IST 664
Natural Language Processing

Assignment:
Homework 1

Rashmitha Varma Pandati
SUID: 622666081

Table of Contents

CORPUS STATISTICS AND PYTHON PROGRAMMING

1. **Introduction:**
   Corpus Statistics and Python Programming document describe the analysis of State of the Union Addresses dataset which is a collection of annual speeches delivered by the presidents of the United States. The final objective of this analysis is to find the similarities or dissimilarities of the most common words and bigrams in the content of the two files.

2. **Descriptive Analysis of State of the Union Addresses dataset**
   The analysis of State of the Union Addresses is based on three documents which are:
   - state_union_part1.txt
   - state_union_part2.txt
   - state_union_policy.txt

   History:
   These files are a part of Project Gutenberg. It is an open-source e-book by volunteers from all parts of the world helping in digitizing the famous books of all centuries. The content from Gutenberg corpus can be accessed via their official link, i.e., https://www.gutenberg.org where the text files can be read in using a Plain text corpus reader for further processing and analysis. The State of the Union Addresses dataset is a collection of annual speeches delivered by the presidents of the United States, from George Washington to Barack Obama, to a joint session of the United States Congress for the span of 1790-2016. The modified version of this dataset contains three files:

   - **The addresses delivered between 1790 and 1860**
     The "state_union_part1.txt" document contains presidential addresses to the society spoken by each president from year 1790 - 1860. This document was released in Feb' 2004.
   - **The address delivered between 1946 and 2016**
     The "state_union_part2.txt" document contains presidential addresses to the society spoken by each president from year 1946 - 2016. This document was released after 2016.
   - **A policy description of the Project Gutenberg Ebook**
     The "state_union_policy.txt" document contains policies regarding its usage, copyrights, licensing, redistributing etc. These are produced by "James Linden".

   Naming Conventions:
   The files have been named in the format 'state_union_∗.txt', where '∗' varies depending on the type of file it is. The contents of "state_union_part1.txt" and "state_union_part2.txt" are separated by "***" to denote the different addresses from different year.

   Number of documents contained in the corpus:
   There are 72 speeches addressed in the 'state_union_part1.txt' document and 70 speeches addressed in the 'state_union_part2.txt' document.

Related policies:
There are 5 Sections in the Policies and they are as below,

- Section 1: General Terms of Use and Redistributing Project Gutenberg-tm electronic works
- Section 2: Information about the Mission of Project Gutenberg-tm
- Section 3: Information about the Project Gutenberg Literary Archive Foundation
- Section 4: Information about Donations to the Project Gutenberg Literary Archive Foundation
- Section 5: General Information About Project Gutenberg-tm electronic works.

3. **Analysis of State of the Union Addresses 1 dataset**
Here we will analyze the "state_union_part1.txt" document.

a) **List Top 50 words by frequency (normalized by the length of the document)**
The process followed to find the list of top 50 words by frequency is as follows:

- **Import the NLTK package and FreqDist from NLTK**
Load the "state_union1.txt" text document and read it as a raw string by replacing new lines with space

- **Perform tokenization using NLTK tokenizer**
The above obtained result is in the form of a single string and for the analysis it needs to be tokenized and hence tokenization is performed.

- **Convert tokenized words to lower case**
The tokenized words are then converted to its respective lower-case words since our analysis over text is case sensitive and same words with different case should be considered as one single word.

- **Find the total number of words in the document**
The total number of words in state_union_part1.txt are 244419. This is needed to normalize the document by its length.

- **Filter the words using alpha_filter function to remove non-alphabetical characters**
Many punctuations would have been considered as a word which do not help in the analysis of the document so are removed by using the alpha_filter function.

- **Remove the Stop Words from the list**
The common words consisting in the grammar that have less information to analyze the document are called stop words. There is no single universal list of stop words used by all-natural language processing tools, so we use the list of stop words from NLTK package with some additional updated stop words provided in the course content.

- **Calculate the Frequency Distribution of the updated word list**
  Frequency Distribution tells us the frequency of each vocabulary item in the text using FreqDist. It is a "distribution" because it tells us how the total number of words in the text are distributed across the vocabulary items.

- **Find and display the top 50 words based on the frequency of the words**

b) List Top 50 Bigrams by frequencies

The process followed to find the list of top 50 bigrams by frequency is as follows:

- **Import collocation finder package from nltk to calculate bigram measure**
- **Consider the list of words obtained after the tokenizing and lowercasing for the bigram freuqency analysis**
  If we consider the words after the filtering of non-alphabetical characters or stop words, then the bigrams obtained might not be the pair of bigrams in the original text.

- **Create a finder by the Bigram Collocation finder package**
  Finder will help to apply the word filter for the list of bigrams after the bigrams and its frequency is found.

- **Apply the alpha filter**
  For any finder, we can apply various filter functions, so apply the alpha_filter that we created earlier. It uses a filter that is applied to the individual words.

- **Apply the stop words filter**
  Using the same technique as the previous step we can apply the stopwords filter to remove all the bigrams which consists of stopwords.

- **Remove low frequency words**
  It is often important to remove low frequency words, as we lack sufficient evidence about their significance as collocations.

- **Find and display the top 50 bigrams by the frequency of the words**

c) List Top 50 bigrams by their Mutual Information scores (using min frequency 5)

The process followed to find the list of top 50 bigrams by their mutual individual scores is as follows:

- **Consider the already obtained bigrams after the filtering of non-alphabetical characters and stop words**

- **Apply a minimum frequency of 5 to the finder**
  If you apply the Mutual Information score to all the bigrams, the results are not considered accurate because uniquely occurring pairs of words get high scores. It is recommended to run the PMI scorer with a minimum frequency of 5, which

make more sense on very large documents. The Church and Hanks paper has more discussion of this.

- **Find the Mutual Information Scores**
  Use the score_ngrams function to find the mutual information scores which is provided by the nltk libraries.

- **Find and display the top 50 bigrams by its Mutual information scores**

4. **Analysis of State of the Union Addresses 2 dataset**
   Here we will analyze the "state_union_part2.txt" document.

   a) **List Top 50 words by frequency (normalized by the length of the document)**
   The process followed to find the list of top 50 words by frequency is as follows:

- **Import the NLTK package and FreqDist from NLTK**
  Load the "state_union2.txt" text document and read it as a raw string by replacing new lines with space

- **Perform tokenization using NLTK tokenizer**
  The above obtained result is in the form of a single string and for the analysis it needs to be tokenized and hence tokenization is performed.

- **Convert tokenized words to lower case**
  The tokenized words are then converted to its respective lower-case words since our analysis over text is case sensitive and same words with different case should be considered as one single word.

- **Find the total number of words in the document**
  The total number of words in state_union_part1.txt are 200425. This is needed to normalize the document by its length.

- **Filter the words using alpha_filter function to remove non-alphabetical characters**
  Many punctuations would have been considered as a word which do not help in the analysis of the document so are removed by using the alpha_filter function.

- **Remove the Stop Words from the list**
  The common words consisting in the grammar that have less information to analyze the document are called stop words. There is no single universal list of stop words used by all-natural language processing tools, so we use the list of stop words from NLTK package with some additional updated stop words provided in the course content.

- **Calculate the Frequency Distribution of the updated word list**
  Frequency Distribution tells us the frequency of each vocabulary item in the

text using FreqDist. It is a "distribution" because it tells us how the total number of words in the text are distributed across the vocabulary items.

- **Find and display the top 50 words based on the frequency of the words**

b) List Top 50 Bigrams by frequencies

The process followed to find the list of top 50 bigrams by frequency is as follows:

- **Import collocation finder package from nltk to calculate bigram measure**
- **Consider the list of words obtained after the tokenizing and lowercasing for the bigram freuqency analysis**
  If we consider the words after the filtering of non-alphabetical characters or stop words, then the bigrams obtained might not be the pair of bigrams in the original text.

- **Create a finder by the Bigram Collocation finder package**
  Finder will help to apply the word filter for the list of bigrams after the bigrams and its frequency is found.

- **Apply the alpha filter**
  For any finder, we can apply various filter functions, so apply the alpha_filter that we created earlier. It uses a filter that is applied to the individual words.

- **Apply the stop words filter**
  Using the same technique as the previous step we can apply the stopwords filter to remove all the bigrams which consists of stopwords.

- **Remove low frequency words**
  It is often important to remove low frequency words, as we lack sufficient evidence about their significance as collocations.

- **Find and display the top 50 bigrams by the frequency of the word**

c) List Top 50 bigrams by their Mutual Information scores (using min frequency 5)

The process followed to find the list of top 50 bigrams by their mutual individual scores is as follows:

- **Consider the already obtained bigrams after the filtering of non-alphabetical characters and stop words**

- **Apply a minimum frequency of 5 to the finder**
  If you apply the Mutual Information score to all the bigrams, the results are not considered accurate because uniquely occurring pairs of words get high scores. It is recommended to run the PMI scorer with a minimum frequency of 5, which make more sense on very large documents. The Church and Hanks paper has more discussion of this.

- **Find the Mutual Information Scores**
  Use the score_ngrams function to find the mutual information scores which is provided by the nltk libraries.

- **Find and display the top 50 bigrams by its Mutual information scores**

5. Comparison Analysis

A. Comparison of results from text 1 and text 2 based on the use of the language

i. Comparing the results based on the frequency of words

There were 23 out of 50 words which were common in both the text files (state_union_part1.txt & state_union_part2.txt) which accounts to a similarity of 46% in the usage of the words. The 23 common words in the top 50 words by frequency are as follows:

{‘year’, ‘congress’, ‘government’, ‘people’, ‘war’, ‘us’, ‘nations’, ‘peace’, ‘united’, ‘national’, ‘great’, ‘states’, ‘time’, ‘shall’, ‘state’, ‘one’, also’, every’, ‘power’, ‘union’, ‘made’, ‘public’, ‘last’}

ii. Comparing the results based on bigrams using their frequency

There were 5 out of 50 bigrams which were common in both the text files (state_union_part1.txt & state_union_part2.txt) which accounts to a similarity of 10% in the usage of words. The 5 common bigrams in the top 50 words by frequency are as follows:
{(‘united’, ‘states’), (‘union’, ‘address’), (‘federal’, ‘government’), (‘last’, ‘year’), (‘american’, ‘people’)}

iii. **Comparing the results based on bigrams using their mutual information scores**

There was 1 out of 50 bigrams which was common in both the text files (state_union_part1.txt & state_union_part2.txt) which accounts to a similarity of 2% in the usage of words The 1 common bigram in the top 50 words by mutual information scores is as follows:

{(‘project’, ‘gutenberg’)}

Inference

From the analysis of the most common words based on the frequency and their mutual information score we can infer that both the files have similar language because they are written by the same author and thus have a similar writing style with similar kind of words. Also, the author has used the same words in both the files while writing the speeches. Because of the same reason, the Top-50 frequency words also give similar results for both the documents.

After an analysis over the common words or bigrams, it can be noticed that they are related to the government and people. These are common words that would be spoken in any presidential address. The first text is concentrated on law, citizen, land, treaty, bank, territory, constitution etc., whereas in the second text it is concentrated on economy, freedom, security, budget, federal, defense, health, economy, tax etc. The common words infer that after the second world war the US presidential addresses concentrate more on security, economy, and health aspects.

B. Problems with the word or bigram lists found and possible solutions

i. Auxiliary verbs

Auxiliary verbs help in emphasizing the main verb which is makes it futile to be considerate as we cannot infer more information from those words.

Example: ('shall', 0.0013970313084695023)

Solution: More stop words need to be defined and considered.

ii. Misleading Inference

If there are words which serve as a description of non-alphabetic characters, then they are equal to having the non-alphabetic characters and the filtering does not turn out to be useful.

Example: ('million', 0.0013720843208182613)
('one', 0.0012872645628040414)
('billion', 0.0012722963702132967)

iii. Improper filtering

If the word has numbers along with alphabetic characters, then the bigram was not properly filtered.

Example: (('june', '24th'), 12.651069398933583)
(('october', '20th'), 11.95413846656963)

Solution: This can be solved by using the custom-made tokenizer which will filter the words attached to alphanumeric characters.

C. Comparison of top 50 bigrams by frequency and top 50 bigrams scored by Mutual Information

i. In "state_union_part1.txt" document

After the analysis, the bigram {('andrew', 'jackson'), 6.54613538695438e-5} of top 50 bigrams by frequency and {('andrew', 'jackson'), 13.489605976239465} of top 50 bigrams scored by mutual information remain common in between them. Apart from this we do not find any similarity between the two. This can be attributed to the frequency that was set (i.e. 5) during the generation of bigrams with mutual information scores. The pointwise mutual score is a measure of association. Unlike the bigrams by frequency the PMI

measure is symmetric (pmi(x;y) = pmi(y;x)).

ii. **In "state_union_part2.txt" document**
After the analysis, the bigram {('mr.', 'speaker'), 0.00011974554072595734} of top 50 bigrams by frequency and {('mr.', 'speaker'), 11.56285439922045} of top 50 bigrams scored by mutual information remain common in between them. Apart from this we do not find any similarity between the two. This can be attributed to the frequency that was set (i.e. 5) during the generation of bigrams with mutual information scores. The pointwise mutual score is a measure of association. Unlike the bigrams by frequency the PMI measure is symmetric (pmi(x;y) = pmi(y;x)).

6. Appendix One: Outputs

❖ **Output of Analysis 1:**

Analysis Results for State_Union_Part1.txt Document		
Top 50 words by their normalized frequency	**Top 50 Bigrams**	**Top 50 Bigrams by their Mutual Information Score**
('states', 0.004390002413887628)	(('united', 'states'), 0.0032403372896542414)	(('project', 'gutenberg'), 14.729071910934852)
('government', 0.003346711998658042)	(('last', 'session'), 0.0006382482540228051)	(('gun', 'boats'), 14.577068817489806)
('united', 0.003301707314079511)	(('union', 'address'), 0.0004909601954021578)	(('sublime', 'porte'), 14.21717287240342)
('may', 0.0031339625806504406)	(('great', 'britain'), 0.0004705035205937345)	(('precious', 'metals'), 14.028632192793763)
('congress', 0.002512079666474374)	(('fellow', 'citizens'), 0.0003845854863983569)	(('circulating', 'medium'), 13.976164772899626)
('public', 0.002487531656704266)	(('public', 'debt'), 0.0003804941514366723)	(('buenos', 'ayres'), 13.577068817489804)
('upon', 0.002254325563888241)	(('present', 'year'), 0.00031094145708803324)	(('quincy', 'adams'), 13.506679489598405)
('made', 0.0020743068255741166)	(('general', 'government'), 0.00019229274319917846)	(('andrew', 'jackson'), 13.489605976239465)

('great', 0.002062032820689063)	(('two', 'countries'), 0.00018820140823749381)	(('grateful', 'acknowledgments'), 13.388034993099787)
('state', 0.0019802061214553696)	(('british', 'government'), 0.00018001873831412452)	(('thomas', 'jefferson'), 13.314034411656012)
('country', 0.001812461388026299)	(('french', 'government'), 0.00018001873831412452)	(('john', 'quincy'), 13.255140722602441)
('last', 0.0016528993245205979)	(('public', 'lands'), 0.00016365339846738592)	(('chief', 'magistrate'), 13.246920215797473)
('present', 0.0015424332805551123)	(('ensuing', 'year'), 0.00015137939358233197)	(('cumberland', 'road'), 13.121389333713614)
('war', 0.001395145221934465)	(('commercial', 'intercourse'), 0.00014728805862064732)	(('lake', 'erie'), 13.044573736662784)
('year', 0.0013828712170494111)	(('indian', 'tribes'), 0.00013092271877390875)	(('supreme', 'ruler'), 13.021252662428164)
('citizens', 0.0013583232072793031)	(('january', '1st'), 0.00012274004885053945)	(('navigable', 'rivers'), 12.729071910934852)
('time', 0.00133786653247088)	(('public', 'service'), 0.00012274004885053945)	(('charge', "d'affaires"), 12.689543546748217)
('every', 0.0013215011926241413)	(('taken', 'place'), 0.00011864871388885479)	(('circuit', 'courts'), 12.651069398933583)
('union', 0.0012846791779689796)	(('federal', 'government'), 0.00011455737892717014)	(('june', '24th'), 12.651069398933583)
('subject', 0.0012846791779689796)	(('mean', 'time'), 0.00011455737892717014)	(('north', 'carolina'), 12.483959413098322)
('part', 0.0012069438136969712)	(('several', 'states'), 0.00010637470900380085)	(('sinking', 'fund'), 12.476091169764985)
('general', 0.0011783044689651786)	(('two', 'nations'), 0.00010637470900380085)	(('st.', 'augustine'), 12.47273215767507)
('commerce', 0.001174213134003494)	(('september', 'last'), 0.0001022833740421162)	(('st.', 'croix'), 12.47273215767507)
('necessary', 0.0011660304640801248)	(('two', 'governments'), 0.0001022833740421162)	(('st.', 'petersburg'), 12.47273215767507)

('power', 0.0011537564591950708)	(('naval', 'force'), 9.819203908043156e-05)	(('st.', 'marys'), 12.472732157675068)
('shall', 0.0011414824543100168)	(('present', 'session'), 9.819203908043156e-05)	(('central', 'america'), 12.45937377482005)
('us', 0.0011210257795015936)	(('public', 'interest'), 9.819203908043156e-05)	(('george', 'washington'), 12.35210245248953)
('people', 0.001116934444539909)	(('every', 'part'), 9.410070411874691e-05)	(('beg', 'leave'), 12.34440806069953)
('treaty', 0.0011087517746165396)	(('foreign', 'nations'), 9.410070411874691e-05)	(('james', 'madison'), 12.269640292297558)
('one', 0.001104660439654855)	(('american', 'people'), 9.000936915706226e-05)	(('james', 'monroe'), 12.269640292297558)
('without', 0.0010842037648464318)	(('last', 'year'), 9.000936915706226e-05)	(('britannic', 'majesty'), 12.255140722602443)
('act', 0.0010760210949230624)	(('friendly', 'relations'), 8.591803419537761e-05)	(('catholic', 'majesty'), 12.255140722602443)
('nations', 0.001022833740421162)	(('good', 'faith'), 8.591803419537761e-05)	(('deeply', 'impressed'), 12.176530887906075)
('session', 0.0010023770656127388)	(('buenos', 'ayres'), 8.182669923369296e-05)	(('manufacturing', 'establishments'), 12.028632192793761)
('peace', 0.000941007041187469)	(('catholic', 'majesty'), 8.182669923369296e-05)	(('sensibly', 'felt'), 12.021252662428166)
('treasury', 0.0009205503663790458)	(('minister', 'plenipotentiary'), 8.182669923369296e-05)	(('west', 'india'), 11.992106316768647)
('within', 0.0009123676964556765)	(('two', 'years'), 8.182669923369296e-05)	(('october', '20th'), 11.95413846656963)
('attention', 0.000883728351723884)	(('charge', "d'affaires"), 7.773536427200831e-05)	(('port', 'towns'), 11.953553075999256)
('duties', 0.0008796370167621993)	(('late', 'war'), 7.773536427200831e-05)	(('mutually', 'beneficial'), 11.90464347551831)
('national', 0.0008755456818005147)	(('sea', 'men'), 7.773536427200831e-05)	(('john', 'adams'), 11.862823299823683)

('interest', 0.0008591803419537761)	(('state', 'governments'), 7.773536427200831e-05)	(('divine', 'providence'), 11.854602793018712)
('powers', 0.0008550890069920914)	(('discriminating', 'duties'), 7.364402931032366e-05)	(('west', 'indies'), 11.822181315326336)
('new', 0.0008509976720304068)	(('beloved', 'country'), 6.955269434863901e-05)	(('black', 'sea'), 11.689543546748217)
('effect', 0.0008469063370687221)	(('foreign', 'powers'), 6.955269434863901e-05)	(('ship', 'building'), 11.661957715076316)
('important', 0.0008387236671453528)	(('foreign', 'relations'), 6.955269434863901e-05)	(('post', 'master'), 11.620547454156686)
('interests', 0.0008387236671453528)	(('internal', 'improvement'), 6.955269434863901e-05)	(('post', 'masters'), 11.620547454156686)
('also', 0.0008264496622602989)	(('new', 'york'), 6.955269434863901e-05)	(('human', 'race'), 11.61359469351492)
('laws', 0.0008223583272986142)	(('public', 'revenue'), 6.955269434863901e-05)	(('six', 'months'), 11.44943553776393)
('system', 0.0008182669923369296)	(('address', 'james'), 6.546135938695438e-05)	(('supreme', 'court'), 11.428676977597133)
('force', 0.0008141756573752449)	(('andrew', 'jackson'), 6.546135938695438e-05)	(('royal', 'order'), 11.407143816047492)

- **Output of Analysis 2:**

Analysis Results of State_Union_Part2.txt Document		
Top 50 words by their normalized frequency	**Top 50 Bigrams**	**Top 50 Bigrams by their Mutual Individual Scores**
('world', 0.003497567668704004)	(('united', 'states'), 0.0010278158912311339)	(('j.', 'clinton'), 14.805348026613416)
('year', 0.002893850567543969)	(('fiscal', 'year'), 0.0008232505924909567)	(('ronald', 'reagan'), 14.805348026613416)
('congress', 0.0028339777971809906)	(('million', 'dollars'), 0.0006087064986902832)	(('william', 'j.'), 14.805348026613416)
('government', 0.0026793064737432957)	(('billion', 'dollars'), 0.0005787701135087938)	(('barack', 'obama'), 14.612702948671021)

('people', 0.002669327678682799)	(('united', 'nations'), 0.0005538231258575527)	(('w.', 'bush'), 14.290774853783656)
('new', 0.002664338281152551)	(('free', 'world'), 0.0004989397530248223)	(('floor', 'appears'), 14.197665449392176)
('war', 0.0022502182861419483)	(('union', 'address'), 0.0004989397530248223)	(('lyndon', 'b.'), 14.027740447949864)
('us', 0.00210552575776475)	(('federal', 'government'), 0.00047898216290382934)	(('iron', 'curtain'), 13.912263230529929)
('nations', 0.00210552575776475)	(('last', 'year'), 0.000404141199950106)	(('george', 'w.'), 13.912263230529927)
('nation', 0.0020256953972807783)	(('free', 'nations'), 0.00036921541723836847)	(('john', 'f.'), 13.912263230529927)
('program', 0.0019807908195085444)	(('let', 'us'), 0.0003542472246476238)	(('richard', 'nixon'), 13.857815446507551)
('years', 0.0018959710614943246)	(('soviet', 'union'), 0.00029936385181489333)	(('f.', 'kennedy'), 13.805348026613414)
('free', 0.0018710240738430836)	(('world', 'war'), 0.0002644380691031558)	(('mass', 'transit'), 13.764706042116067)
('federal', 0.0018460770861918423)	(('social', 'security'), 0.0002594486715729076)	(('b.', 'johnson'), 13.705812353062502)
('economic', 0.0018011725084196083)	(('american', 'people'), 0.00024946987651241115)	(('project', 'gutenberg'), 13.61270294867102)
('peace', 0.0017612573281776225)	(('armed', 'forces'), 0.00024946987651241115)	(('harry', 's.'), 13.44277794722871)
('united', 0.0017313209429961332)	(('years', 'ago'), 0.0002444804789821629)	(('dwight', 'd.'), 13.290774853783658)
('national', 0.0016115754022701758)	(('economic', 'growth'), 0.00020456529874017713)	(('river', 'basins'), 13.290774853783656)
('great', 0.0015966072096794313)	(('civil', 'rights'), 0.00019458650367968068)	(('s.', 'truman'), 13.22038552589226)
('states', 0.0015566920294374455)	(('foreign', 'policy'), 0.000179618311088936)	(('d.', 'eisenhower'), 12.912263230529929)

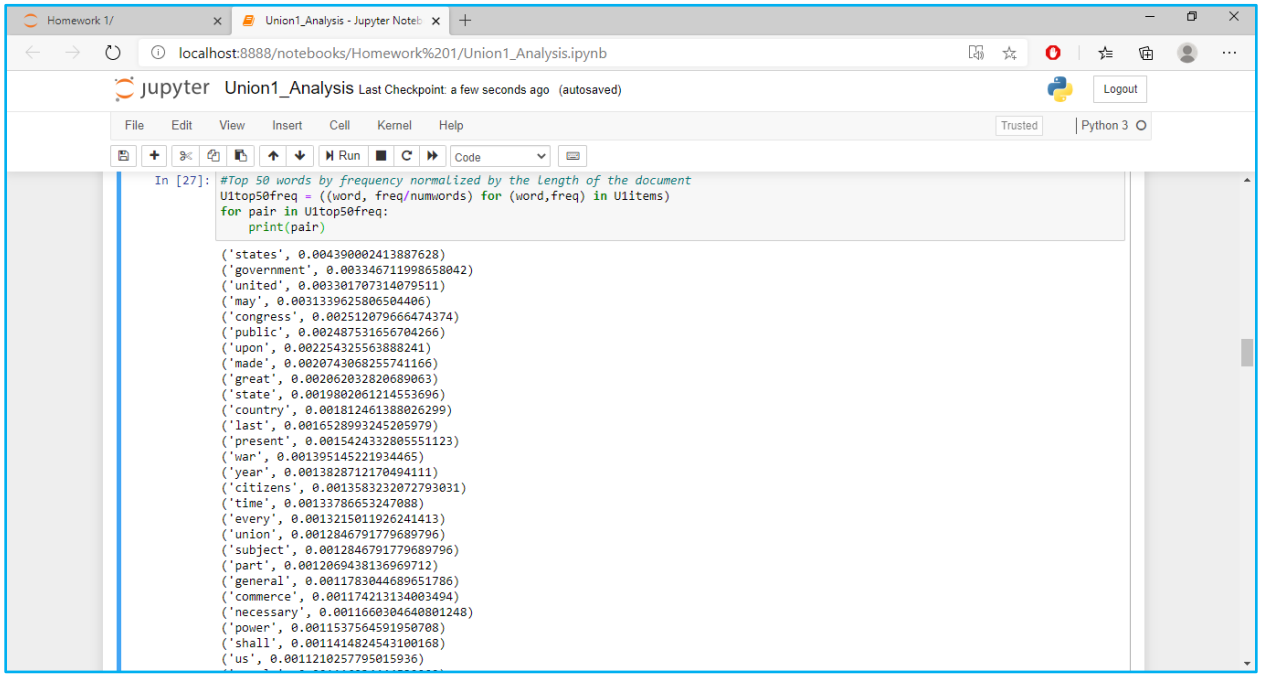
('time', 0.0014519146813022328)	(('medical', 'care'), 0.000179618311088936)	(('displaced', 'persons'), 12.602718860098399)
('shall', 0.0013970313084695023)	(('atomic', 'energy'), 0.0001746289135586878)	(('rural', 'electrification'), 12.568308829312567)
('million', 0.0013720843208182613)	(('full', 'employment'), 0.00015966072096794311)	(('test', 'ban'), 12.49168754770965)
('security', 0.001367094923288013)	(('past', 'year'), 0.00015966072096794311)	(('ballistic', 'missiles'), 12.483419931726056)
('state', 0.0013571161282275165)	(('every', 'american'), 0.0001546713234376949)	(('raw', 'materials'), 12.442777947228706)
('dollars', 0.0013072221529250342)	(('war', 'ii'), 0.0001546713234376949)	(('export-import', 'bank'), 12.40324958304207)
('american', 0.0012922539603342897)	(('shall', 'continue'), 0.00014968192590744667)	(('unliquidated', 'obligations'), 12.395472232450349)
('one', 0.0012872645628040414)	(('local', 'governments'), 0.00014469252837719844)	(('inflationary', 'pressures'), 12.29077485378366)
('billion', 0.0012722963702132967)	(('public', 'works'), 0.00014469252837719844)	(('nursing', 'homes'), 12.220385525892262)
('make', 0.0012074342023200698)	(('natural', 'resources'), 0.00013970313084695022)	(('outer', 'space'), 12.220385525892258)
('many', 0.0012024448047898216)	(('executive', 'branch'), 0.0001297243357864538)	(('southeast', 'asia'), 12.058114096993384)
('defense', 0.0011824872146688288)	(('national', 'defense'), 0.0001297243357864538)	(('mobilization', 'base'), 11.968846758896296)
('also', 0.0011774978171385805)	(('eight', 'years'), 0.00012473493825620557)	(('ballistic', 'missile'), 11.930878908697276)
('every', 0.0011774978171385805)	(('military', 'strength'), 0.00012473493825620557)	(('intergovernmental', 'relations'), 11.831343235146356)
('freedom', 0.0011525508294873395)	(('small', 'business'), 0.00012473493825620557)	(('white', 'house'), 11.788274513254473)
('america', 0.0011475614319570913)	(('mr.', 'speaker'), 0.00011974554072595734)	(('middle', 'east'), 11.784286411085587)

('power', 0.0011325932393663465)	(('western', 'europe'), 0.00011974554072595734)	(('marshall', 'plan'), 11.71961815258753)
('military', 0.0011325932393663465)	(('national', 'security'), 0.00011475614319570911)	(('contract', 'authorizations'), 11.7058123530625)
('help', 0.0011176250467756017)	(('shall', 'recommend'), 0.0001097667456654609)	(('head', 'start'), 11.686703530114794)
('union', 0.0011076462517151055)	(('minimum', 'wage'), 0.00010477734813521268)	(('open', 'spaces'), 11.6354230251711)
('fiscal', 0.001097667456654609)	(('private', 'enterprise'), 0.00010477734813521268)	(('mr.', 'speaker'), 11.562854399220457)
('made', 0.0010926780591243607)	(('cold', 'war'), 9.978795060496445e-05)	(('maritime', 'commission'), 11.546613758213248)
('economy', 0.0010627416739428714)	(('collective', 'bargaining'), 9.978795060496445e-05)	(('collective', 'bargaining'), 11.53588735162019)
('public', 0.0010627416739428714)	(('collective', 'security'), 9.978795060496445e-05)	(('urban', 'renewal'), 11.49168754770965)
('first', 0.0010577522764126232)	(('every', 'citizen'), 9.978795060496445e-05)	(('legislative', 'branches'), 11.460699855225972)
('programs', 0.0010377946862916304)	(('first', 'time'), 9.978795060496445e-05)	(('monetary', 'fund'), 11.437064296034313)
('last', 0.0010328052887613821)	(('free', 'men'), 9.978795060496445e-05)	(('universal', 'training'), 11.364775435227433)
('progress', 0.0010228264937008856)	(('next', 'fiscal'), 9.978795060496445e-05)	(('surprise', 'attack'), 11.30527442347877)
('strength', 0.0010128476986403891)	(('purchasing', 'power'), 9.978795060496445e-05)	(('north', 'atlantic'), 11.290774853783656)
('system', 0.0010028689035798926)	(('world', 'peace'), 9.978795060496445e-05)	(('wage', 'earners'), 11.25835337609128)

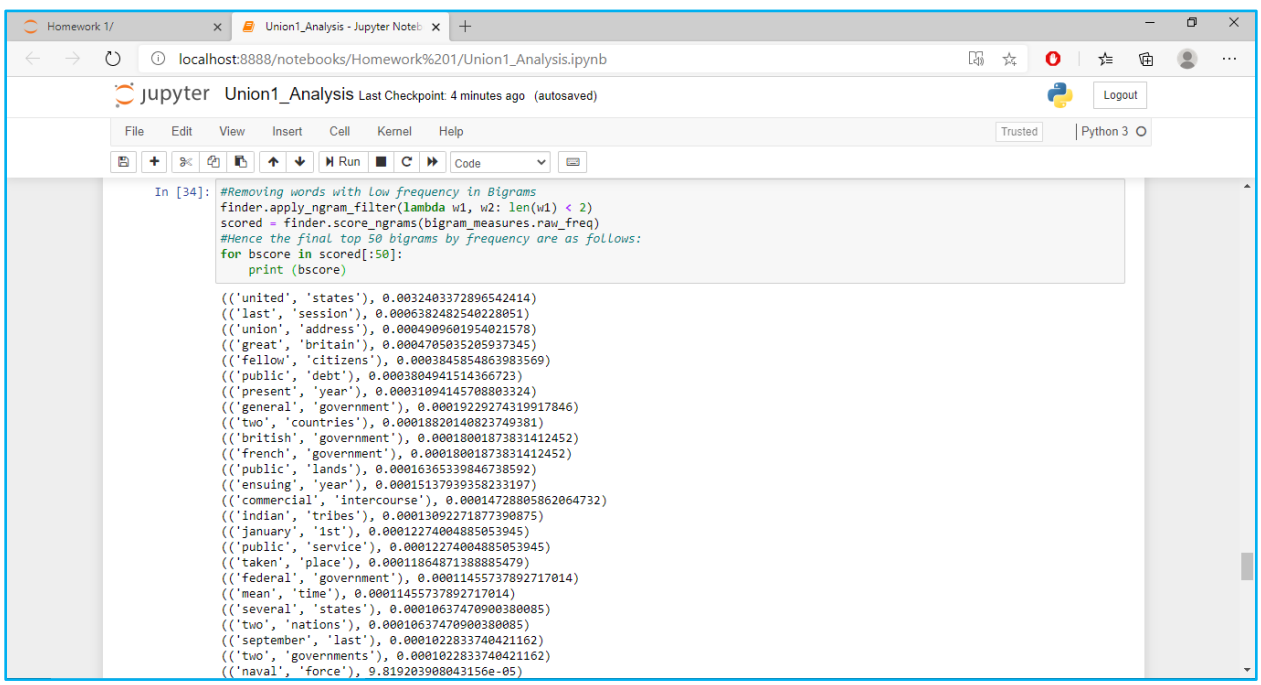
7. Appendix Two: IDE Screenshots:

a) Analysis of “state_union-part1.txt” document:

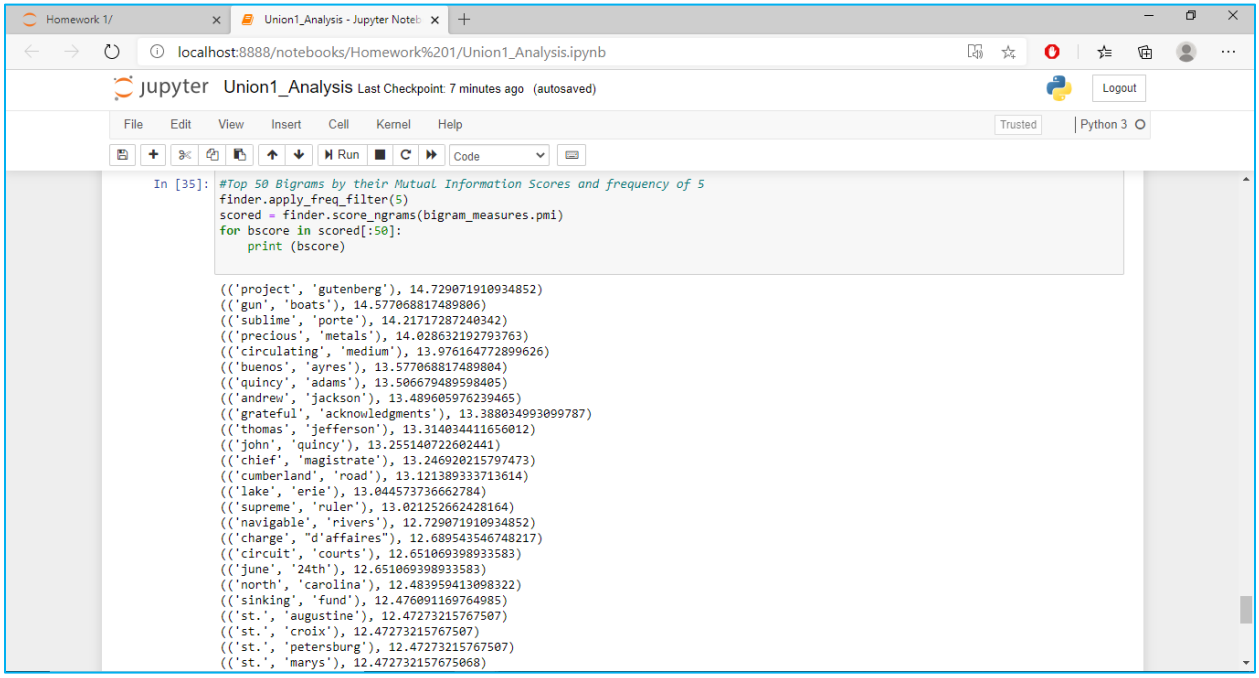
- **Top 50 words by frequency (normalized by the length of the document)**



- **Top 50 bigrams by frequency**

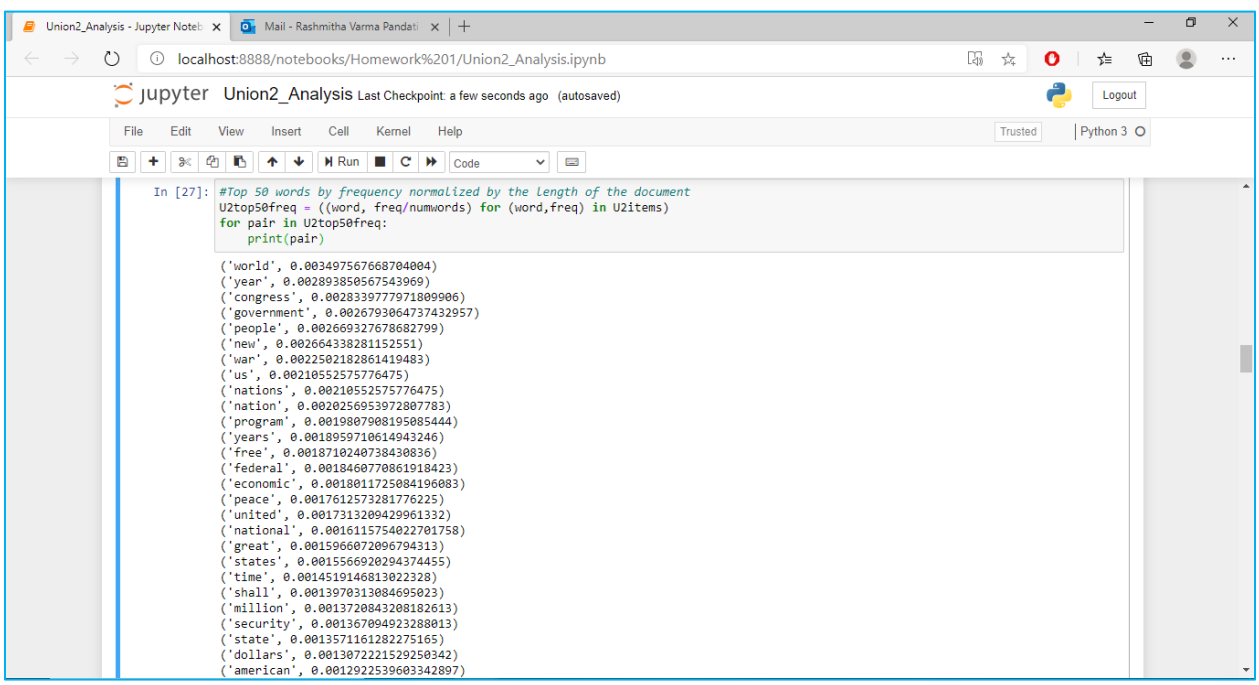


- **Top 50 bigrams by Mutual Information Scores**

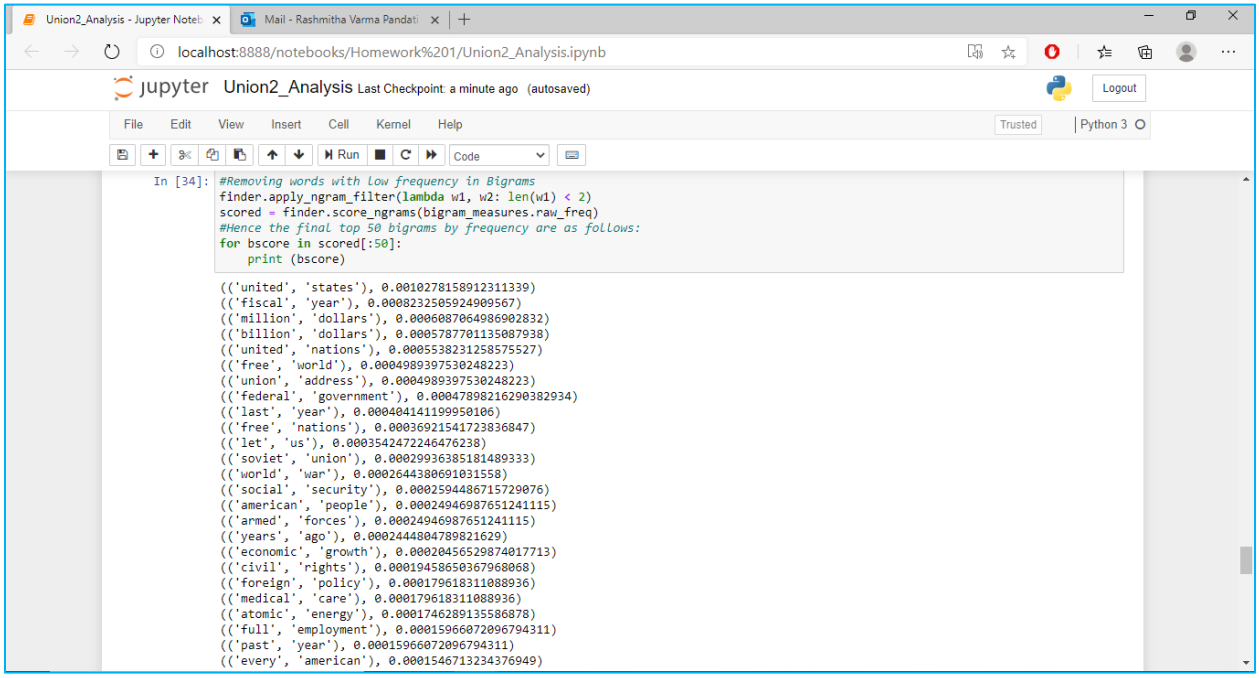


b) Analysis of "state_union_part2.txt" document:

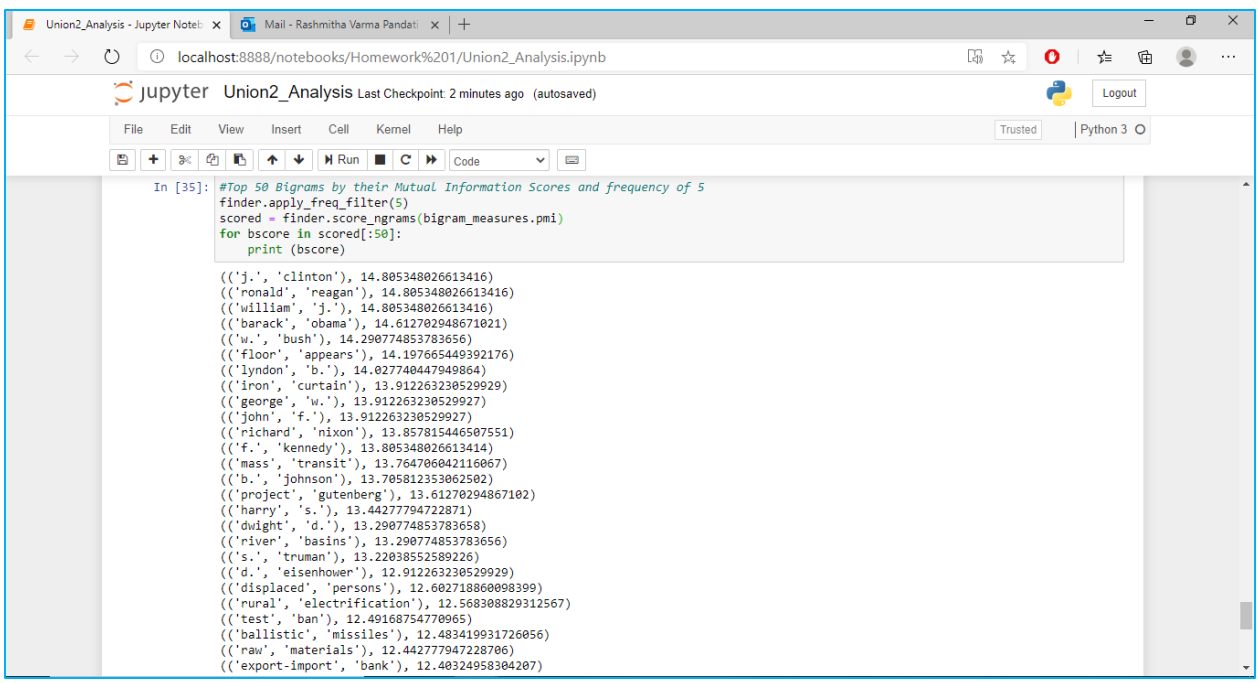
- **Top 50 words by frequency (normalized by the length of the document)**



- **Top 50 bigrams by frequency**



- **Top 50 bigrams by Mutual Information Scores**



8. References:

- Course content
- https://en.wikipedia.org/wiki/Project_Gutenberg