

CORPUS LINGUISTICS AND PYTHON PROGRAMMING

**Submitted By
Sanjana Rajagopala
SUID - 607219462**

INDEX

1. Description of State Union Addresses Dataset.....3

2. Analysis of State Union Addresses – Part 1.....4

3. Analysis of State Union Addresses - Part25

4. Comparison.....5

5. Appendix.....7

DESCRIPTION OF STATE UNION ADDRESSES DATASET

History

The dataset of the State of the Union Addresses is derived from the Project Gutenberg-tm which is a collection of electronic books. It was initiated by Professor Michael S. Hart who distributed and shared it for over forty years. The creation of these eBooks involved the efforts of several volunteers and is freely available for use. It is part of the non-profit organization Project Gutenberg Literary Archive Foundation. While the entire Project Gutenberg eBook was produced by Al Haines the two-individual works containing the Complete State of the Union Addresses from 1790 to 2016 were produced by James Linden.

Brief Description

The corpus mainly revolves around the concept of Presidential address and their plan for America in their respective tenure of power. Mostly, it deals with terms related to the progress so far and measures that government would take in the direction of employment, foreign policies, budget, military, health et al.

Naming Convention of the Files

As depicted at the end of the project, the Gutenberg eBook State Union Addresses is named 50950.txt or 50950.zip. The eBook containing the State Union Addresses is named Project Gutenberg eBook of Complete Addresses. The contents are provided in two files as – ‘state_union_part1’ and ‘state_union_part2’. The first file contains the Complete State of the Union Addresses from 1790 to 1860. Similarly, the second file contains the Complete State of the Union Addresses from 1946 to 2016. Both the files included various authors and was released as Edition 12 in February, 2004. They are named as per the initial date to ending date of the addresses included in them.

Number of Documents

The file “Complete State of the Union Addresses from 1790 to 1860” contains a total of 82 union addresses. The file “Complete State of the Union Addresses from 1946 to the Present” contains a total of 70 union addresses. In the Contents section, the name of the President, the type and the date of address are mentioned.

Related Policy

According to the stated license in the Project, the Project Gutenberg Literary Archive Foundation (PGAF) holds the compilation copyright. However, because it is not protected by U.S. copyright law it is free for use – download, copy, distribute, re-use without restriction anywhere in the United States; if located outside the U.S. the respective laws of the country apply.

ANALYSIS OF STATE UNION ADDRESSES: PART1

The analysis task of listing the top 50 words by frequency or the bigrams requires certain steps of pre-processing to ensure that we obtain a meaningful list. The common steps for both these listings are as follows –

Step 1 – Tokenize the input raw text file

To generate the list of individual tokens that can be used for further processing, as the first step we apply word tokenizer on the raw text.

Step 2 – Conversion of all the tokens into lower case tokens

Because the current analysis is to list the most frequently occurring words, it is better to convert into lower case so that the occurrence of every token is considered similarly irrespective of the capitalization.

Step 3 – Removal of punctuations and numbers

Considering the current analysis, the inclusion of punctuations and digits does not make sense. Hence, with the help of regular expressions these characters are removed.

Step 4 – Retention of words with hyphen in between

To handle the cases wherein the words with special characters are removed, another Regular expression is utilized so that such words are not eliminated.

Step 5 – Removal of Stop words in English

The inclusion of stop words in our analysis causes the top 50 to be filled with them and that does not add explicit interpretation. Hence, it is better to remove them. For the current analysis, the stop words list is retained as is and not modified. In case of the bigram list, we remove the candidates having both the words as stop words otherwise they are retained.

Step 6 – Lemmatization

To overcome the problem of counting the words such as ‘interest’ and ‘interests’ separately and including them as keys in the frequency table we make use of lemmatization. So that only the base core words are retained, and their forms can be stripped out. This is to ensure that families of derivationally related words with similar meanings can be counted together as one form in the frequency table.

Additionally, for the list of top 50 bigrams we need the following steps,

Step 7 – Removal of too short words

The candidates with both the words whose length is less than 3 are removed so that we can avoid the occurrence of common but unnecessary words for analysis. Example – ‘so’, ‘to’.

Step 8 – For bigrams by mutual information score,

We perform the frequency filter so that we maintain only those candidates whose minimum frequency is at least 5. Hence, ensuring that the words below the minimum threshold are removed.

ANALYSIS OF STATE UNION ADDRESSES: PART2

The steps of Pre-processing for the second text file remain the same as that of first text file. Precisely, the entire analysis is carried out in the same manner as that in the Step 1 except for the change in the names of the respective files.

COMPARISON

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

Based on the list of frequently occurring words

Although the two documents depict the similarity in terms of the theme of development of the United States, the topics of significance can be observed to be different. The usage of words such as 'nation', 'united states', 'government', 'congress' and 'state' prove the portion of similarity between the two documents. However, in the first document the heavy occurrence of 'commerce', 'law', 'treaty' represent the areas of interest undertaken or stressed by Presidents during those years. On the other hand, the second document represents 'security', 'tax', 'budget' as the areas of concentration. Also, first document shows that the addresses actions were mostly anticipatory or futuristic because of the high presence of verbs- 'may' and 'would'. The second document shows emphasis on compulsory actions in the form of verbs such as 'must', 'force'.

Based on the list of bigrams (Raw frequency) and list of Bigrams (PMI measure)

The respective lists of bigrams for the first and second documents upholds the ideas as illustrated above. Additionally, with the predominant occurrence of non-English words – (van, buren), (punta,arena) and locations – (san, francisco), (eastern, asia) etc it can prove the inclination towards locations and related measures. Conversely, in the second document there is comparatively more usage of person names (abraham, lincoln), (barack, obama), (ronald, reagan) hence, proving the references to past leaders in the union addresses.

The complexity of the texts in the two documents in terms of the used n-letter words remains largely similar. Approximately on an average, the number of letters used across the texts is four.

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

Yes, there are problems with the obtained list of words and bigrams.

In case of the bigram list,

1. The problem of having 'the' before most of the words makes it redundant - the best example is (united, states) and (the, united) which are obvious to be "the united states". Hence, counting them separately is redundant and unnecessary for current analysis.
2. The listing of ambiguous bigrams such as (ha, been) and (which, ha); (the, america) and (the, american) - whether reference is to – 'has' or 'had' or 'have' or some other word such as 'hatch'.

3. The lists of bigrams based on the Mutual information score are majorly filled with proper nouns and not particularly interesting for my current analysis.

In case of the word lists,

4. The occurrence of 'u' in the top 50 list presents ambiguity as to which word it exactly refers to - you or your or yours etc.
5. The occurrence of prepositions such as 'upon', 'within' in the top 50 list even after the removal of stop words. These do not greatly contribute to the analysis or predictions.
6. Despite the applied lemmatization on each of the documents, the listing of 'economy' and 'economical' not only presents the problem of redundancy but also difficulty in understanding the usage of complex or diverse vocabulary in the document.

To handle these and generate more elegant bigrams list,

1. We can recursively grow the Stopwords list. This avoids the counting of non-essential words with only commonly used verbs, prepositions or conjunctions and hence, concentrating on the actual aspects depicted by the other words.
2. Decide whether the retaining the bigrams with 'the' in the beginning are required for your current analysis so that counting of (the, united) and (united, states) separately is avoided.
3. Decision about the preferred method of lemmatization by trial and error to ensure that most of the canonical forms of words and their corresponding inflections are not treated separately and increase the redundancy.
4. Performing POS – Part of Speech tagging will help in reducing the density of proper nouns in the bigrams lists.

3. How are the top 50 bigrams by frequency different from the top 50 bigrams scored by Mutual Information?

The top 50 bigrams by frequency are different from that scored by Mutual Information in terms of the listing of words that often appear together versus how often they appear separately. As a result, in the list of bigrams by frequency, the phrases with terms appearing together a lot and mostly with determiners such as 'the', 'a' as the first term can be observed. Moreover, the bigrams with no distance meaning zero tokens in between them are enlisted here. Whereas in the list of bigrams by mutual information the names of people, places, lakes are predominantly observed. Apart from this, the non-substitutional consecutive words such as ('prime', 'minister'), ('rocky', 'mountain'), ('ballot', 'box') can also be seen in this list scored by PMI.

APPENDIX

RESULTS FOR PART-1

1. The top 50 words by frequency (Normalized by the length of the document)

[('state', 1.5771089125478466), ('government', 1.0713463992135372), ('united', 0.7797694994666277), ('may', 0.6534334539521848), ('congress', 0.6279152460833735), ('upon', 0.6086720073626305), ('country', 0.5969587316195696), ('would', 0.5777154928988266), ('public', 0.5752055052395992), ('power', 0.48484594950741494), ('year', 0.47898931163588443), ('great', 0.4488694597251564), ('made', 0.4438494844067017), ('duty', 0.4254429082390345), ('law', 0.4053630069652157), ('time', 0.3823547867556318), ('last', 0.3810997929260181), ('war', 0.37566148633102553), ('interest', 0.3744064925014119), ('subject', 0.3714781735656466), ('present', 0.35641824761028257), ('nation', 0.347214959526449), ('act', 0.3375933401660775), ('citizen', 0.3350833525068502), ('people', 0.3288083833587818), ('treaty', 0.312911794850342), ('part', 0.30998347591457676), ('shall', 0.2903219059172959), ('without', 0.2773536363446213), ('union', 0.2689870108138635), ('right', 0.2660586918780983), ('one', 0.2652220293250225), ('general', 0.25309042230542367), ('mexico', 0.25309042230542367), ('treasury', 0.2476521157104311), ('every', 0.2468154531573553), ('necessary', 0.24472379677466585), ('constitution', 0.2384488276265975), ('territory', 0.23510217741429437), ('new', 0.22924553954276392), ('object', 0.2225522391181577), ('foreign', 0.2171139325231651), ('measure', 0.21418561358739988), ('two', 0.21334895103432408), ('system', 0.21293061975778618), ('commerce', 0.2116756259281725), ('peace', 0.20958396954548308), ('consideration', 0.20623731933317996), ('within', 0.20038068146164947), ('service', 0.20038068146164947)]

2. The top 50 Bigrams by Frequency

[(('united', 'state'), 0.0032673822765571145), (('the', 'united'), 0.0032422899277519584), (('ha', 'been'), 0.002213503626740558), (('the', 'public'), 0.0016417565361087861), (('the', 'government'), 0.001589779527869534), (('may', 'be'), 0.0011936788788738553), (('of', 'congress'), 0.0011596249769240006), (('the', 'state'), 0.001138117249376724), (('state', 'of'), 0.0010879325517664116), (('upon', 'the'), 0.001021617058495642), (('the', 'present'), 0.0009965247096904858), (('part', 'of'), 0.0009266245951618367), (('the', 'people'), 0.0009230399739039572), (('the', 'country'), 0.0008871937613251628), (('the', 'union'), 0.0008674783444068258), (('the', 'treasury'), 0.0008477629274884888), (('act', 'of'), 0.0008137090255386342), (('would', 'be'), 0.0007796551235887794), (('the', 'last'), 0.0007760705023309), (('the', 'constitution'), 0.0007653166385572617), (('the', 'subject'), 0.0006380625839025413), (('the', 'two'), 0.0006326856520157221), (('the', 'act'), 0.0006255164094999632), (('the', 'law'), 0.0006219317882420838), (('the', 'secretary'), 0.0006093856138395057), (('secretary', 'of'), 0.0005914625075501086), (('state', 'and'), 0.0005896701969211688), (('the', 'treaty'), 0.0005681624693738922), (('government', 'of'), 0.0005663701587449524), (('the', 'whole'), 0.0005556162949713141), (('the', 'great'), 0.0005520316737134346), (('interest', 'of'), 0.0005359008780529771), (('the', 'general'), 0.0005341085674240374), (('portion', 'of'), 0.000530523946166158), (('the', 'first'), 0.0005251470142793388), (('be', 'made'), 0.0005090162186188813), (('our', 'citizen'), 0.0005054315973610018), (('great', 'britain'), 0.000491093112329484), (('of', 'war'), 0.000491093112329484), (('shall', 'be'), 0.0004660007635243279), (('the', 'power'), 0.0004642084528953882), (('condition', 'of'), 0.00046241614226644846), (('the', 'war'),

0.0004498699678638704), (('within', 'the'), 0.0004498699678638704), (('the', 'right'), 0.0004409084147191718), (('last', 'session'), 0.00043732379346129237), (('been', 'made'), 0.0004229853084297746), (('right', 'of'), 0.00041223144465613623), (('the', 'part'), 0.0003996852702535582), (('which', 'ha'), 0.0003996852702535582)]

3. The top 50 bigrams by Mutual Information Scores

[('bona', 'fide'), 16.767819779008715), (('del', 'norte'), 16.50478537317492), (('millard', 'fillmore'), 16.50478537317492), (('punta', 'arena'), 16.50478537317492), (('ballot', 'box'), 16.28239295183847), (('guadalupe', 'hidalgo'), 15.919822872453764), (('porto', 'rico'), 15.919822872453764), (('franklin', 'pierce'), 15.767819779008715), (('la', 'plata'), 15.630316255258778), (('vera', 'cruz'), 15.504785373174922), (('entangling', 'alliance'), 15.282392951838471), (('costa', 'rica'), 15.089747873896076), (('nucleus', 'around'), 15.089747873896076), (('santa', 'anna'), 15.002285032645737), (('santa', 'fe'), 15.002285032645737), (('van', 'buren'), 15.002285032645737), (('sublime', 'porte'), 14.96046485695111), (('martin', 'van'), 14.832360031203425), (('ad', 'valorem'), 14.76781977900871), (('quincy', 'adam'), 14.63031625525878), (('buenos', 'ayres'), 14.50478537317492), (('de', 'facto'), 14.356393533282247), (('project', 'gutenberg'), 14.334860371732606), (('gun', 'boat'), 14.219383154312672), (('andrew', 'jackson'), 14.199930791646498), (('retired', 'list'), 14.08075909066882), (('circulating', 'medium'), 14.045353754537622), (('rocky', 'mountain'), 14.045353754537622), (('john', 'quincy'), 14.002285032645739), (('thomas', 'jefferson'), 13.914822191395396), (('precious', 'metal'), 13.844743755659255), (('almighty', 'god'), 13.832360031203425), (('john', 'tyler'), 13.832360031203425), (('san', 'francisco'), 13.80434565503383), (('san', 'jacinto'), 13.804345655033828), (('san', 'juan'), 13.804345655033828), (('rio', 'grande'), 13.663483119193975), (('inferior', 'quality'), 13.318918827863586), (('cut', 'off'), 13.308388160371416), (('james', 'buchanan'), 13.256857859731333), (('predatory', 'incursion'), 13.187674294585332), (('hudson', 'bay'), 13.138463158929104), (('water', 'witch'), 13.045353754537622), (('council', 'bluff'), 13.019358546004678), (('lake', 'erie'), 12.967844387316708), (('topographical', 'engineer'), 12.96046485695111), (('posse', 'comitatus'), 12.919822872453762), (('eastern', 'asia'), 12.914822191395396), (('argentine', 'confederation'), 12.907850230787687), (('catholic', 'majesty'), 12.822961333201174)]

RESULTS FOR PART-2

1. The top 50 words by frequency (Normalized by the length of the document)

[('year', 1.0699859046469633), ('american', 0.737641797900558), ('must', 0.7331384902210674), ('people', 0.7191782364146465), ('nation', 0.6781981365312822), ('world', 0.6709928442440973), ('new', 0.6489266366145935), ('america', 0.5800260291183874), ('congress', 0.5566088291850364), ('state', 0.554357175345291), ('government', 0.5489532061299024), ('u', 0.5476022138260552), ('program', 0.49536384474396444), ('time', 0.39223809888363004), ('country', 0.3850328065964451), ('make', 0.38413214506054694), ('one', 0.3782778450772092), ('work', 0.37557586046951486), ('need', 0.37017189125412614), ('every', 0.3512579990002657), ('federal', 0.3350460913540996), ('help', 0.3242381529233222), ('war', 0.3161321991002391), ('million', 0.31252955295664664), ('security', 0.31027789911690135), ('tax', 0.30847657604510514), ('job', 0.3053242606694617), ('economic', 0.3021719452938183), ('peace', 0.30082095298997114), ('united', 0.2931653299348371), ('also', 0.2877613607194484), ('economy', 0.2801057376643144), ('right', 0.27875474536046724), ('national', 0.27470176844892574), ('child', 0.27425143768097665), ('great', 0.26254283771430115), ('last', 0.2584898608027596), ('many', 0.25353622235532003), ('free', 0.25173489928352377), ('let', 0.24948324544377845), ('first', 0.2490329146758294), ('would', 0.24678126083608412), ('effort', 0.24633093006813503), ('know', 0.24137729162069538), ('budget', 0.2391256377809501), ('system', 0.2391256377809501), ('life', 0.2368739839412048), ('family', 0.23642365317325578), ('force', 0.23327133779761236), ('freedom', 0.23192034549376514)]

2. The top 50 bigrams by frequency

[(('we', 'must'), 0.0019082195939457397), (('the', 'world'), 0.0017202888763601745), (('the', 'congress'), 0.0013568184775133668), (('the', 'united'), 0.0011895394871350065), (('united', 'state'), 0.0009541097969728698), (('ha', 'been'), 0.0008363949518918015), (('state', 'of'), 0.0007909611520359505), (('the', 'american'), 0.000764113906666584), (('must', 'be'), 0.0006773766523963232), (('to', 'make'), 0.0006711811342341616), (('the', 'people'), 0.0006587900979098387), (('to', 'help'), 0.0006525945797476772), (('the', 'federal'), 0.0005865090526846212), (('the', 'first'), 0.0005844438799639008), (('this', 'year'), 0.0005555314618738138), (('the', 'past'), 0.0005390100801080498), (('continue', 'to'), 0.0005245538710630063), (('the', 'next'), 0.0005183583529008448), (('the', 'nation'), 0.0004977066256936399), (('the', 'union'), 0.0004977066256936399), (('american', 'people'), 0.0004935762802521989), (('the', 'state'), 0.0004791200712071554), (('last', 'year'), 0.0004646638621621119), (('our', 'nation'), 0.0004481424803963479), (('the', 'future'), 0.000439881789513466), (('need', 'to'), 0.00043781661679274546), (('our', 'country'), 0.00043781661679274546), (('part', 'of'), 0.00043368627135130446), (('the', 'soviet'), 0.000431621098630584), (('one', 'of'), 0.00042955592590986346), (('our', 'people'), 0.00042955592590986346), (('congress', 'to'), 0.00041716488958554046), (('want', 'to'), 0.00041303454414409947), (('fiscal', 'year'), 0.0004006435078197765), (('of', 'american'), 0.0004006435078197765), (('to', 'work'), 0.000398578335099056), (('year', 'ago'), 0.000394447989657615), (('of', 'america'), 0.000386187298774733), (('to', 'meet'), 0.000386187298774733), (('federal', 'government'), 0.00038412212605401253), (('the', 'government'), 0.00038412212605401253), (('our', 'economy'), 0.000382056953333292), (('member',


'of'), 0.00037379626245041006), (('social', 'security'), 0.00037379626245041006), (('health', 'care'), 0.00036760074428824854), (('the', 'last'), 0.00036140522612608707), (('we', 'need'), 0.00035727488068464607), (('at', 'home'), 0.00035520970796392554), (('the', 'new'), 0.00035520970796392554), (('effort', 'to'), 0.00035314453524320507)]

3. The top 50 Bigrams by Mutual Information Scores

[('el', 'salvador'), 16.30034362211258), (('bin', 'laden'), 16.077951200776134), (('saudi', 'arabia'), 16.07795120077613), (('sam', 'rayburn'), 15.62227171699994), (('jimmy', 'carter'), 15.425874504196438), (('northern', 'ireland'), 15.300343622112582), (('iron', 'curtain'), 14.97841552722522), (('floor', 'appears'), 14.885306122833738), (('red', 'tape'), 14.885306122833738), (('jill', 'biden'), 14.814916794942338), (('thomas', 'jefferson'), 14.814916794942338), (('barack', 'obama'), 14.797843281583395), (('teen', 'pregnancy'), 14.662913701497288), (('abraham', 'lincoln'), 14.627918280141085), (('ronald', 'reagan'), 14.425874504196438), (('small-business', 'owner'), 14.21545472452607), (('intercontinental', 'ballistic'), 14.139351745440276), (('grass', 'root'), 14.098709760942928), (('status', 'quo'), 14.027325127706163), (('empowerment', 'zone'), 13.97841552722522), (('nationwide', 'radio'), 13.937773542727872), (('al', 'qaeda'), 13.75602310588877), (('al', 'qaida'), 13.756023105888769), (('richard', 'nixon'), 13.73894959252983), (('line-item', 'veto'), 13.675852757204785), (('saddam', 'hussein'), 13.66647152091048), (('prime', 'minister'), 13.637378609390149), (('persian', 'gulf'), 13.573248271970437), (('carbon', 'pollution'), 13.563378027946376), (('synthetic', 'fuel'), 13.390250594465716), (('panama', 'canal'), 13.312416454413153), (('per', 'caput'), 13.24144993305901), (('baby', 'boom'), 13.119771376470757), (('steam', 'coal'), 13.037309216278786), (('catastrophic', 'illness'), 13.030882947119355), (('franklin', 'roosevelt'), 13.020235702919846), (('hardest', 'hit'), 12.978415527225218), (('supreme', 'court'), 12.792548981913884), (('greenhouse', 'gas'), 12.780270652135448), (('distinguished', 'guest'), 12.673560945696796), (('pell', 'grant'), 12.637378609390149), (('honored', 'guest'), 12.624778572610516), (('river', 'basin'), 12.599903903971487), (('indian', 'ocean'), 12.563378027946374), (('collective', 'bargaining'), 12.563378027946372), (('mass', 'transit'), 12.518983908587922), (('ballistic', 'missile'), 12.44236262698501), (('rural', 'electrification'), 12.425874504196441), (('mental', 'illness'), 12.364883874307289), (('north', 'carolina'), 12.315450514502787)]


SCREENSHOTS OF PYTHON PROCESSING

1. The top 50 words by frequency – Part 1

 jupyter











assignment_1_a

Last Checkpoint: Last Tuesday at 1:22 AM (autosaved)

 Logout

FileEditViewInsertCellKernelWidgetsHelp

TrustedPython 3



Code

In [10]:

```
from operator import itemgetter
freq_dist_1 = FreqDist(state_revised_lemma)
len(freq_dist_1)

#Finding the List of top 50 frequently occurring words normalized by the length of document
freqWords = []
for key in freq_dist_1:
    freqWords.append((key, freq_dist_1.freq(key)*100))

#Sort the List in descending order
freqWords = sorted(freqWords, key = itemgetter(1), reverse=True)

#Print the words
print(freqWords[:50])
```

[('state', 1.5771089125478466), ('government', 1.0713463992135372), ('united', 0.7797694994666277), ('may', 0.6534334539521848), ('congress', 0.6279152460833735), ('upon', 0.6086720073626305), ('country', 0.5969587316195696), ('would', 0.5777154928988266), ('public', 0.5752055052395992), ('power', 0.48484594950741494), ('year', 0.47898931163588443), ('great', 0.4488694597251564), ('made', 0.44384948444067017), ('duty', 0.4254429082390345), ('law', 0.4053630069652157), ('time', 0.3823547867556318), ('last', 0.3810997929260181), ('war', 0.37566148633102553), ('interest', 0.3744064925014119), ('subject', 0.3714781735656466), ('present', 0.35641824761028257), ('nation', 0.347214959526449), ('act', 0.3375933401660775), ('citizen', 0.3350833525068502), ('people', 0.3288083833587818), ('treaty', 0.312911794850342), ('part', 0.30998347591457676), ('shall', 0.2903219059172959), ('with out', 0.2773536363446213), ('union', 0.2689870108138635), ('right', 0.2660586918780983), ('one', 0.2652220293250225), ('general', 0.25309042230542367), ('mexico', 0.25309042230542367), ('treasury', 0.2476521157104311), ('every', 0.2468154531573553), ('necessary', 0.24472379677466585), ('constitution', 0.2384488276265975), ('territory', 0.23510217741429437), ('new', 0.22924553954276392), ('object', 0.2225522391181577), ('foreign', 0.2171139325231651), ('measure', 0.21418561358739988), ('two', 0.21334895103432408), ('system', 0.21293061975778618), ('commerce', 0.2116756259281725), ('peace', 0.20958396954548308), ('consideration', 0.20623731933317996), ('within', 0.20038068146164947), ('service', 0.20038068146164947)]

2. The top 50 words by frequency – Part2

```
jupyter assignment_1_a Last Checkpoint: Last Tuesday at 1:22 AM (autosaved)
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3
In [11]: #Similarly for the second document
freq_dist_2 = FreqDist(state_revised_lemma_2)
len(freq_dist_2)

#Finding the list of top 50 frequently occurring words normalized by the length of document
freqWords_2 = []
for key in freq_dist_2:
    freqWords_2.append((key,freq_dist_2.freq(key)*100))

#Sort the list in descending order
freqWords_2 = sorted(freqWords_2, key = itemgetter(1),reverse=True)

#Print the words
print(freqWords_2[:50])

[('year', 1.0699859046469633), ('american', 0.737641797900558), ('must', 0.7331384902210674), ('people', 0.7191782364146465),
('nation', 0.6781981365312822), ('world', 0.6709928442440973), ('new', 0.6489266366145935), ('america', 0.5800260291183874),
('congress', 0.5566088291850364), ('state', 0.554357175345291), ('government', 0.5489532061299024), ('u', 0.5476022138260552),
('program', 0.49536384474396444), ('time', 0.39223809888363004), ('country', 0.3850328065964451), ('make', 0.3841321450605469
4), ('one', 0.3782778450772092), ('work', 0.37557586046951486), ('need', 0.37017189125412614), ('every', 0.3512579990002657),
('federal', 0.3350460913540996), ('help', 0.3242381529233222), ('war', 0.3161321991002391), ('million', 0.31252955295664664),
('security', 0.31027789911690135), ('tax', 0.30847657604510514), ('job', 0.3053242606694617), ('economic', 0.3021719452938183),
('peace', 0.30082095298997114), ('united', 0.2931653299348371), ('also', 0.2877613607194484), ('economy', 0.2801057376643144),
('right', 0.27875474536046724), ('national', 0.27470176844892574), ('child', 0.27425143768097665), ('great', 0.2625428377143011
5), ('last', 0.2584898608027596), ('many', 0.25353622235532003), ('free', 0.25173489928352377), ('let', 0.24948324544377845),
('first', 0.2490329146758294), ('would', 0.24678126083608412), ('effort', 0.24633093006813503), ('know', 0.24137729162069538),
('budget', 0.2391256377809501), ('system', 0.2391256377809501), ('life', 0.2368739839412048), ('family', 0.23642365317325578),
('force', 0.23327133779761236), ('freedom', 0.23192034549376514)]
```

3. The top 50 bigrams by frequency – Part 1

```
jupyter assignment_1_a Last Checkpoint: Last Tuesday at 1:22 AM (unsaved changes)
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3
state_finder_2.apply_ngram_filter(lambda w1,w2: len(w1)<3 and len(w2)<3)


#Step 3 - Remove only if both the words in the bigram are stop words

state_finder_1.apply_ngram_filter(lambda w1,w2: w1 in stopwords and w2 in stopwords)
state_finder_2.apply_ngram_filter(lambda w1,w2:w1 in stopwords and w2 in stopwords)

In [14]: bigram_measures = nltk.collocations.BigramAssocMeasures()
score_raw_1 = state_finder_1.score_ngrams(bigram_measures.raw_freq)
print(score_raw_1[:50])










[[('united', 'state'), 0.0032673822765571145], (('the', 'united'), 0.0032422899277519584), (('ha', 'been'), 0.00221350362674055
8), (('the', 'public'), 0.0016417565361087861), (('the', 'government'), 0.001589779527869534), (('may', 'be'), 0.00119367887887
38553), (('of', 'congress'), 0.0011596249769240006), (('the', 'state'), 0.001138117249376724), (('state', 'of'), 0.001087932551
7664116), (('upon', 'the'), 0.001021617058495642), (('the', 'present'), 0.0009965247096904858), (('part', 'of'), 0.000926624595
1618367), (('the', 'people'), 0.0009230399739039572), (('the', 'country'), 0.0008871937613251628), (('the', 'union'), 0.0008674
783444068258), (('the', 'treasury'), 0.0008477629274884888), (('act', 'of'), 0.0008137090255386342), (('would', 'be'), 0.000779
6551235887794), (('the', 'last'), 0.0007760705023309), (('the', 'constitution'), 0.0007653166385572617), (('the', 'subject'),
0.0006380625839025413), (('the', 'two'), 0.0006326856520157221), (('the', 'act'), 0.0006255164094999632), (('the', 'law'), 0.00
06219317882420838), (('the', 'secretary'), 0.0006093856138395057), (('secretary', 'of'), 0.0005914625075501086), (('state', 'an
d'), 0.0005896701969211688), (('the', 'treaty'), 0.0005681624693738922), (('government', 'of'), 0.0005663701587449524), (('th
e', 'whole'), 0.0005556162949713141), (('the', 'great'), 0.0005520316737134346), (('interest', 'of'), 0.0005359008780529771),
(('the', 'general'), 0.0005341085674240374), (('portion', 'of'), 0.000530523946166158), (('the', 'first'), 0.000525147014279338
8), (('be', 'made'), 0.0005090162186188813), (('our', 'citizen'), 0.0005054315973610018), (('great', 'britain'), 0.000491093112
329484), (('of', 'war'), 0.000491093112329484), (('shall', 'be'), 0.0004660007635243279), (('the', 'power'), 0.0004642084528953
882), (('condition', 'of'), 0.00046241614226644846), (('the', 'war'), 0.0004498699678638704), (('within', 'the'), 0.00044986996
78638704), (('the', 'right'), 0.0004409084147191718), (('last', 'session'), 0.00043732379346129237), (('been', 'made'), 0.00042
29853084297746), (('right', 'of'), 0.00041223144465613623), (('the', 'part'), 0.0003996852702535582), (('which', 'ha'), 0.00039
96852702535582)]
```

4. The top 50 bigrams by frequency – Part 2

 assignment_1_a Last Checkpoint: Last Tuesday at 1:22 AM (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help

Trusted Python 3



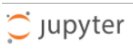
Code

```
In [15]: score_raw_2 = state_finder_2.score_ngrams(bigram_measures.raw_freq)
print(score_raw_2[:50])

[('we', 'must'), 0.0019082195939457397), (('the', 'world'), 0.0017202888763601745), (('the', 'congress'), 0.001356818477513366
8), (('the', 'united'), 0.0011895394871350065), (('united', 'state'), 0.0009541097969728698), (('ha', 'been'), 0.00083639495189
18015), (('state', 'of'), 0.0007909611520359505), (('the', 'american'), 0.000764113906666584), (('must', 'be'), 0.0006773766523
963232), (('to', 'make'), 0.0006711811342341616), (('the', 'people'), 0.0006587900979098387), (('to', 'help'), 0.00065259457974
76772), (('the', 'federal'), 0.0005865090526846212), (('the', 'first'), 0.0005844438799639008), (('this', 'year'), 0.0005555314
618738138), (('the', 'past'), 0.0005390100801080498), (('continue', 'to'), 0.0005245538710630063), (('the', 'next'), 0.00051835
83529008448), (('the', 'nation'), 0.0004977066256936399), (('the', 'union'), 0.0004977066256936399), (('american', 'people'),
0.0004935762802521989), (('the', 'state'), 0.0004791200712071554), (('last', 'year'), 0.0004646638621621119), (('our', 'natio
n'), 0.0004481424803963479), (('the', 'future'), 0.000439881789513466), (('need', 'to'), 0.00043781661679274546), (('our', 'cou
ntry'), 0.00043781661679274546), (('part', 'of'), 0.00043368627135130446), (('the', 'soviet'), 0.000431621098630584), (('one',
'of'), 0.00042955592590986346), (('our', 'people'), 0.00042955592590986346), (('congress', 'to'), 0.00041716488958554046), (('w
ant', 'to'), 0.00041303454414409947), (('fiscal', 'year'), 0.0004006435078197765), (('of', 'american'), 0.0004006435078197765),
(('to', 'work'), 0.000398578335099056), (('year', 'ago'), 0.000394447989657615), (('of', 'america'), 0.000386187298774733),
(('to', 'meet'), 0.000386187298774733), (('federal', 'government'), 0.00038412212605401253), (('the', 'government'), 0.00038412
212605401253), (('our', 'economy'), 0.000382056953333292), (('member', 'of'), 0.00037379626245041006), (('social', 'security'),
0.00037379626245041006), (('health', 'care'), 0.00036760074428824854), (('the', 'last'), 0.00036140522612608707), (('we', 'nee
d'), 0.00035727488068464607), (('at', 'home'), 0.00035520970796392554), (('the', 'new'), 0.00035520970796392554), (('effort',
'to'), 0.00035314453524320507)]
```










```
In [12]: #Step 5 - Filtering with the minimum frequency of 5
```

5. The top 50 bigrams by Mutual Information – Part 1

 assignment_1_a Last Checkpoint: Last Tuesday at 1:22 AM (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Trusted Python 3



Code

```
212605401253), (('our', 'economy'), 0.000382056953333292), (('member', 'of'), 0.00037379626245041006), (('social', 'security'),
0.00037379626245041006), (('health', 'care'), 0.00036760074428824854), (('the', 'last'), 0.00036140522612608707), (('we', 'nee
d'), 0.00035727488068464607), (('at', 'home'), 0.00035520970796392554), (('the', 'new'), 0.00035520970796392554), (('effort',
'to'), 0.00035314453524320507)]

In [16]: #Step 5 - Filtering with the minimum frequency of 5
state_finder_1.apply_freq_filter(5)
state_finder_2.apply_freq_filter(5)

#Listing the top 50 bigrams by Mutual Information scores
score_pmi_1 = state_finder_1.score_ngrams(bigram_measures.pmi)
print(score_pmi_1[:50])

[('bona', 'fide'), 16.767819779008715), (('del', 'norte'), 16.50478537317492), (('millard', 'fillmore'), 16.50478537317492),
(('punta', 'arena'), 16.50478537317492), (('ballot', 'box'), 16.28239295183847), (('guadalupe', 'hidalgos'), 15.91982287245376
4), (('porto', 'rico'), 15.919822872453764), (('franklin', 'pierce'), 15.767819779008715), (('la', 'plata'), 15.63031625525877
8), (('vera', 'cruz'), 15.504785373174922), (('entangling', 'alliance'), 15.282392951838471), (('costa', 'rica'), 15.0897478738
96076), (('nucleus', 'around'), 15.089747873896076), (('santa', 'anna'), 15.002285032645737), (('santa', 'fe'), 15.002285032645
737), (('van', 'buren'), 15.002285032645737), (('sublime', 'porte'), 14.96046485695111), (('martin', 'van'), 14.83236003120342
5), (('ad', 'valorem'), 14.76781977900871), (('quincy', 'adam'), 14.63031625525878), (('buenos', 'ayres'), 14.50478537317492),
(('de', 'facto'), 14.356393533282247), (('project', 'guttenberg'), 14.334860371732606), (('gun', 'boat'), 14.219383154312672),
(('andrew', 'jackson'), 14.199930791646498), (('retired', 'list'), 14.08075909066882), (('circulating', 'medium'), 14.045353754
537622), (('rocky', 'mountain'), 14.045353754537622), (('john', 'quincy'), 14.002285032645739), (('thomas', 'jefferson'), 13.91
4822191395396), (('precious', 'metal'), 13.844743755659255), (('almighty', 'god'), 13.832360031203425), (('john', 'tyler'), 13.
832360031203425), (('san', 'francisco'), 13.80434565503383), (('san', 'jacinto'), 13.804345655033828), (('san', 'juan'), 13.804
345655033828), (('rio', 'grande'), 13.663483119193975), (('inferior', 'quality'), 13.318918827863586), (('cut', 'off'), 13.3083
88160371416), (('james', 'buchanan'), 13.256857859731333), (('predatory', 'incursion'), 13.187674294585332), (('hudson', 'ba
y'), 13.138463158929104), (('water', 'witch'), 13.045353754537622), (('council', 'bluff'), 13.019358546004678), (('lake', 'eri
e'), 12.967844387316708), (('topographical', 'engineer'), 12.96046485695111), (('posse', 'comitatus'), 12.919822872453762),
(('eastern', 'asia'), 12.914822191395396), (('argentine', 'confederation'), 12.907850230787687), (('catholic', 'majesty'), 12.8
22961333201174)]
```

6. The top 50 bigrams by Mutual Information – Part 2

```
jupyter assignment_1_a Last Checkpoint: Last Tuesday at 1:22 AM (unsaved changes) Logout
```

```
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3
```

```
88160371416), (('james', 'buchanan'), 13.256857859731333), (('predatory', 'incursion'), 13.187674294585332), (('hudson', 'ba  
y'), 13.138463158929104), (('water', 'witch'), 13.045353754537622), (('council', 'bluff'), 13.019358546004678), (('lake', 'eri  
e'), 12.967844387316708), (('topographical', 'engineer'), 12.96046485695111), (('posse', 'comitatus'), 12.919822872453762),  
 (('eastern', 'asia'), 12.914822191395396), (('argentine', 'confederation'), 12.907850230787687), (('catholic', 'majesty'), 12.8  
22961333201174)]
```

```
In [17]: score_pmi_2 = state_finder_2.score_ngrams(bigram_measures.pmi)  
print(score_pmi_2[:50])
```

```
[ (('el', 'salvador'), 16.30034362211258), (('bin', 'laden'), 16.077951200776134), (('saudi', 'arabia'), 16.07795120077613),  
 (('sam', 'rayburn'), 15.62227171699994), (('jimmy', 'carter'), 15.425874504196438), (('northern', 'ireland'), 15.30034362211258  
2), (('iron', 'curtain'), 14.97841552722522), (('floor', 'appears'), 14.885306122833738), (('red', 'tape'), 14.88530612283373  
8), (('jill', 'biden'), 14.814916794942338), (('thomas', 'jefferson'), 14.814916794942338), (('barack', 'obama'), 14.7978432815  
83395), (('teen', 'pregnancy'), 14.662913701497288), (('abraham', 'lincoln'), 14.627918280141085), (('ronald', 'reagan'), 14.42  
5874504196438), (('small-business', 'owner'), 14.21545472452607), (('intercontinental', 'ballistic'), 14.139351745440276), (('g  
rass', 'root'), 14.098709760942928), (('status', 'quo'), 14.027325127706163), (('empowerment', 'zone'), 13.97841552722522),  
 (('nationwide', 'radio'), 13.937773542727872), (('al', 'qaeda'), 13.75602310588877), (('al', 'qaida'), 13.756023105888769),  
 (('richard', 'nixon'), 13.73894959252983), (('line-item', 'veto'), 13.675852757204785), (('saddam', 'hussein'), 13.666471520910  
48), (('prime', 'minister'), 13.637378609390149), (('persian', 'gulf'), 13.573248271970437), (('carbon', 'pollution'), 13.56337  
8027946376), (('synthetic', 'fuel'), 13.390250594465716), (('panama', 'canal'), 13.312416454413153), (('per', 'caput'), 13.2414  
4993305901), (('baby', 'boom'), 13.119771376470757), (('steam', 'coal'), 13.037309216278786), (('catastrophic', 'illness'), 13.  
030882947119355), (('franklin', 'roosevelt'), 13.020235702919846), (('hardest', 'hit'), 12.978415527225218), (('supreme', 'cour  
t'), 12.792548981913884), (('greenhouse', 'gas'), 12.780270652135448), (('distinguished', 'guest'), 12.673560945696796), (('pel  
l', 'grant'), 12.637378609390149), (('honored', 'guest'), 12.624778572610516), (('river', 'basin'), 12.599903903971487), (('ind  
ian', 'ocean'), 12.563378027946374), (('collective', 'bargaining'), 12.563378027946372), (('mass', 'transit'), 12.5189839085879  
22), (('ballistic', 'missile'), 12.44236262698501), (('rural', 'electrification'), 12.425874504196441), (('mental', 'illness'),  
12.364883874307289), (('north', 'carolina'), 12.315450514502787)]
```