Analyzing Presidential Inaugural Speeches

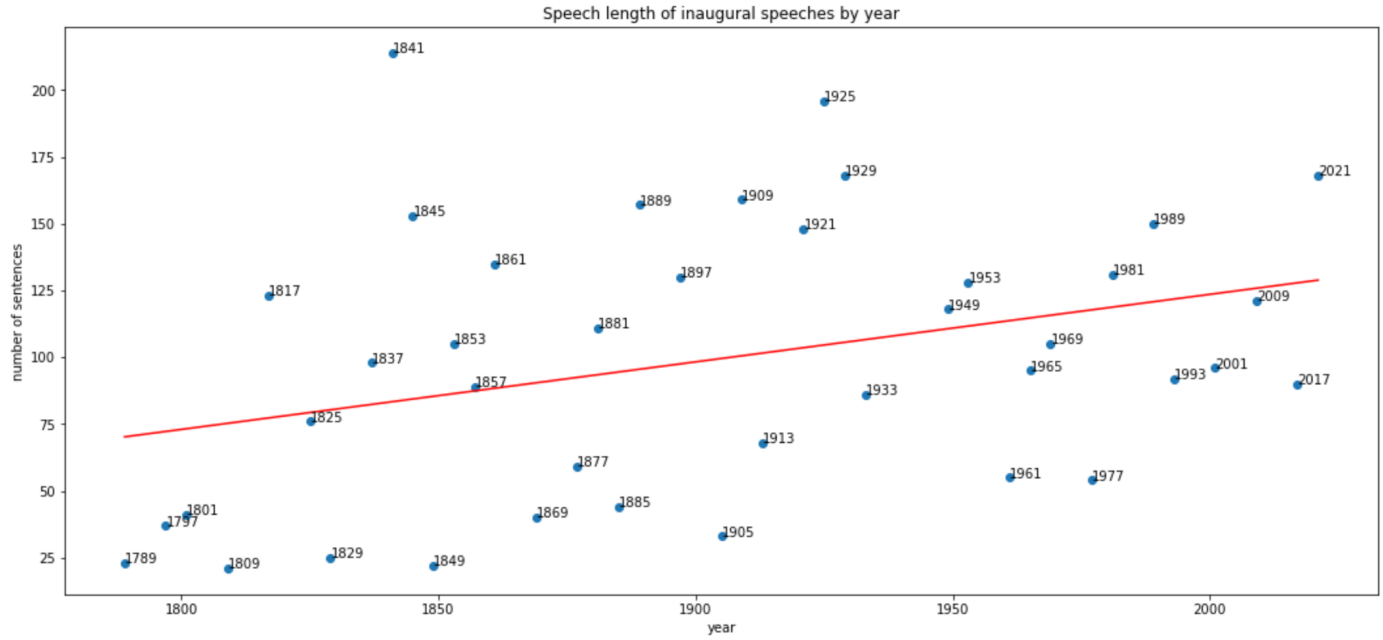
a) **Text length**

Although there isn't a strong pattern, the best fit lines show a slight upward trend in the number of sentences and a very slight downward trend in the number of words over time. This would presumably mean that the average sentence length has been decreasing over time (see part (c)). The longest speech as determined by both the number of sentences and the number of words is William Henry Harrison's 1841 inaugural speech. His speech contained 8469 words, which is a clear outlier in the plot (ii). However, the number of sentences in his speech is not an outlier in the plot (i). The shortest speech in terms of sentences was James Madison's 1809 speech with 21 sentences. The shortest speech in terms of words was Theodore Roosevelt's 1905 speech, the only speech under 1000 words. However, neither one of these speeches seem to be outliers as there are other data points on the graphs with similar y-values.

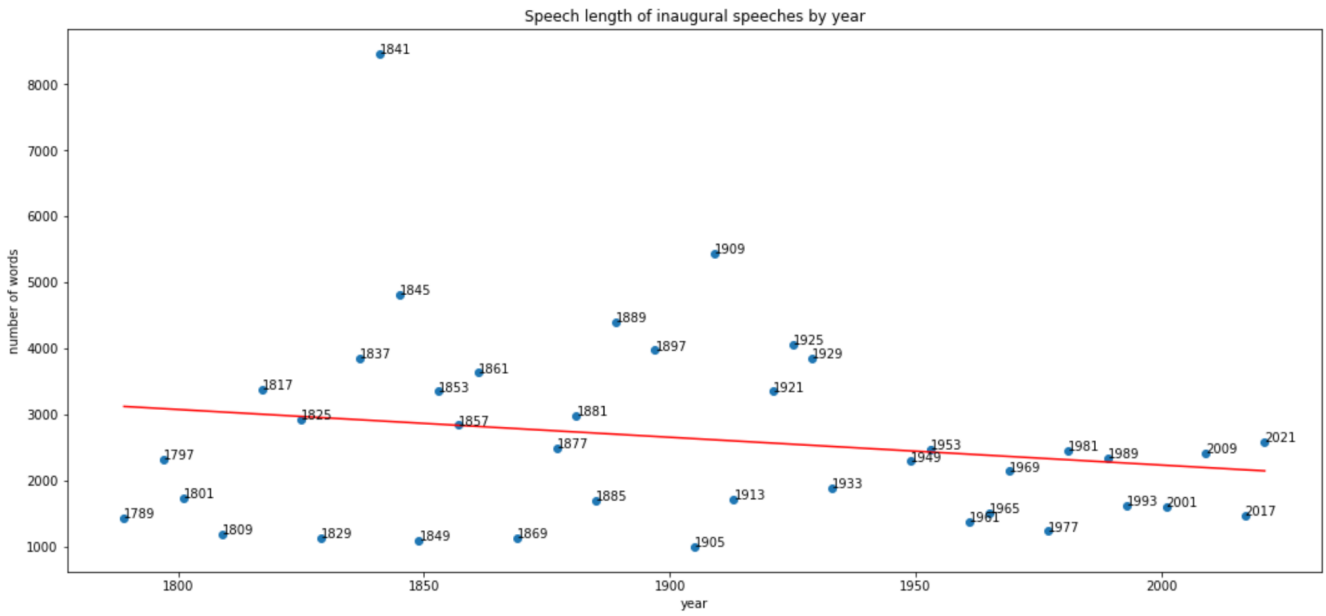
As for whether the length of these speeches correlate with how influential or popular the Presidents were, I think that it is hard to tell. For one, approval rating statistics vary greatly from source to source and only concern the last 10-15 presidents[1]. As for correlations with influence, presidents like Lincoln, Trump, and George Washington all have varying sentence and word counts that do not form any pattern.

[1] https://www.presidency.ucsb.edu/statistics/data/final-presidential-job-approval-ratings

(i)


Speech length of inaugural speeches by year
number of sentences
year
1800
1850
1900
1950
2000
25
50
75
100
125
150
175
200
1789
1797
1801
1809
1817
1825
1829
1837
1841
1845
1849
1853
1857
1861
1869
1877
1881
1885
1889
1897
1905
1909
1913
1921
1925
1929
1933
1949
1953
1961
1965
1969
1977
1981
1989
1993
2001
2009
2017
2021

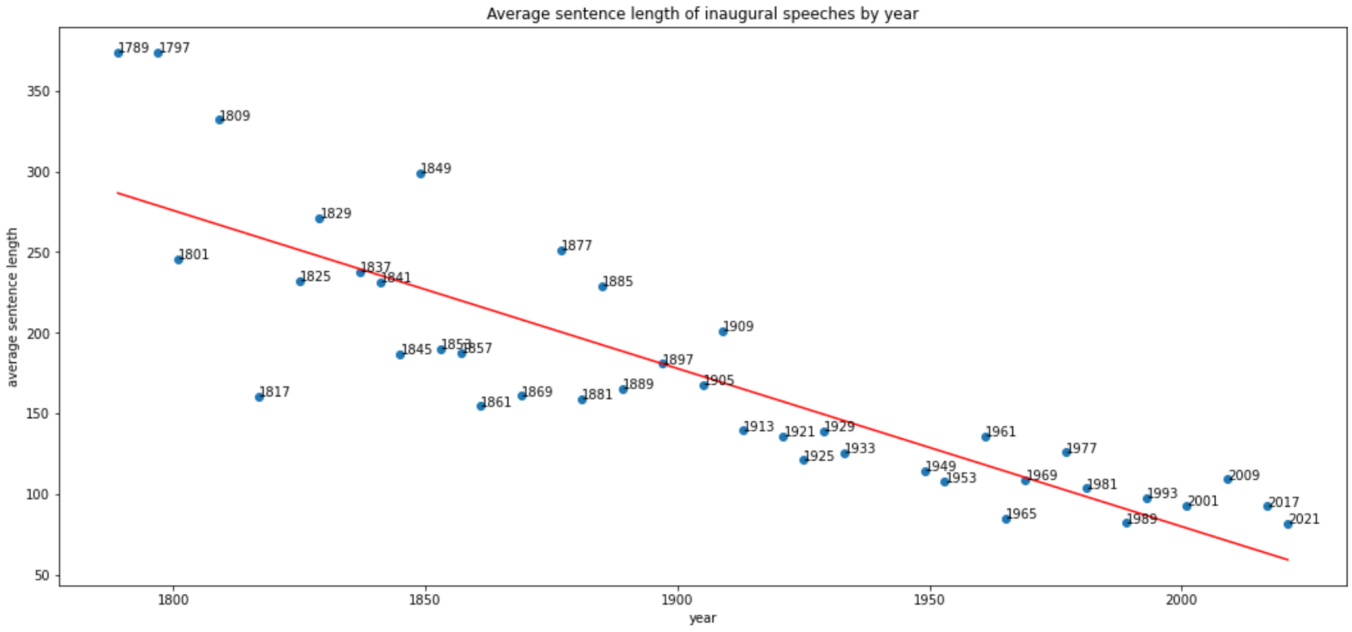
(ii)


Speech length of inaugural speeches by year
number of words
year
1000
2000
3000
4000
5000
6000
7000
8000
1800
1850
1900
1950
2000
1789
1797
1801
1809
1817
1825
1829
1837
1841
1845
1849
1853
1857
1861
1869
1877
1881
1885
1889
1897
1905
1909
1913
1921
1925
1929
1933
1949
1953
1961
1965
1969
1977
1981
1989
1993
2001
2009
2017
2021

b) **Sentence length**: What is the average sentence length (in words) of all speeches?

There is a clear trend of the average sentence length decreasing in inaugural speeches over time. This is illustrated by the best fit line in red. By using the interquartile range, we can see that there are no clear outliers. It is apparent that earlier presidents used much longer sentences– George Washington and John Adams having the highest average sentence length. Whereas George Bush had the lowest sentence length. It is interesting that at the turn of the twentieth century and throughout the world wars and recessions, average sentence length stayed low.

(i)

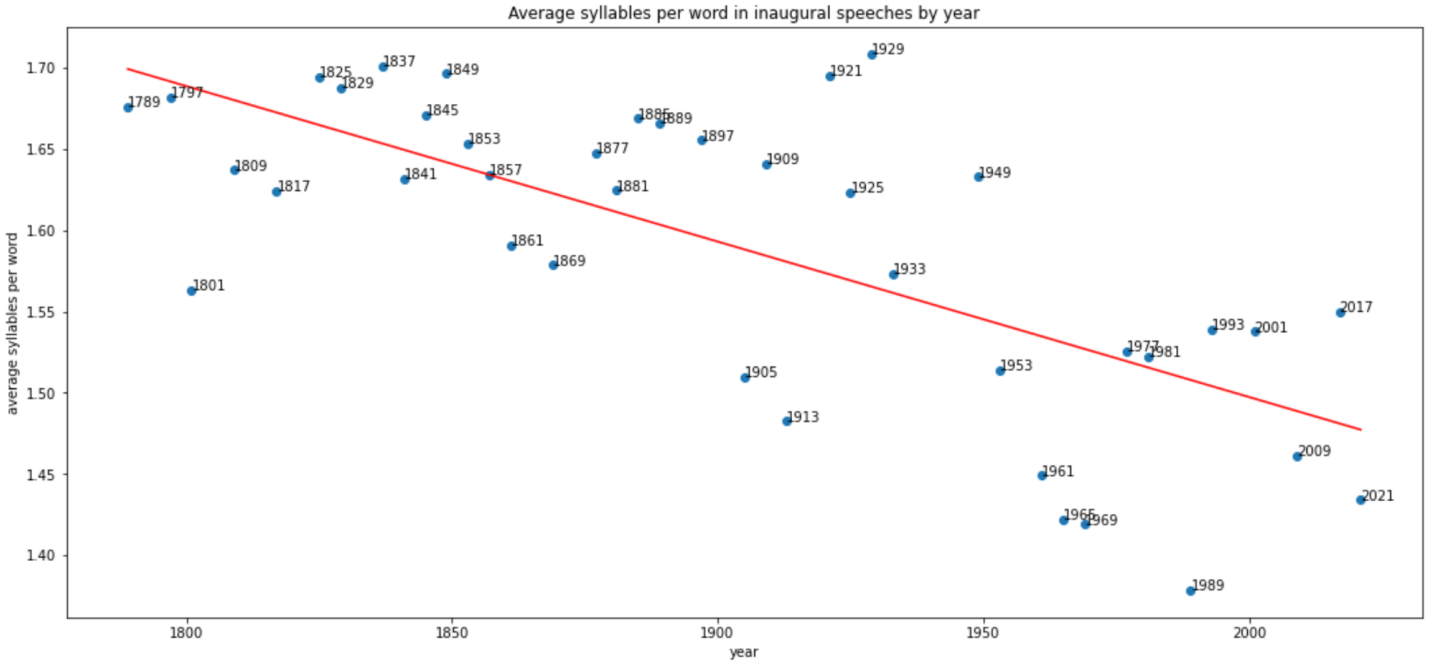


c) **Number of syllables per word** by each president in his speech.

The president that used the shortest words was George Bush (1989), followed by Nixon (1969) and Lyndon B.Johnson (1965). All of these presidents operated in a period of economic growth

following a recession in the US. The president that used the longest words was Herbert Hoover in 1929. A dip in average syllable count can be seen between 2001 and 2009 as well as 1929 and 1933, both of which are time periods in which the US economy collapsed. There is a trend showing that the length of words in syllables has been decreasing over time. It's hard to determine outliers from this plot as the data points do not stay tight around the best fit line. When using the interquartile method, we can see that there are no outliers.

(i)

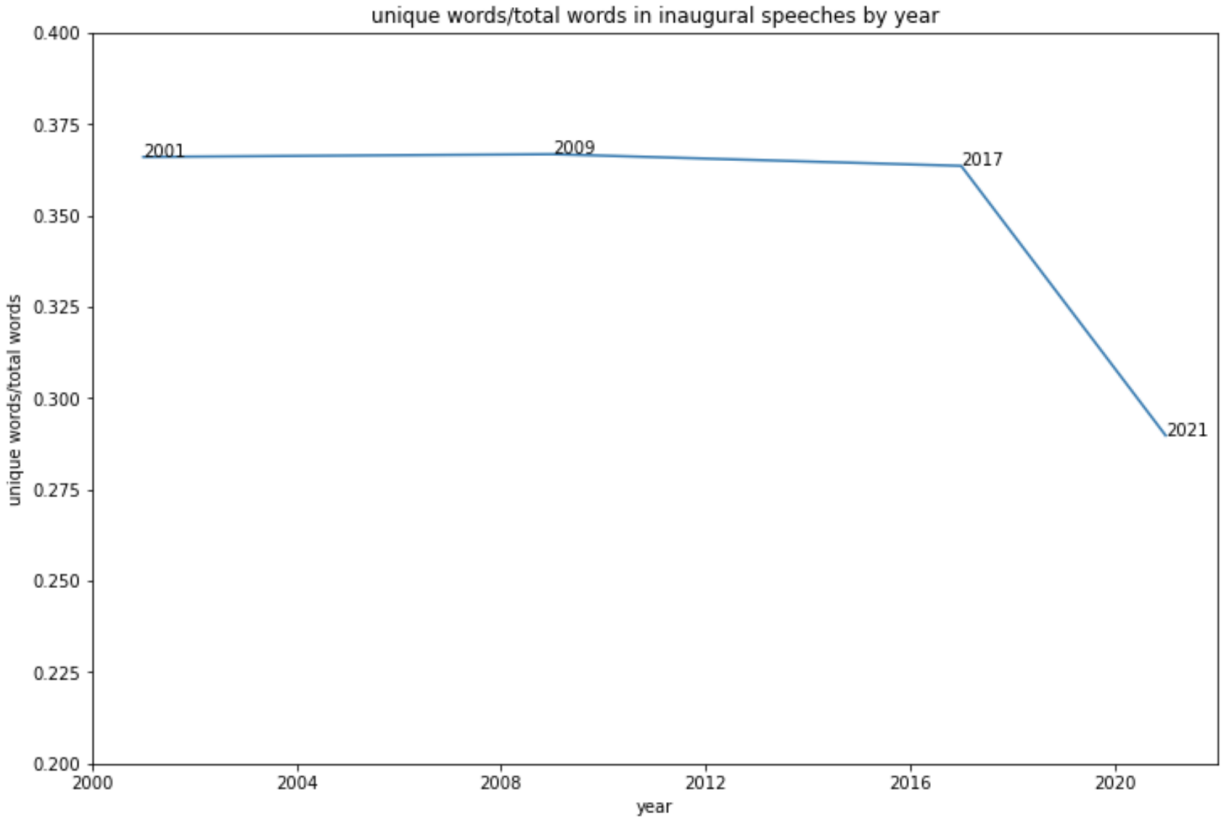


d) **Vocabulary: Number of unique words per speech**:

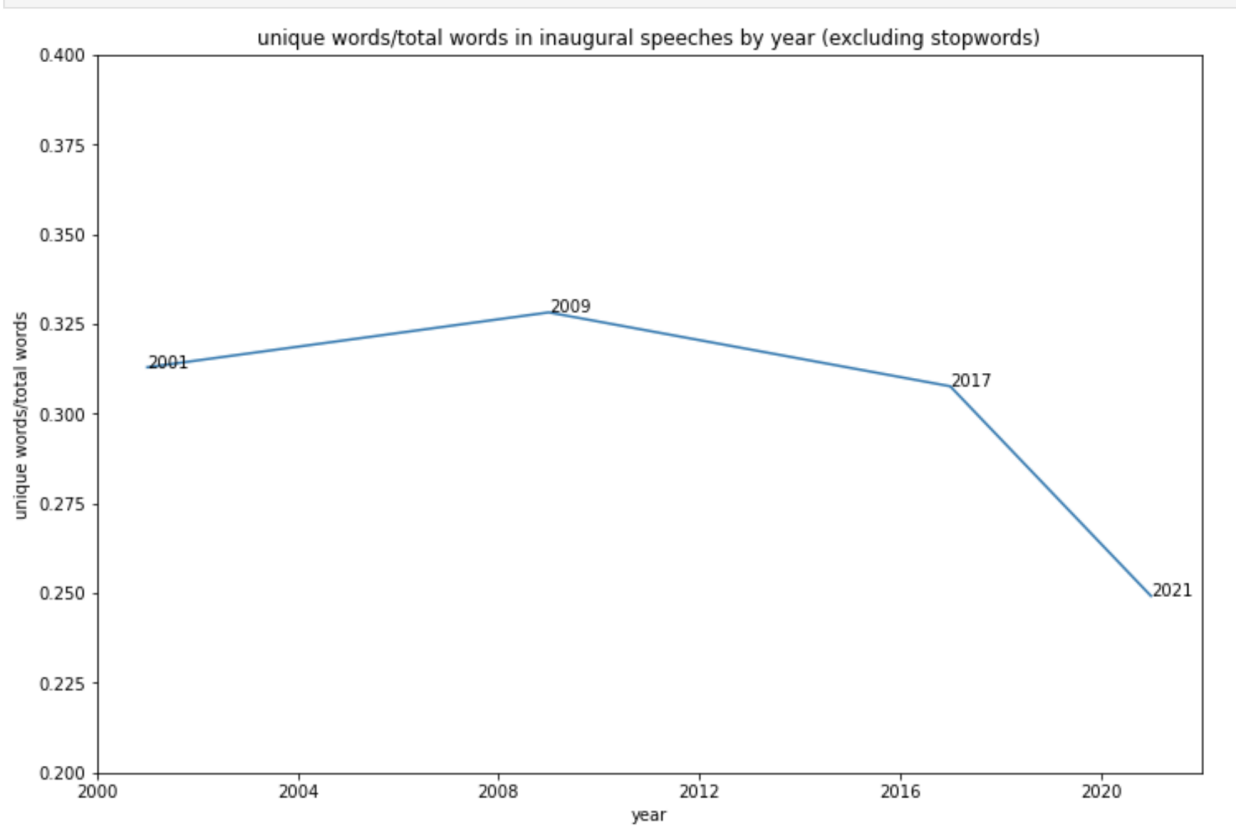
Initially, when plotting the unique words per speech by year, we see that Bush, Obama, and Trump have a very similar number of unique words. We also see that Biden has a considerably lower amount. Once we remove the stopwords (plot(ii)), we see all four points shifting lower. This makes sense because common stopwords such as "the" and "of" are not taken into

consideration. However, we see the same pattern where Biden keeps a considerably lower number of unique words. Looking at the 20 most frequent non-stop words we can see that there are a lot of recurring words including “America”, “nation”, “us, and “must”. We can also notice language that indicates whether a certain president prefers aggressive versus peaceful language or the type of politics that they employ. For example, Trump’s campaign relied a lot on a narrative that separated Americans from an “other”. We can see from the frequent use of words like “America”, “Americans”, “country”, “one”, and “protected” that there is grouping together of Americans in order to protect themselves from an “other”. Peaceful and hopeful language is more apparent in Bush and Biden’s speeches with common words including “unity”, “promise”, “freedom”, “us”, and “together”.

(i)



(ii)



(iii)

George W. Bush

	Word	Frequency
0	nation	11
1	america	11
2	us	11
3	story	10
4	citizens	9
5	country	9
6	must	6
7	every	6
8	president	5
9	common	5
10	new	5
11	many	5
12	freedom	5
13	promise	5
14	americans	5
15	know	5
16	never	5
17	courage	5
18	justice	4
19	yet	4

Barack Obama

	Word	Frequency
0	us	23
1	nation	12
2	new	11
3	america	10
4	every	8
5	must	8
6	today	7
7	people	7
8	less	7
9	world	7
10	let	7
11	time	6
12	work	6
13	common	6
14	generation	5
15	day	5
16	know	5
17	spirit	5
18	god	5
19	words	4

Donald J. Trump

	Word	Frequency
0	america	20
1	american	11
2	people	10
3	country	9
4	one	8
5	nation	8
6	every	7
7	world	6
8	great	6
9	back	6
10	never	6
11	new	6
12	president	5
13	many	5
14	today	5
15	protected	5
16	across	5
17	right	5
18	dreams	5
19	god	5

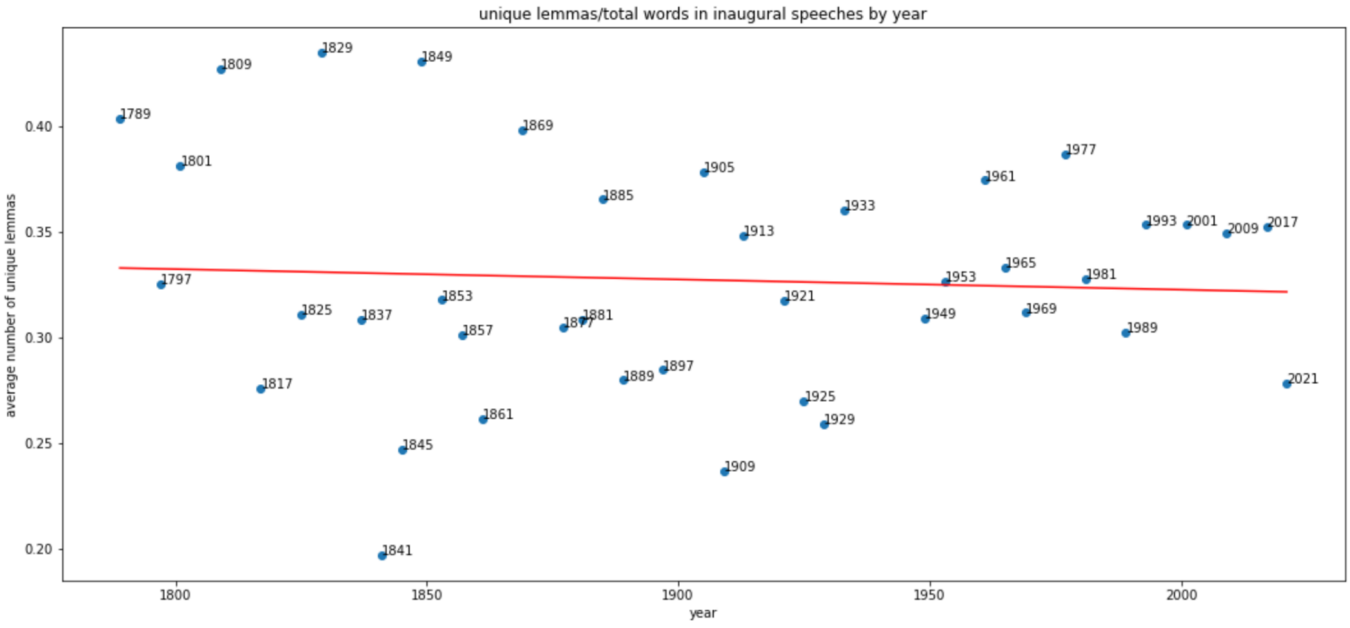
Joseph R. Biden

	Word	Frequency
0	us	26
1	america	20
2	one	15
3	nation	15
4	americans	11
5	democracy	11
6	people	11
7	today	10
8	much	10
9	know	9
10	story	9
11	another	9
12	history	8
13	american	8
14	must	8
15	world	8
16	unity	8
17	president	7
18	days	7
19	together	7

e) **Active Vocabulary of the US Presidents**:

There isn't a clear correlation between the active vocabulary (unique lemmas/total words) of the US presidents as illustrated by the line of best fit which doesn't necessarily follow a lot of the points. Using the interquartile method, we find that William Herry Harrison's 1841 speech is an outlier as it is much lower than the other speeches. The president who had the highest percentage of unique lemmas was Andrew Jackson who had an active vocabulary of 43.5%. The president who had the lowest percentage of unique lemmas was William Herry Harrison who had an active vocabulary of 19.7%. Although there is no clear correlation between active vocabulary and time, we can notice that the data points have gotten generally closer to the best fit line as time passed. This may be explained by the fact that there is more of a blueprint for what an inaugural speech looks like and therefore speeches are becoming more standardized in terms of their active vocabulary.

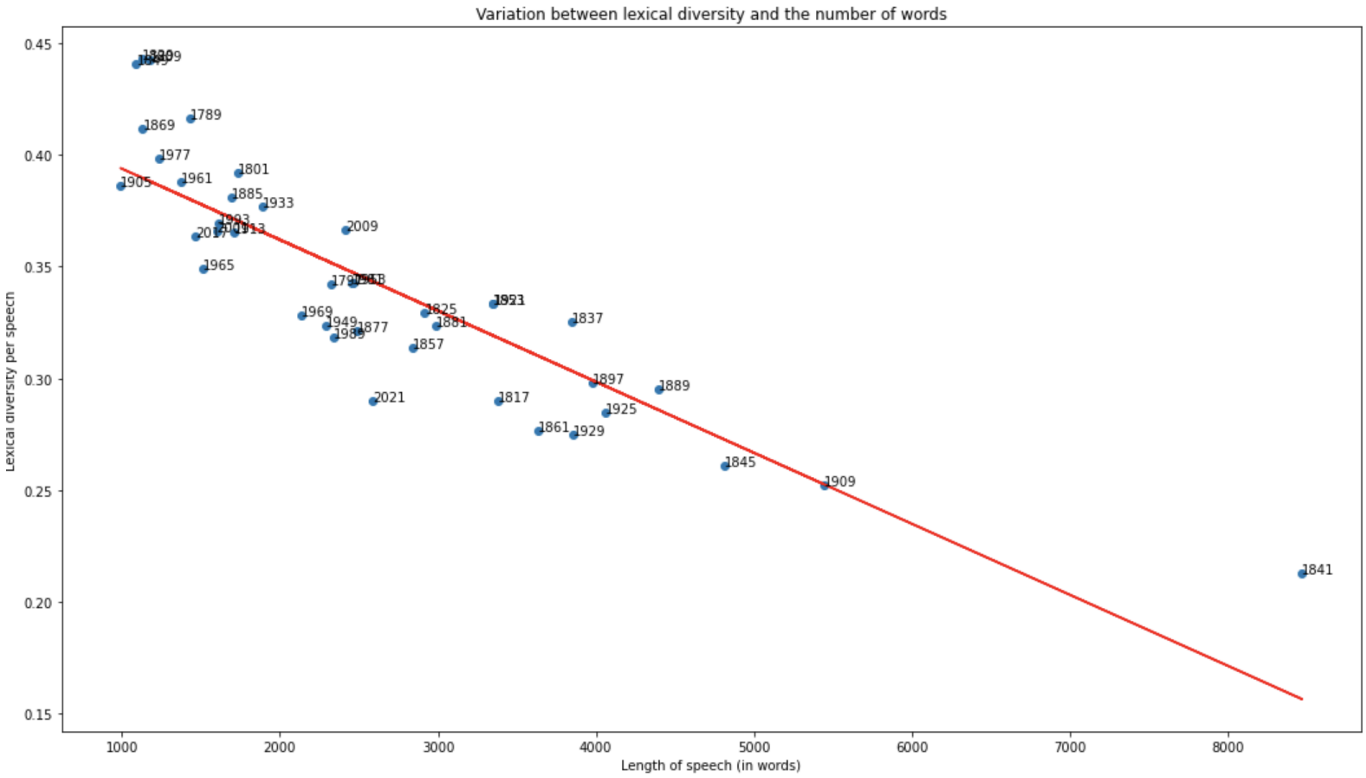
(i)



f) **Lexical Diversity**

The lexical diversity in inaugural speeches decreases with the number of words used. This is illustrated by the downward sloping line of best fit. It shows that the speeches that are higher in length have lower lexical diversity. For example, William Henry Harrison who had the longest speech in words had the lowest lexical diversity score of 0.29. The opposite can be said about those speeches that are lower in length. For example, Andrew Jackson who has the shortest speech in words (refer to part(a)), had one of the most diverse vocabularies (a lexical diversity score of around 0.44).

(i)

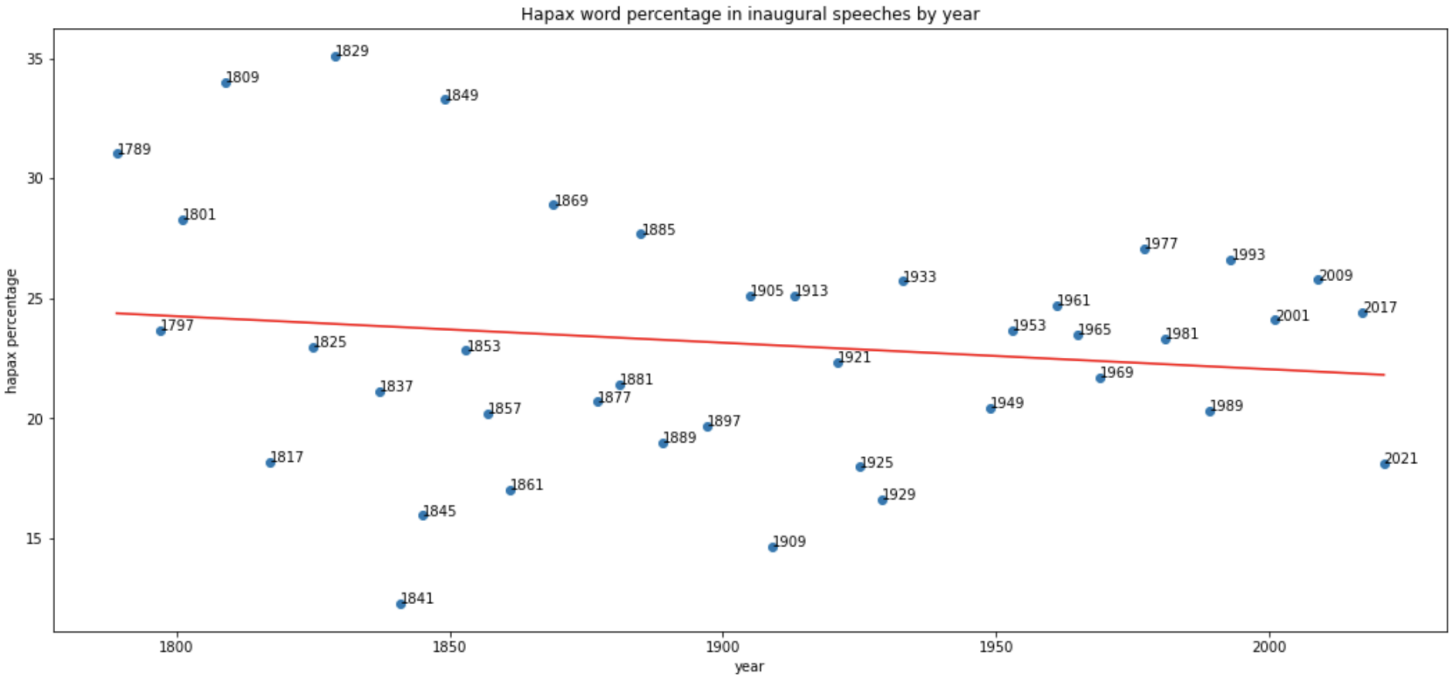


g) **Hapax Legomena and Words with a Frequency of at most 4:**

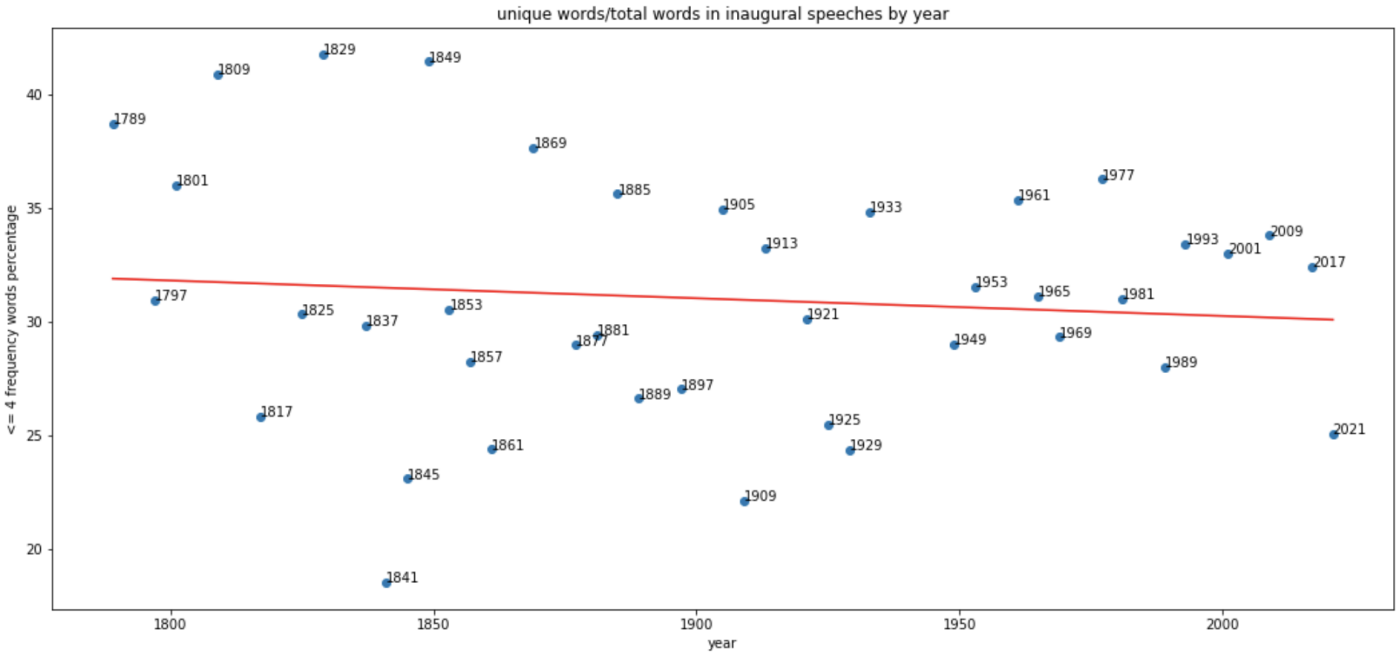
Some of the hapax words used in the speeches are “demonization”, “distinguished”, and “joblessness”. These words tend to have a higher number of syllables, often three or more. These words also tend to be more uncommon in everyday speech. Another category of hapax words are names of important figures or politicians such as “Pelosi”, “Clinton”, and “Bush”. It makes sense that a president wouldn’t want to mention other politicians too much in a speech where they must make themselves sound presidential. Using the interquartile range method, we can see that there is one outlier-Andrew Jackson. His 1829 speech has an hapax percentage of 35.11%. This can be partially explained by the fact that Jackson has a very low word count (as seen in plot (a)(ii)). It is more likely that a word he used is used once as there are simply fewer words in his speech. The president with the lowest percentage of hapax words is William Henry Harrison in 1841 with 12.2% Hapax words. This makes sense because he has the longest speech as mentioned in part (a).

The plot for words with frequency <=4 has no outliers and a best fit line that doesn’t provide a clear relationship between the frequency of unique words and the year. A few examples of words in this class include “elected”, “distinguished”, as well as names of other politicians. These words have either the same number of syllables as the hapax words or fewer. The president with the highest number of words with frequency <=4 is Andrew Jackson with 41.76%. The president with the lowest number of such words is William Henry Harrison with 18.51%. This makes sense considering they also had the highest and lowest percentage of Hapax words

(i)


Hapax word percentage in inaugural speeches by year
hapax percentage
year
35
30
25
20
15
1800
1850
1900
1950
2000
1789
1797
1801
1809
1817
1825
1829
1837
1841
1845
1849
1853
1857
1861
1869
1877
1881
1885
1889
1897
1905
1909
1913
1921
1925
1929
1933
1949
1953
1961
1965
1969
1977
1981
1989
1993
2001
2009
2017
2021

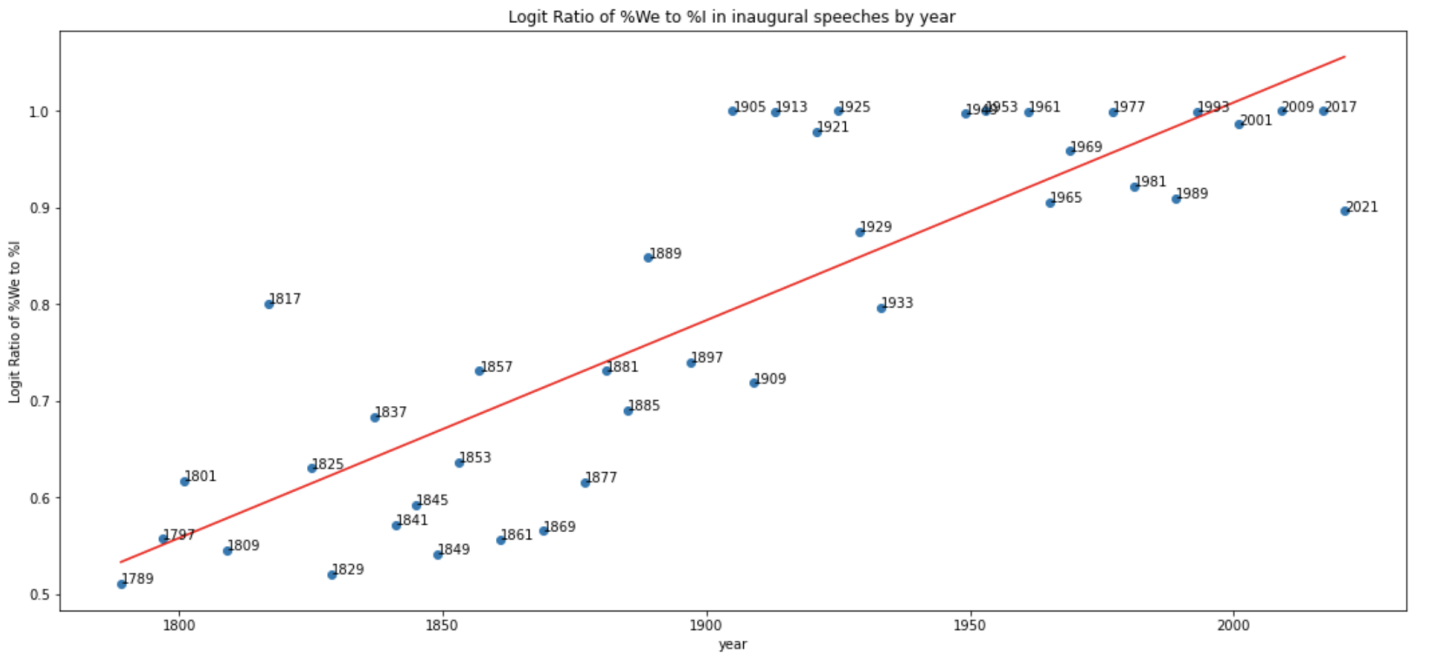
(ii)


unique words/total words in inaugural speeches by year
<= 4 frequency words percentage
year
40
35
30
25
20
1800
1850
1900
1950
2000
1789
1797
1801
1809
1817
1825
1829
1837
1841
1845
1849
1853
1857
1861
1869
1877
1881
1885
1889
1897
1905
1909
1913
1921
1925
1929
1933
1949
1953
1961
1965
1969
1977
1981
1989
1993
2001
2009
2017
2021

h) **The Ratio of Percentage Usage of "We" to Percentage Usage of "I"**

The presidents with the lowest “We” to “I” ratio are “ George Washington”, “John Adams” and “Andrew Jackson”. It makes sense that these presidents who were elected during a young democracy would want to distinguish themselves as capable and would therefore refer to themselves often. The presidents with the highest “We” to “I” ratio are Trump, Obama, and Clinton. This can be explained by how recent presidents have tried to create ingroups in order to increase unity. There is a positive correlation between this ratio and time. This means that as time progressed, presidents adopted the use of “we” over “I”. Using the interquartile method, we see that there are no outliers in the data.

(i)



i) **Keywords Over Time**

When looking at the trends for keywords (plot (i)), we can see that it varies between each of the four words. There is a positive trend in the use of "America" with large increases in Bush Sr's and Trump's speeches. There is an overall positive trend with the word "America" throughout the presidential speeches. We can see that the president who used the word "America" most frequently is Donald Trump, a president that pushed for the success of an individual "America" against the rest of the world. The increase in the use of "America" can also be related to the growing patriotism and nationalism in the population. The word "democracy" can be seen to have an uptick around the time of Americas strong opposition to communism (eg: the Cold War in Russia). The same explanation can be given for the word "freedom" which many Americans associate with democracy. The word "protection" isn't used as commonly as the other keywords, especially in recent years. This may be because current America prides itself on being offensive and strong as opposed to defensive.

(i)