road'), 13.245516779875564),
(('cut', 'off'), 13.180364100350312)]
```

In [ ]:

```
### use all lower case, use the stop word list
```

In [121]:

```
# get a list of stopwords from nltk
from nltk.corpus import *
nltkstopwords = nltk.corpus.stopwords.words('english')
print(len(nltkstopwords))
nltkstopwords[:30]
```

```
179
```

Out[121]:

```
['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']
```

In [122]:

```
# remove all the stop words from text1
stoppedsup1words = [w for w in words1 if not w in nltkstopwords]
len(set(stoppedsup1words))
```

Out[122]:

```
11461
```

In [14]:

```
## list the top 50 words by frequency
```

In [123]:

```
# Creating a frequency distribution of words
fdist1stoppedsup1words = nltk.FreqDist(stoppedsup1words)
# print the top 50 tokens by frequency
fdist1stoppedsup1words.most_common(50)
```

Out[123]:

```
[('states', 2725),
 ('government', 2220),
 ('united', 1864),
 ('may', 1562),
 ('congress', 1500),
 ('upon', 1455),
 ('would', 1381),
 ('public', 1375),
 ('country', 1163),
 ('great', 1073),
 ('made', 1061),
 ('state', 1045),
 ('last', 911),
 ('war', 834),
 ('present', 812),
 ('time', 808),
 ('people', 786),
 ('year', 785),
 ('power', 744),
 ('citizens', 723),
 ('subject', 711),
 ('shall', 694),
 ('without', 663),
 ('union', 643),
 ('act', 627),
 ('treaty', 624),
 ('one', 620),
 ('part', 618),
 ('mexico', 605),
 ('general', 601),
 ('every', 590),
 ('treasury', 590),
 ('necessary', 575),
 ('constitution', 557),
 ('new', 548),
 ('duty', 529),
 ('foreign', 519),
 ('two', 510),
 ('commerce', 506),
 ('nations', 502),
 ('peace', 501),
 ('system', 494),
 ('laws', 492),
 ('duties', 488),
 ('within', 479),
 ('law', 477),
 ('us', 463),
 ('interests', 451),
 ('interest', 444),
 ('amount', 443)]
```

In [124]:

```
## list the top 50 bigrams by frequencies
bigrams1stoppedsup1words=list(nltk.bigrams(stoppedsup1words))
```

In [125]:

```
# create the bigram finder and score the bigrams by frequency
# scored is a list of bigram pairs with their score
finder1stoppedsup1words = BigramCollocationFinder.from_words(stoppedsup1words)
scored1stoppedsup1words = finder1stoppedsup1words.score_ngrams(bigram_measures.raw_freq)
```

In [126]:

```
scored1stoppedsup1words[:50]
```

Out[126]:

```
[(('united', 'states'), 0.007648905979748451),
 (('great', 'britain'), 0.0011502745545834663),
```

```
 (('last', 'session'), 0.0010159359204715286),
 (('public', 'debt'), 0.0007514567345636513),
 (('state', 'union'), 0.0007262682406676631),
 (('house', 'representatives'), 0.0006213161827677117),
 (('fiscal', 'year'), 0.0006045238535037195),
 (('union', 'address'), 0.0006045238535037195),
 (('report', 'secretary'), 0.0005835334419237293),
 (('public', 'lands'), 0.0005457507010797468),
 (('two', 'countries'), 0.0005121660425517623),
 (('present', 'year'), 0.0004449967254957935),
 (('within', 'limits'), 0.0004198023159980523),
 (('secretary', 'treasury'), 0.00041561014928380717),
 (('fellow', 'citizens'), 0.00040721398465181106),
 (('session', 'congress'), 0.00040721398465181106),
 (('act', 'congress'), 0.0003946197377038169),
 (('general', 'government'), 0.00039042165538781884),
 (('year', 'ending'), 0.00039042165538781884),
 (('british', 'government'), 0.0003862235730718208),
 (('two', 'governments'), 0.0003736293261238266),
 (('citizens', 'united'), 0.0003610350791758325),
 (('federal', 'government'), 0.00035683699685983445),
 (('secretary', 'war'), 0.0003526389145438364),
 (('annual', 'message'), 0.0003400446675958422),
 (('public', 'service'), 0.00033584658527984417),
 (('senate', 'house'), 0.00033584658527984417),
 (('consideration', 'congress'), 0.00032325233833185),
 (('ending', 'june'), 0.0003148561736998539),
 (('last', 'annual'), 0.0003148561736998539),
 (('attention', 'congress'), 0.00031065809138385584),
 (('government', 'united'), 0.0003064600090678578),
 (('public', 'money'), 0.0002896676798038656),
 (('indian', 'tribes'), 0.00027707343285587145),
 (('mexican', 'government'), 0.0002728753505398734),
 (('part', 'united'), 0.0002728753505398734),
 (('treasury', 'notes'), 0.0002728753505398734),
 (('upon', 'subject'), 0.00026867726822387534),
 (('commercial', 'intercourse'), 0.0002644791859078773),
 (('several', 'states'), 0.0002644791859078773),
 (('secretary', 'state'), 0.0002602811035918792),
 (('provision', 'made'), 0.00024768685664388506),
 (('article', 'treaty'), 0.00023929069201188898),
 (('claims', 'citizens'), 0.00023929069201188898),
 (('address', 'december'), 0.00023509260969589092),
 (('new', 'mexico'), 0.00023509260969589092),
 (('favorable', 'consideration'), 0.00023089452737989286),
 (('naval', 'force'), 0.00023089452737989286),
 (('bank', 'united'), 0.0002266964450638948),
 (('people', 'united'), 0.0002266964450638948)]
```

In [127]:

```
# list the top 50 bigrams by their Mutual Information scores (using min frequency 5)
finder11stoppedsup1words = BigramCollocationFinder.from_words(stoppedsup1words)
finder11stoppedsup1words.apply_freq_filter(5)
scored11stoppedsup1words = finder11stoppedsup1words.score_ngrams(bigram_measures.pmi)
scored11stoppedsup1words[:50]
```

Out[127]:

```
[(('bona', 'fide'), 15.539910019165042),
 (('posse', 'comitatus'), 15.539910019165042),
 (('punta', 'arenas'), 15.539910019165042),
 (('ballot', 'box'), 15.276875613331246),
 (('del', 'norte'), 15.276875613331246),
 (('millard', 'fillmore'), 15.276875613331246),
 (('clayton', 'bulwer'), 14.861838114052404),
 (('guadalupe', 'hidalgo'), 14.691913112610091),
 (('porto', 'rico'), 14.691913112610091),
 (('writ', 'mandamus'), 14.598803708218608),
 (('franklin', 'pierce'), 14.539910019165042),
 (('la', 'plata'), 14.402406495415105),
 (('vera', 'cruz'), 14.276875613331246),
 (('entangling', 'alliances'), 14.206486285439848),
 (('seminaries', 'learning'), 14.013841207497453),
 (('gun', 'boats'), 13.884558190552486),
 (('nucleus', 'around'), 13.861838114052404)
```

```
(('nucleus', 'around'), 13.661838114052404),
(('ruler', 'universe'), 13.861838114052404),
(('costa', 'rica'), 13.8618381140524),
(('santa', 'anna'), 13.774375272802065),
(('santa', 'fe'), 13.774375272802065),
(('van', 'buren'), 13.774375272802065),
(('project', 'gutenberg'), 13.774375272802063),
(('sublime', 'porte'), 13.732555097107436),
(('tea', 'coffee'), 13.613910600608818),
(('martin', 'van'), 13.604450271359752),
(('ad', 'valorem'), 13.53991001916504),
(('beacons', 'buoys'), 13.402406495415105),
(('water', 'witch'), 13.402406495415105),
(('quincy', 'adams'), 13.402406495415104),
(('statute', 'book'), 13.338276157995391),
(('buenos', 'ayres'), 13.276875613331244),
(('indiana', 'illinois'), 13.139372089581311),
(('de', 'facto'), 13.128483773438575),
(('franking', 'privilege'), 13.106950611888934),
(('rocky', 'mountains'), 13.054483191994798),
(('andrew', 'jackson'), 12.972021031802825),
(('retired', 'list'), 12.916979668244863),
(('sooner', 'later'), 12.876945006442611),
(('circulating', 'medium'), 12.81744399469395),
(('intent', 'meaning'), 12.798828316526604),
(('th', 'jefferson'), 12.774375272802065),
(('john', 'quincy'), 12.774375272802063),
(('precious', 'metals'), 12.715481583748494),
(('thomas', 'jefferson'), 12.686912431551724),
(('lake', 'erie'), 12.633019423556524),
(('almighty', 'god'), 12.604450271359752),
(('john', 'tyler'), 12.604450271359752),
(('san', 'jacinto'), 12.576435895190155),
(('san', 'juan'), 12.576435895190155)]
```

In [ ]:

```
### use all lower case, use the stop word list, and lemmatization
```

In [128]:

```
# NLTK has a lemmatizer that uses WordNet as a dictionary
wnl = nltk.WordNetLemmatizer()
```

In [129]:

```
Lemmasup1words = [wnl.lemmatize(t) for t in stoppedsup1words]
Lemmasup1words[:30]
```

Out[129]:

```
['project',
 'gutenberg',
 'ebook',
 'complete',
 'state',
 'union',
 'address',
 'present',
 'series',
 'u',
 'presidential',
 'state',
 'union',
 'address',
 'copyright',
 'law',
 'changing',
 'world',
 'sure',
 'check',
 'copyright',
 'law',
 'country',
```

```
 'downloading',
 'redistributing',
 'project',
 'gutenberg',
 'ebook',
 'header',
 'first']
```

In [14]:

```
## list the top 50 words by frequency
```

In [130]:

```
# Creating a frequency distribution of words
fdist1Lemmasup1words = nltk.FreqDist(Lemmasup1words)
# print the top 50 tokens by frequency
fdist1Lemmasup1words.most_common(50)
```

Out[130]:

```
[('state', 3770),
 ('government', 2561),
 ('united', 1864),
 ('may', 1562),
 ('congress', 1501),
 ('upon', 1455),
 ('country', 1427),
 ('would', 1381),
 ('public', 1375),
 ('power', 1159),
 ('year', 1145),
 ('great', 1073),
 ('made', 1061),
 ('duty', 1017),
 ('law', 969),
 ('time', 914),
 ('last', 911),
 ('war', 898),
 ('interest', 895),
 ('subject', 888),
 ('present', 852),
 ('nation', 830),
 ('act', 807),
 ('citizen', 801),
 ('people', 786),
 ('treaty', 748),
 ('part', 741),
 ('shall', 694),
 ('without', 663),
 ('union', 643),
 ('right', 636),
 ('one', 634),
 ('general', 605),
 ('mexico', 605),
 ('treasury', 592),
 ('every', 590),
 ('necessary', 585),
 ('constitution', 570),
 ('territory', 562),
 ('new', 548),
 ('object', 532),
 ('foreign', 519),
 ('measure', 512),
 ('two', 510),
 ('system', 509),
 ('commerce', 506),
 ('peace', 501),
 ('consideration', 493),
 ('within', 479),
 ('service', 479)]
```

In [136]:

```
## list the top 50 bigrams by frequencies
bigrams1Lemmasup1words=list(nltk.bigrams(Lemmasup1words))
```

In [148]:

```
# create the bigram finder and score the bigrams by frequency
# scored is a list of bigram pairs with their score
finder1Lemmasup1words = BigramCollocationFinder.from_words(Lemmasup1words)
scored1Lemmasup1words = finder1Lemmasup1words.score_ngrams(bigram_measures.raw_freq)
scored1Lemmasup1words[:50]
```

Out[148]:

```
[(('united', 'state'), 0.007653104062064449),
 (('great', 'britain'), 0.0011502745545834663),
 (('last', 'session'), 0.0010243320851035247),
 (('state', 'union'), 0.000885795368675589),
 (('public', 'debt'), 0.0007598528991956474),
 (('house', 'representative'), 0.0006213161827677117),
 (('union', 'address'), 0.0006213161827677117),
 (('report', 'secretary'), 0.0006171181004517136),
 (('fiscal', 'year'), 0.0006087219358197175),
 (('public', 'land'), 0.000562543030343739),
 (('act', 'congress'), 0.0005121660425517623),
 (('two', 'country'), 0.0005121660425517623),
 (('present', 'year'), 0.0004491948078117956),
 (('within', 'limit'), 0.0004407986431797945),
 (('public', 'money'), 0.0004366005608637974),
 (('session', 'congress'), 0.0004324024785477934),
 (('secretary', 'treasury'), 0.0004156101492838071),
 (('fellow', 'citizen'), 0.0004072139846518106),
 (('general', 'government'), 0.000403015902335813),
 (('british', 'government'), 0.0003904216553878184),
 (('secretary', 'war'), 0.0003904216553878184),
 (('year', 'ending'), 0.0003904216553878184),
 (('citizen', 'united'), 0.0003778274084398247),
 (('two', 'government'), 0.0003736293261238266),
 (('federal', 'government'), 0.0003652331614918305),
 (('annual', 'message'), 0.0003568369968598345),
 (('public', 'service'), 0.0003358465852798417),
 (('senate', 'house'), 0.0003358465852798417),
 (('consideration', 'congress'), 0.0003316485029638461),
 (('ending', 'june'), 0.0003148561736998539),
 (('last', 'annual'), 0.0003148561736998539),
 (('attention', 'congress'), 0.0003106580913838584),
 (('government', 'united'), 0.0003064600090678578),
 (('part', 'united'), 0.0002896767679803656),
 (('upon', 'subject'), 0.0002854695975874756),
 (('indian', 'tribe'), 0.0002770734328558714),
 (('mexican', 'government'), 0.0002770734328558714),
 (('naval', 'force'), 0.0002770734328558714),
 (('several', 'state'), 0.0002728753505398734),
 (('treasury', 'note'), 0.0002728753505398734),
 (('article', 'treaty'), 0.0002644791859078773),
 (('commercial', 'intercourse'), 0.0002644791859078773),
 (('state', 'would'), 0.0002644791859078773),
 (('secretary', 'state'), 0.0002602811035918792),
 (('provision', 'made'), 0.0002560830212758811),
 (('american', 'citizen'), 0.0002476868566438506),
 (('claim', 'citizen'), 0.0002476868566438506),
 (('public', 'interest'), 0.0002476868566438506),
 (('state', 'government'), 0.0002434887743278703),
 (('foreign', 'power'), 0.0002392906920118898)]
```

In [149]:

```
## list the top 50 bigrams by their Mutual Information scores (using min frequency 5)
finder11Lemmasup1words = BigramCollocationFinder.from_words(Lemmasup1words)
finder11Lemmasup1words.apply_freq_filter(5)
scored11Lemmasup1words = finder11Lemmasup1words.score_ngrams(bigram_measures.pmi)
scored11Lemmasup1words[:50]
```

Out[149]:

```
[(('bona', 'fide'), 15.539910019165042)
```

```
[(('bona', 'fide'), 15.559910019165042),
 (('del', 'norte'), 15.276875613331246),
 (('millard', 'fillmore'), 15.276875613331246),
 (('punta', 'arena'), 15.276875613331246),
 (('ballot', 'box'), 15.054483191994798),
 (('clayton', 'bulwer'), 14.861838114052404),
 (('guadalupe', 'hidalgo'), 14.691913112610091),
 (('porto', 'rico'), 14.691913112610091),
 (('writ', 'mandamus'), 14.598803708218608),
 (('franklin', 'pierce'), 14.539910019165042),
 (('la', 'plata'), 14.402406495415105),
 (('vera', 'cruz'), 14.276875613331246),
 (('entangling', 'alliance'), 14.054483191994798),
 (('seminary', 'learning'), 14.0398364160304),
 (('nucleus', 'around'), 13.861838114052404),
 (('costa', 'rica'), 13.8618381140524),
 (('santa', 'anna'), 13.774375272802065),
 (('santa', 'fe'), 13.774375272802065),
 (('van', 'buren'), 13.774375272802065),
 (('sublime', 'porte'), 13.732555097107436),
 (('tea', 'coffee'), 13.613910600608818),
 (('martin', 'van'), 13.604450271359752),
 (('ad', 'valorem'), 13.53991001916504),
 (('quincy', 'adam'), 13.402406495415104),
 (('buenos', 'ayres'), 13.276875613331244),
 (('beacon', 'buoy'), 13.264902971665169),
 (('ruler', 'universe'), 13.198873101329973),
 (('indiana', 'illinois'), 13.139372089581311),
 (('de', 'facto'), 13.128483773438575),
 (('project', 'gutenberg'), 13.106950611888934),
 (('gun', 'boat'), 12.991473394468999),
 (('andrew', 'jackson'), 12.972021031802825),
 (('sooner', 'later'), 12.876945006442611),
 (('retired', 'list'), 12.852849330825148),
 (('circulating', 'medium'), 12.81744399469395),
 (('rocky', 'mountain'), 12.81744399469395),
 (('intent', 'meaning'), 12.798828316526604),
 (('th', 'jefferson'), 12.774375272802065),
 (('john', 'quincy'), 12.774375272802063),
 (('thomas', 'jefferson'), 12.686912431551724),
 (('precious', 'metal'), 12.616833995815583),
 (('almighty', 'god'), 12.604450271359752),
 (('john', 'tyler'), 12.604450271359752),
 (('san', 'jacinto'), 12.576435895190155),
 (('san', 'juan'), 12.576435895190155),
 (('san', 'francisco'), 12.576435895190153),
 (('seizure', 'confiscation'), 12.539910019165044),
 (('rio', 'grande'), 12.435573359350302),
 (('effusion', 'blood'), 12.38935034258966),
 (('inferior', 'quality'), 12.091009068019913)]
```

In [139]:

```
### Lowercase + Punctuation Removed + Stop Words Removed + Self-defined Stop Words Removed+ Lemmat
ization
#self define more stop words
morestopwords = ['would',"upon","shall"]
```

In [140]:

```
stopwords = nltkstopwords + morestopwords
len(stopwords)
```

Out[140]:

```
182
```

In [141]:

```
# remove all the stop words from text1
morestoppedsup1words = [w for w in words1 if not w in stopwords]
len(set(morestoppedsup1words))
```

Out[141]:

Out[141]:

```
11458
```

In [142]:

```
moreLemmasup1words = [wnl.lemmatize(t) for t in morestoppedsup1words]
moreLemmasup1words[:30]
```

Out[142]:

```
['project',
 'gutenberg',
 'ebook',
 'complete',
 'state',
 'union',
 'address',
 'present',
 'series',
 'u',
 'presidential',
 'state',
 'union',
 'address',
 'copyright',
 'law',
 'changing',
 'world',
 'sure',
 'check',
 'copyright',
 'law',
 'country',
 'downloading',
 'redistributing',
 'project',
 'gutenberg',
 'ebook',
 'header',
 'first']
```

In [14]:

```
## list the top 50 words by frequency
```

In [143]:

```
# Creating a frequency distribution of words
fdist1moreLemmasup1words = nltk.FreqDist(moreLemmasup1words)
# print the top 50 tokens by frequency
fdist1moreLemmasup1words.most_common(50)
```

Out[143]:

```
[('state', 3770),
 ('government', 2561),
 ('united', 1864),
 ('may', 1562),
 ('congress', 1501),
 ('country', 1427),
 ('public', 1375),
 ('power', 1159),
 ('year', 1145),
 ('great', 1073),
 ('made', 1061),
 ('duty', 1017),
 ('law', 969),
 ('time', 914),
 ('last', 911),
 ('war', 898),
 ('interest', 895),
 ('subject', 888),
 ('present', 852),
 ('nation', 830)
```

```
('nation', 830),
('act', 807),
('citizen', 801),
('people', 786),
('treaty', 748),
('part', 741),
('without', 663),
('union', 643),
('right', 636),
('one', 634),
('general', 605),
('mexico', 605),
('treasury', 592),
('every', 590),
('necessary', 585),
('constitution', 570),
('territory', 562),
('new', 548),
('object', 532),
('foreign', 519),
('measure', 512),
('two', 510),
('system', 509),
('commerce', 506),
('peace', 501),
('consideration', 493),
('within', 479),
('service', 479),
('condition', 476),
('relation', 476),
('effect', 469)]
```

In [144]:

```
## list the top 50 bigrams by frequencies
bigrams1moreLemmasup1words=list(nltk.bigrams(moreLemmasup1words))
```

In [146]:

```
# create the bigram finder and score the bigrams by frequency
# scored is a list of bigram pairs with their score
finder1moreLemmasup1words = BigramCollocationFinder.from_words(moreLemmasup1words)
scored1moreLemmasup1words = finder1moreLemmasup1words.score_ngrams(bigram_measures.raw_freq)
scored1moreLemmasup1words[:50]
```

Out[146]:

```
[(('united', 'state'), 0.007768223152117406),
 (('great', 'britain'), 0.0011675771495777121),
 (('last', 'session'), 0.0010397402353903714),
 (('state', 'union'), 0.0008991196297842965),
 (('public', 'debt'), 0.0007712827155969558),
 (('house', 'representative'), 0.0006306621099908809),
 (('union', 'address'), 0.0006306621099908809),
 (('report', 'secretary'), 0.0006264008795179696),
 (('fiscal', 'year'), 0.0006178784185721469),
 (('public', 'land'), 0.0005710048833701219),
 (('act', 'congress'), 0.0005198701176951857),
 (('two', 'country'), 0.0005198701176951857),
 (('present', 'year'), 0.0004559516606015153),
 (('within', 'limit'), 0.0004474291996556926),
 (('public', 'money'), 0.0004431679691827812 4),
 (('session', 'congress'), 0.0004389067387098699),
 (('secretary', 'treasury'), 0.00042186181681822443),
 (('fellow', 'citizen'), 0.00041333935587240173),
 (('general', 'government'), 0.0004090781253994904),
 (('british', 'government'), 0.0003962944339807563),
 (('secretary', 'war'), 0.0003962944339807563),
 (('year', 'ending'), 0.0003962944339807563),
 (('citizen', 'united'), 0.0003835107425620222),
 (('two', 'government'), 0.0003792495120891109),
 (('federal', 'government'), 0.0003707270511432881),
 (('annual', 'message'), 0.0003622045901974654),
 (('public', 'service'), 0.0003408984378329086),
 (('senate', 'house'), 0.0003408984378329086),
```

```
 (('consideration', 'congress'), 0.00033663720735999727),
 (('ending', 'june'), 0.00031959228546835186),
 (('last', 'annual'), 0.00031959228546835186),
 (('attention', 'congress'), 0.00031533105499544046),
 (('government', 'united'), 0.0003110698245225291),
 (('part', 'united'), 0.0002940249026308837),
 (('indian', 'tribe'), 0.00028212412112121496),
 (('mexican', 'government'), 0.00028212412112121496),
 (('naval', 'force'), 0.00028212412112121496),
 (('several', 'state'), 0.00027697998073923825),
 (('treasury', 'note'), 0.00027697998073923825),
 (('article', 'treaty'), 0.00026845751979341555),
 (('commercial', 'intercourse'), 0.00026845751979341555),
 (('provision', 'made'), 0.00026845751979341555),
 (('secretary', 'state'), 0.00026419962893205042),
 (('state', 'government'), 0.00025567382837468145),
 (('american', 'citizen'), 0.0002514125979017701),
 (('claim', 'citizen'), 0.0002514125979017701),
 (('public', 'interest'), 0.0002514125979017701),
 (('foreign', 'power'), 0.0002428901369559474),
 (('address', 'december'), 0.00023862890648303605),
 (('new', 'mexico'), 0.00023862890648303605)]
```

In [147]:

```
# list the top 50 bigrams by their Mutual Information scores (using min frequency 5)
finder11moreLemmasup1words = BigramCollocationFinder.from_words(moreLemmasup1words)
finder11moreLemmasup1words.apply_freq_filter(5)
scored11moreLemmasup1words = finder11moreLemmasup1words.score_ngrams(bigram_measures.pmi)
scored11moreLemmasup1words[:50]
```

Out[147]:

```
[(('bona', 'fide'), 15.518370391206467),
 (('del', 'norte'), 15.255335985372675),
 (('millard', 'fillmore'), 15.255335985372675),
 (('punta', 'arena'), 15.255335985372671),
 (('ballot', 'box'), 15.032943564036223),
 (('clayton', 'bulwer'), 14.840298486093833),
 (('guadalupe', 'hidalgo'), 14.670373484651517),
 (('porto', 'rico'), 14.670373484651517),
 (('writ', 'mandamus'), 14.577264080260033),
 (('franklin', 'pierce'), 14.51837039120647),
 (('la', 'plata'), 14.380866867456534),
 (('vera', 'cruz'), 14.255335985372675),
 (('entangling', 'alliance'), 14.032943564036227),
 (('seminary', 'learning'), 14.018296788071826),
 (('costa', 'rica'), 13.840298486093829),
 (('nucleus', 'around'), 13.840298486093829),
 (('santa', 'anna'), 13.75283564484349),
 (('santa', 'fe'), 13.75283564484349),
 (('van', 'buren'), 13.75283564484349),
 (('sublime', 'porte'), 13.711015469148865),
 (('tea', 'coffee'), 13.592370972650247),
 (('martin', 'van'), 13.582910643401181),
 (('ad', 'valorem'), 13.518370391206469),
 (('quincy', 'adam'), 13.380866867456533),
 (('buenos', 'ayres'), 13.255335985372673),
 (('beacon', 'buoy'), 13.243363343706594),
 (('ruler', 'universe'), 13.177333473371398),
 (('indiana', 'illinois'), 13.117832461622736),
 (('de', 'facto'), 13.10694414548),
 (('project', 'gutenberg'), 13.085410983930363),
 (('gun', 'boat'), 12.969933766510428),
 (('andrew', 'jackson'), 12.95048140384425),
 (('sooner', 'later'), 12.85540537848404),
 (('retired', 'list'), 12.831309702866573),
 (('circulating', 'medium'), 12.795904366735378),
 (('rocky', 'mountain'), 12.795904366735378),
 (('intent', 'meaning'), 12.77728868856803),
 (('john', 'quincy'), 12.752835644843492),
 (('th', 'jefferson'), 12.75283564484349),
 (('thomas', 'jefferson'), 12.665372803593149),
 (('precious', 'metal'), 12.595294367857012),
 (('almighty', 'god'), 12.582910643401181),
 (('john', 'tyler'), 12.582910643401181),
```

```
 (('san', 'francisco'), 12.554896267231582),
 (('san', 'jacinto'), 12.55489626723158),
 (('san', 'juan'), 12.55489626723158),
 (('seizure', 'confiscation'), 12.518370391206469),
 (('rio', 'grande'), 12.414033731391731),
 (('effusion', 'blood'), 12.367810714631089),
 (('inferior', 'quality'), 12.069469440061338)]
```

In [150]:

```
#### Analysis of State of the Union Addresses dataset: Part2
# Read file
sup2=open('state_union_part2.txt')
raw2=sup2.read()
```

In [151]:

```
# check the type and length of the file
print(type(raw2))
print(len(raw2))
```

```
<class 'str'>
2572378
```

In [152]:

```
# show some of the words
raw2[:300]
```

Out[152]:

```
'The Project Gutenberg EBook of Complete State of the Union Addresses,\nfrom 1946 to the
Present\n(#41 in our series of US Presidential State of the Union Addresses)\n\nCopyright laws are
changing all over the world. Be sure to check the\ncopyright laws for your country before
downloading or redistributin'
```

In [153]:

```
# convert string into tokens and then text
tokens2=word_tokenize(raw2)
text2=nltk.Text(tokens2)
```

In [154]:

```
# remove punctuations
words2 = [w.lower() for w in text2 if w.isalpha()]
```

In [156]:

```
print(type(words2))
words2[:30]
```

```
<class 'list'>
```

Out[156]:

```
['the',
 'project',
 'gutenberg',
 'ebook',
 'of',
 'complete',
 'state',
 'of',
 'the',
 'union',
 'addresses',
 'from',
 'to',
```

```
 'the',
 'present',
 'in',
 'our',
 'series',
 'of',
 'us',
 'presidential',
 'state',
 'of',
 'the',
 'union',
 'addresses',
 'copyright',
 'laws',
 'are',
 'changing']
```

In [157]:

```
# remove all the stop words from text1
stoppedsup2words = [w for w in words2 if not w in nltkstopwords]
len(set(stoppedsup2words))
```

Out[157]:

```
13193
```

In [158]:

```
Lemmasup2words = [wnl.lemmatize(t) for t in stoppedsup2words]
Lemmasup2words[:30]
```

Out[158]:

```
['project',
 'gutenberg',
 'ebook',
 'complete',
 'state',
 'union',
 'address',
 'present',
 'series',
 'u',
 'presidential',
 'state',
 'union',
 'address',
 'copyright',
 'law',
 'changing',
 'world',
 'sure',
 'check',
 'copyright',
 'law',
 'country',
 'downloading',
 'redistributing',
 'project',
 'gutenberg',
 'ebook',
 'header',
 'first']
```

In [14]:

```
## list the top 50 words by frequency
```

In [159]:

```
# Creating a frequency distribution of words
```

```
# creating a frequency distribution of words
fdist2Lemmasup2words = nltk.FreqDist(Lemmasup2words)
# print the top 50 tokens by frequency
fdist2Lemmasup2words.most_common(50)
```

Out[159]:

```
[('year', 2376),
 ('american', 1638),
 ('must', 1628),
 ('people', 1597),
 ('nation', 1506),
 ('world', 1490),
 ('new', 1441),
 ('america', 1288),
 ('congress', 1236),
 ('state', 1231),
 ('government', 1219),
 ('u', 1216),
 ('program', 1100),
 ('time', 871),
 ('country', 855),
 ('make', 853),
 ('one', 840),
 ('work', 834),
 ('need', 822),
 ('every', 780),
 ('federal', 744),
 ('help', 720),
 ('war', 702),
 ('million', 694),
 ('security', 689),
 ('tax', 685),
 ('job', 678),
 ('economic', 671),
 ('peace', 668),
 ('united', 651),
 ('also', 639),
 ('economy', 622),
 ('right', 619),
 ('national', 610),
 ('child', 609),
 ('great', 583),
 ('last', 574),
 ('many', 563),
 ('free', 559),
 ('let', 554),
 ('first', 553),
 ('would', 548),
 ('effort', 547),
 ('know', 536),
 ('budget', 531),
 ('system', 531),
 ('life', 526),
 ('family', 525),
 ('force', 518),
 ('freedom', 515)]
```

In [160]:

```
## list the top 50 bigrams by frequencies
bigrams2Lemmasup2words=list(nltk.bigrams(Lemmasup2words))
```

In [161]:

```
# create the bigram finder and score the bigrams by frequency
# scored is a list of bigram pairs with their score
finder2Lemmasup2words = BigramCollocationFinder.from_words(Lemmasup2words)
scored2Lemmasup2words = finder2Lemmasup2words.score_ngrams(bigram_measures.raw_freq)
scored2Lemmasup2words[:50]
```

Out[161]:

```
[(('united', 'state'), 0.002096959408857157),
 (('last', 'year'), 0.0012481901243197363),
```

```
 (('state', 'union'), 0.0012300346316023584),
 (('american', 'people'), 0.001098407309401368),
 (('fiscal', 'year'), 0.0008805413967928322),
 (('year', 'ago'), 0.0008669247772547988),
 (('federal', 'government'), 0.0008442304113580763),
 (('social', 'security'), 0.0008260749186406982),
 (('health', 'care'), 0.0008079194259233203),
 (('let', 'u'), 0.0008079194259233203),
 (('billion', 'dollar'), 0.0006944475964397079),
 (('union', 'address'), 0.0006445199914669184),
 (('united', 'nation'), 0.000612747879211507),
 (('million', 'dollar'), 0.0005991312596734735),
 (('soviet', 'union'), 0.0005673591474180619),
 (('next', 'year'), 0.000558281401059373),
 (('men', 'woman'), 0.0005128926692659281),
 (('past', 'year'), 0.0005083537960865835),
 (('free', 'world'), 0.0004947371765485501),
 (('member', 'congress'), 0.00045842619111379407),
 (('every', 'american'), 0.0004493484447551051),
 (('million', 'american'), 0.0004357318252170716),
 (('economic', 'growth'), 0.0004266540788583826),
 (('middle', 'east'), 0.0004130374593203491),
 (('state', 'local'), 0.00040849858614100463),
 (('free', 'nation'), 0.00040395971296166015),
 (('make', 'sure'), 0.00040395971296166015),
 (('four', 'year'), 0.00038126534706493765),
 (('first', 'time'), 0.00036764872752690416),
 (('small', 'business'), 0.00036764872752690416),
 (('ask', 'congress'), 0.0003585709811682152),
 (('world', 'war'), 0.0003585709811682152),
 (('armed', 'force'), 0.0003449543616301817),
 (('tax', 'cut'), 0.0003449543616301817),
 (('foreign', 'policy'), 0.0003358766152714927),
 (('must', 'continue'), 0.0003358766152714927),
 (('new', 'job'), 0.0003358766152714927),
 (('work', 'together'), 0.0003358766152714927),
 (('two', 'year'), 0.00030410450301608124),
 (('vice', 'president'), 0.00030410450301608124),
 (('local', 'government'), 0.00029956562983673676),
 (('around', 'world'), 0.00029048788347804774),
 (('national', 'security'), 0.00028594901029870327),
 (('must', 'also'), 0.00028141013711935873),
 (('address', 'january'), 0.0002723323907606698),
 (('human', 'right'), 0.0002723323907606698),
 (('health', 'insurance'), 0.00026325464440198076),
 (('nation', 'world'), 0.00026325464440198076),
 (('civil', 'right'), 0.0002587157712226363),
 (('fellow', 'citizen'), 0.0002587157712226363)]
```

In [162]:

```
# list the top 50 bigrams by their Mutual Information scores (using min frequency 5)
finder22Lemmasup2words = BigramCollocationFinder.from_words(Lemmasup2words)
finder22Lemmasup2words.apply_freq_filter(5)
scored22Lemmasup2words = finder22Lemmasup2words.score_ngrams(bigram_measures.pmi)
scored22Lemmasup2words[:50]
```

Out[162]:

```
[(('el', 'salvador'), 15.164271890106203),
 (('bin', 'laden'), 14.94187946876976),
 (('saudi', 'arabia'), 14.941879468769756),
 (('sam', 'rayburn'), 14.749234390827358),
 (('endowed', 'creator'), 14.316274983551253),
 (('jimmy', 'carter'), 14.289802772190063),
 (('northern', 'ireland'), 14.164271890106207),
 (('gerald', 'ford'), 14.097157694247668),
 (('iron', 'curtain'), 13.842343795218842),
 (('floor', 'appears'), 13.74923439082736),
 (('red', 'tape'), 13.74923439082736),
 (('jill', 'biden'), 13.678845062935963),
 (('thomas', 'jefferson'), 13.678845062935963),
 (('barack', 'obama'), 13.66177154957702),
 (('lyndon', 'johnson'), 13.66177154957702),
 (('teen', 'pregnancy'), 13.526841969490913),
 (('abraham', 'lincoln'), 13.49184654813471),
```

(('abraham', 'lincoln'), 13.[illegible]),
(('william', 'clinton'), 13.327770622389085),
(('ronald', 'reagan'), 13.289802772190063),
(('mom', 'dad'), 13.263807563657119),
(('greece', 'turkey'), 13.204913874603552),
(('elementary', 'secondary'), 13.122795254130045),
(('intercontinental', 'ballistic'), 13.003280013433898),
(('feeding', 'hungry'), 12.9678746773027),
(('grass', 'root'), 12.962638028936553),
(('lady', 'gentleman'), 12.941879468769756),
(('status', 'quo'), 12.891253395699788),
(('empowerment', 'zone'), 12.842343795218842),
(('nationwide', 'radio'), 12.801701810721497),
(('radio', 'television'), 12.749234390827361),
(('dwight', 'eisenhower'), 12.643439726804763),
(('al', 'qaeda'), 12.619951373882396),
(('al', 'qaida'), 12.619951373882394),
(('richard', 'nixon'), 12.602877860523451),
(('saddam', 'hussein'), 12.579309389385045),
(('introduced', 'thomas'), 12.563367845516026),
(('harry', 'truman'), 12.539781025198408),
(('prime', 'minister'), 12.501306877383774),
(('reported', 'floor'), 12.501306877383772),
(('persian', 'gulf'), 12.437176539964058),
(('carbon', 'pollution'), 12.427306295940001),
(('capitol', 'introduced'), 12.396717976106578),
(('george', 'bush'), 12.283853505858877),
(('synthetic', 'fuel'), 12.25417886245934),
(('panama', 'canal'), 12.176344722406778),
(('per', 'caput'), 12.105378201052636),
(('franklin', 'roosevelt'), 12.076809048855866),
(('chemical', 'biological'), 12.012268796661154),
(('baby', 'boom'), 11.983699644464382),
(('steam', 'coal'), 11.90123748427241)]