CS 657 Mining Massive Datasets

Assignment 1: Modified WordCount/Pairs counting

Name: ANIKETH SURESH

G#: [redacted]

Introduction

The goal of this assignment was to perform several tasks regarding textual data. The dataset consisted of text transcriptions of the State of the Union Addresses, given by Presidents to Congress from 1790 (George Washington) to 2020 (Donald Trump).

Data Acquisition

To acquire the data, these steps were taken:

1) Wrote a python script to fetch the page source(HTML) of the main index page (http://stateoftheunion.onetwothree.net/texts/index.html).
2) Parsed the HTML to obtain the href (Hypertext REFerence: links) for all the addresses.
3) Parsed the files and renamed each webpage in the format yyyymmdd.txt corresponding to the date of the address. (Note that these are all text files).

Tasks

Part 1

Data Cleaning:

To clean the data, these steps in order were required:

1) The speech data was split into lines.
2) If the line contained characters or special words (from a list), then the line was removed. This accounted for removal of all the html elements.
3) Punctuation was removed.
4) Stop words (checked from a list of words) were removed. (Eg,.: yourself, and, you're, on, of, the a, both, does, have, not, etc)
5) Words of length <=2 were removed to clean up some html elements that snuck through.
6) All the words were then lower-cased.

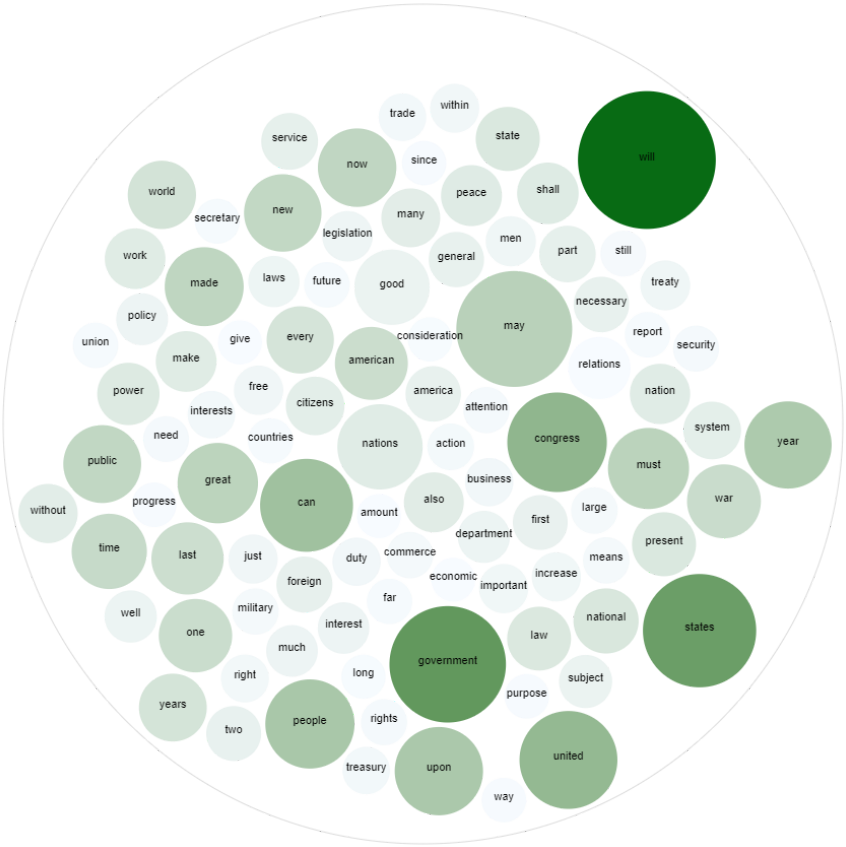
Some words are shown below. Each word is preceded by the year of the speech that it was taken from.

1790	**government**	**1790**	**plenty**	**1942**	**urgent**
1790	**union**	**1790**	**blessed**	**2019**	**giant**
1790	**concord**	**1790**	**circumstances**	**2019**	**leaps**
1790	**peace**	**1942**	**modern**	**2019**	**science**
1790	**plenty**	**1942**	**methods**	**2019**	**discovery**
1790	**blessed**	**1942**	**warfare**	**2019**	**unrivaled**
1790	**circumstances**	**1942**	**make**	**2019**	**progress**
1790	**government**	**1942**	**task**	**2019**	**toward**
1790	**union**	**1942**	**shooting**	**2019**	**equality**
1790	**concord**	**1942**	**fighting**	**2020**	**fellow**
1790	**peace**	**1942**	**even**	**2020**	**americans**

Average:

To compute the average, the counts for all words were calculated and then divided by the total number of speeches.

The top 50 word-averages are shown below. Each circle represents the value of the average use of the word over all the years.



Maximum and Minimum:

To compute the max and min, the counts for all words were first calculated and were then used to find the max and min number of times it appeared over all the years. Therefore, we can know the count that a particular word was used to its maximum and minimum.
The data for these counts are in the Part1-max-count.txt and Part1-min-count.txt.
Also, the max and mix counts were merged and sorted based on the max count to show that it makes sense for the minimum of a word (that is repetitive) would have a min count of ~1. This data can be found in Part1-min-max-together.txt.

Top 10 MAX:					
will	**290**	**year**	**182**	**fiscal**	**129**
dollars	**206**	**mexico**	**158**	**administration**	**128**
war	**195**	**million**	**136**	**congress**	**127**
		states	**135**		

Top 10 MIN:					
bilateral	**9**	**herrera**	**8**	**vj-day**	**7**
deforestation	**9**	**yamen**	**8**	**greytown**	**7**
anna	**8**	**wine**	**7**	**practise**	**7**
		gerrymander	**7**		

Average and Standard Deviation in a window of 4 years:

To accomplish this, all the years in a bracket/window of 4 consecutive years were assigned an id corresponding to its 4-year bracket/window. It was now easy to calculate the average for words in these windows. The standard deviation was also calculated in this manner, after applying the equation for standard deviation.
Some counts for the average are given below on the column in green. The id corresponds to the bracket/window, with 0-indexing.
Some counts for the standard deviation are also given below on the column in blue.

(6, 'will')	**64.5**	**(3, 'can')**	**30.25**
(2, 'people')	**55.25**	**(2, 'must')**	**29.75**
(3, 'must')	**51.25**	**(3, 'every')**	**29.25**
(3, 'will')	**50.25**	**(3, 'years')**	**29.0**
(4, 'will')	**49.4**	**(6, 'can')**	**28.25**
(3, 'new')	**39.25**	**(7, 'new')**	**27.25**
(3, 'america')	**35.75**	**(5, 'america')**	**26.75**
(2, 'work')	**33.25**	**(7, 'america')**	**26.75**
(2, 'can')	**31.0**	**(1, 'must')**	**26.5**
(8, 'american')	**31.0**	**(2, 'challenge')**	**16.60383991732033**
(3, 'now')	**30.5**		

(1, 'must')	14.430869689661812
(2, 'health')	12.90348790056394
(2, 'care')	12.833062767710599
(2, 'people')	12.275483697190918
(2, 'government')	10.207227831296802
(3, 'now')	10.062305898749054
(7, 'help')	10.059199769365355
(2, 'plan')	10.034316120194738
(4, 'must')	9.869143833180262
(2, 'can')	9.695359714832659
(3, 'security')	9.5524865872714
(2, 'new')	9.33742469849155
(4, 'budget')	9.17867092775419
(3, 'new')	9.065732182234372
(4, 'weapons')	8.863859204657981
(3, 'social')	8.774964387392123
(3, 'century')	8.699856320652657
(6, 'right')	8.514693182963201
(5, 'america')	8.407585860400118
(5, 'security')	8.381527307120106
(1, 'plan')	8.37779804005802

Average plus 2 Standard Deviations:

Using the previous result I was able to find the words in the years following each bracket/window that appeared with a frequency that was greater than the average plus two standard deviations. If we were to assume that the data came from a gaussian/normal distribution (maybe unlikely), then in this case the words which satisfied this criteria are in the >95% percentile with a low chance of occurring. These words would have to be at a frequency that was not used as much in the previous year's window/bracket. This would possibly imply a change in tone/words/ideologies used for different issues. Most likely this would relate to the change in the president (term of 4 years). This would be the reason for policy change and a different type of leadership. A few interesting words that I found in the results are shown below. The index relates to the 0-indexed year's bracket/window after which the word shown, satisfies the criteria in this section.

(2, 'challenge')
(2, 'health')
(3, 'security')
(4, 'budget')
(4, 'weapons')
(7, 'jobs')
(2, 'investment')
(4, 'saddam')
(4, 'hussein')
(3, 'gun')
(2, 'crime')
(5, 'iraq')
(5, 'qaeda')
(3, 'education')
(3, 'revolution')
(8, 'healthcare')
(5, 'extremists')
(5, 'oil')
(6, 'research')

Part 2

Co-occurrence of words:

Interestingly, I tried a different approach to cleaning the data in this section.
Instead of separating the sentences based on a new line, I extracted all the text between the paragraph tags (html) and cleaned the data with methods similar to the previous section. I believe this should have produced a slightly better result with respect to the data cleaning. Here are a few highly frequent co-occurring words:

('states', 'united') 5686
('mexico', 'texas') 138
('government', 'indians') 89
('fiscal', 'year') 978
('government', 'people') 902
('must', 'war') 136
('last', 'year') 884
('dollar', 'gold') 37
('congress', 'states') 870
('justice', 'true') 14
('great', 'states') 621
('department', 'post-office') 99
('americans', 'million') 80
('britain', 'great') 585
('federal', 'government') 572
('expenditures', 'treasury') 90
('congress', 'session') 542
('tribes', 'war') 24
('indian', 'interior') 32
('expenditures', 'interest') 41

The pairs represent the co-occuring words and the number is the count with which they reoccur.

Conditional Probability P(B/A):

Both P(B/A) and P(A/B) have been calculated. Here are a few cases whereP(A/B) exceed 0.8:

('mills', 'silk')
('guns', 'rapid-fire')
('indians', 'self-support')
('battleships', 'submarines')
('courts', 'jurors')
('coffee', 'tea')
('missiles', 'warheads')
('harbor', 'pearl')
('soldiers', 'widows')
('measures', 'weights')
('man', 'wageworker')
('death', 'sin')
('tribes', 'uncivilized')
('al', 'qaeda')
('banks', 'discount')
('navy', 'orphans')

Lift:

The co-occurence of words with a lift larger than 1.5 would mean that the pair has a large association within the speeches. In certain cases where the confidence measure is not enough, the lift of the two words would account for the actual ‘list’ that the first term provides to be associated with the confidence that the second term would be related. A low lift would naturally result in two non-related words.
The results for the remaining lift solution is in Part2-lift.txt

('cloth', 'yarn')
('1928', 'department-store')
('linen', 'mills')
('knitting', 'mills')
('mills', 'plush')
('sub', 'submarines')
('complainant', 'irreparable')
('brakes', 'couplers')
('mills', 'silk')
('beacons', 'piers')
('pregnancy', 'teen')
('battleships', 'destroyers')
('ballistic', 'intercontinental')
('daughters', 'sons')

('sheep', 'wool')
('math', 'reading')
('democrats', 'republicans')
('colleges', 'universities')
('beds', 'hospital')
('coin', 'convertibility')
('bonds', 'percents')
('fuels', 'solar')
('killed', 'wounded')
('doctors', 'nurses')
('clothing', 'shelter')
('certificates', 'forged')
('beds', 'veterans')
('coffee', 'sugar')
('latitude', 'parallel')
('energy', 'exploratory')