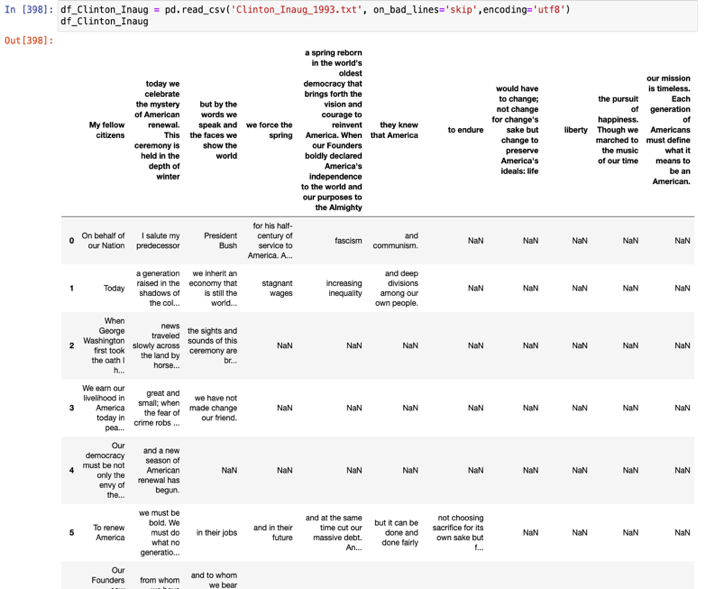
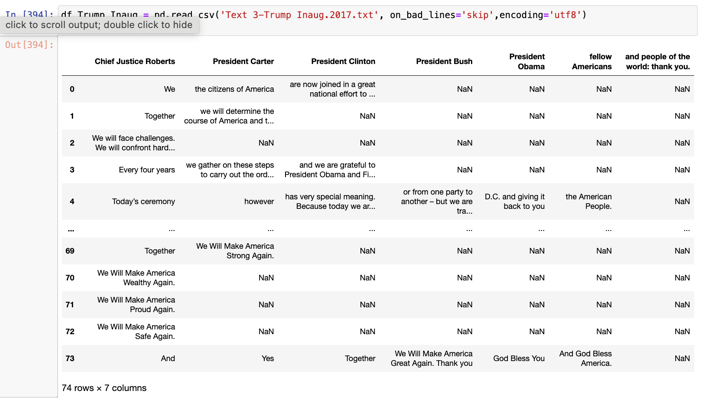
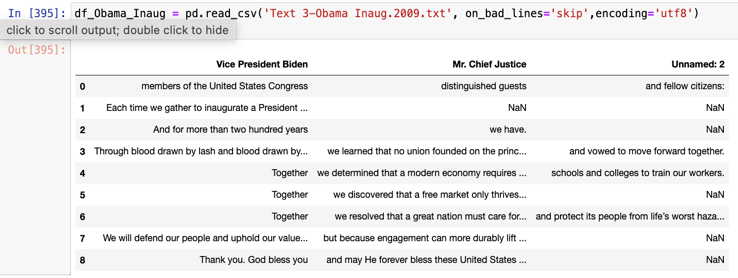
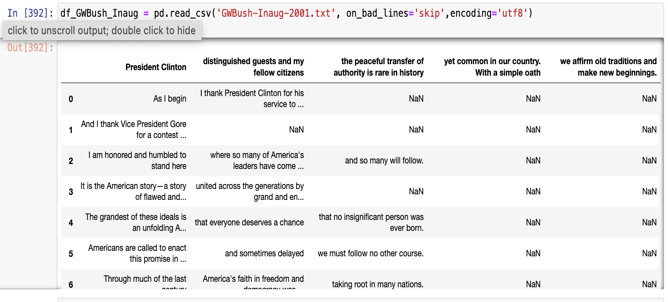
Comparative Text Analysis of U.S. Presidential Inauguration Speeches

I started my analysis by reading in multiple texts. I then settled on the following four texts: [Bill Clinton’s inauguration speech from January 20th 1993](https://millercenter.org/the-presidency/presidential-speeches/january-20-1993-first-inaugural), [G. W. Bush’s inauguration speech from January 20th 2001](https://millercenter.org/the-presidency/presidential-speeches/january-20-2001-first-inaugural-address), Barack Obama’s inauguration speech from January 20th 2009 and lastly Donald Trump’s inauguration speech from January 20th 2017. The reason I chose these speeches is because I wanted to analyze two speeches from former Republican presidents and two speeches from former Democratic presidents at a certain important point in their political careers, and what could possibly be more important than their very first inauguration speech? We didn’t have Clinton’s and Bush’s speeches available on Courseworks so I found them on the UVA Miller Center’s website. They have an amazing collection of speeches of many former presidents so I highly recommend to check them out if interested. I will compare the texts further throughout this lab.





*Creating bag-of-words representations by cleaning and tokenizing the speech texts*

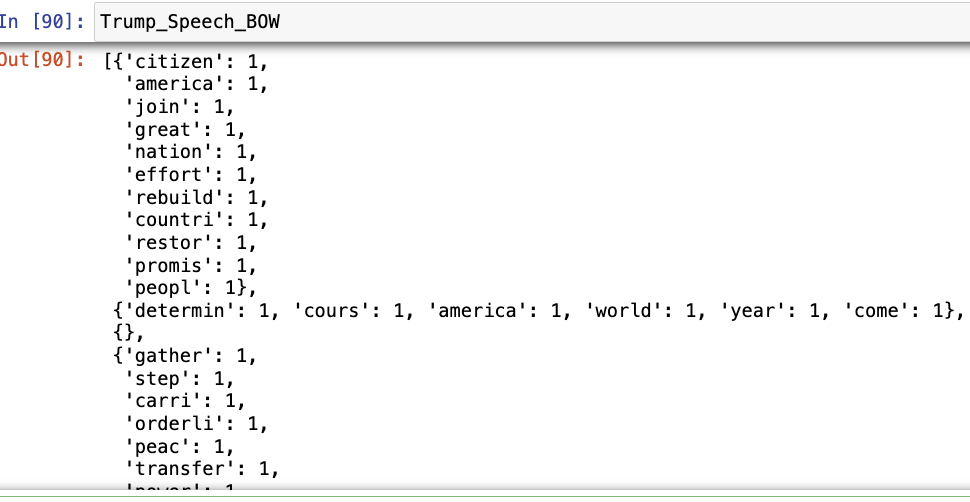
I will now create bag of words out of my text data. Bag of words are a simple way to transform text data into numerical data in order to use it for our analysis moving forward. As seen below, I combined the data preprocessing and bag of words conversion into one code line:

The data preprocessing is done through lower casing, removing punctuation, removing stop words, removing extra white space and stemming. We do that in order to remove the “extra noise” around our data. No valuable information goes missing through the process which is great.

The bag of word model was created with the NLTK library (there are multiple ways to create a model, for example with CountVector). The set is stored in the variable “vocab” as it contains all the individual words in our list that’s called “text\_data”. This is important for our next step as every single word in every single text document in our “text\_data” list is being counted and then stored in and represented by dictionaries. These dictionaries then are being stored as a list in the “bow\_model” variable.



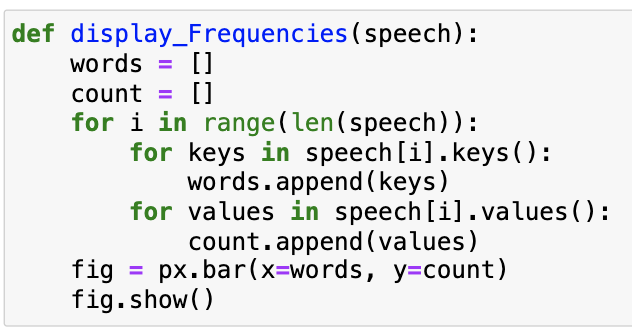
Below you can see the actual bag of words for Trump’s speech. You can see it as a list with dictionaries stored inside (separated by the curly brackets). Every element has a key (word) and a value count (number):



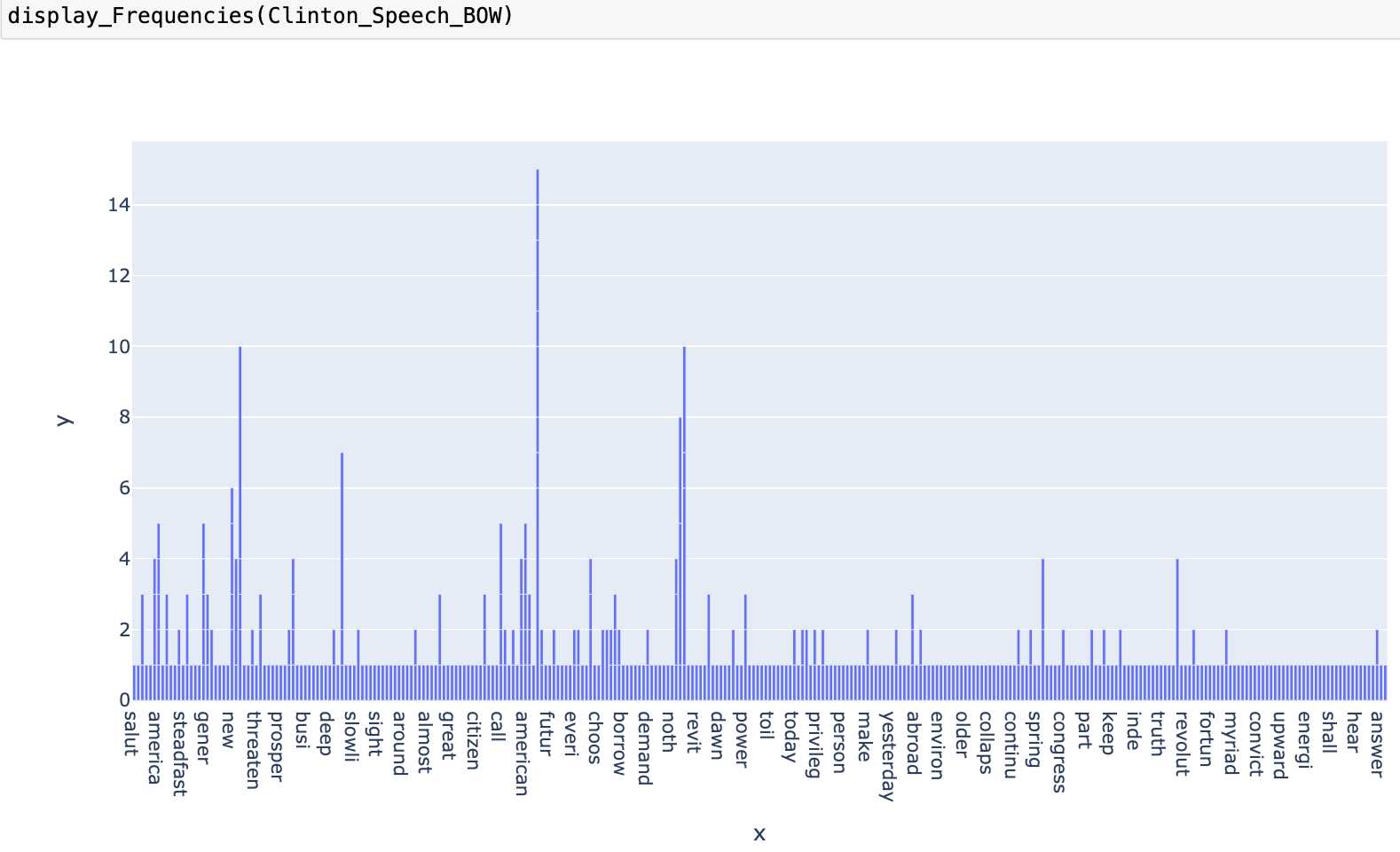
*Calculating and comparing relative word frequencies across speeches*

Next up I will generate the relative word frequencies for the bag of words as this will make it very convenient to compare the bag of words under one another. This is generally a great method to look if there are any key themes or patterns in one of our speeches or just to gain some insight about the general content. Below you can see that I visualized the overall word frequencies for each bag of words individually with a bar chart and then the top 15 most frequent words of every bag of words together in one table. Let’s start with the bar charts:

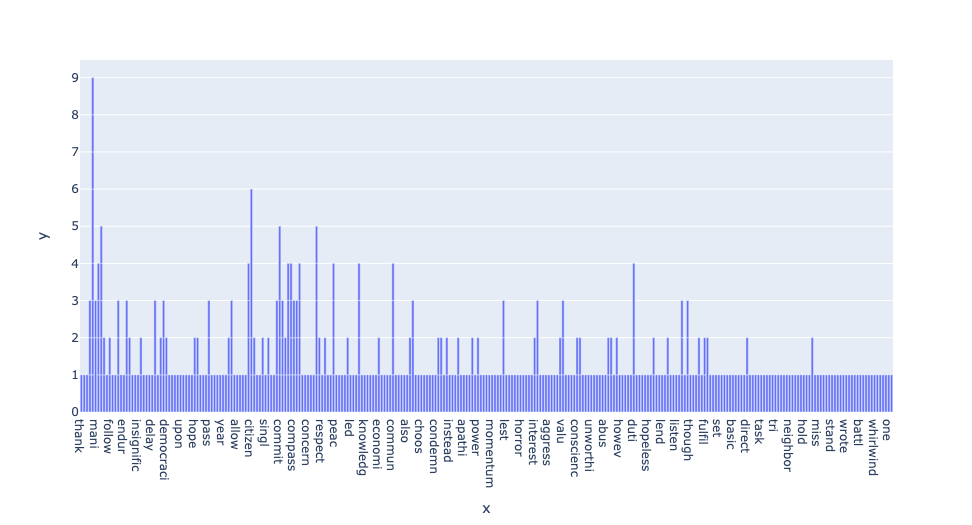
Looking at the diversity of the relative word frequency of each bag of words (i.e., how many different words were used more than once during a speech), it is interesting to see that Barack Obama’s speech is by far the least diverse one whereas Bill Clinton’s seems to be the most diverse one. G.W. Bush and Donald Trump are both in the middle and relatively similar to one another when it comes to their speeches.



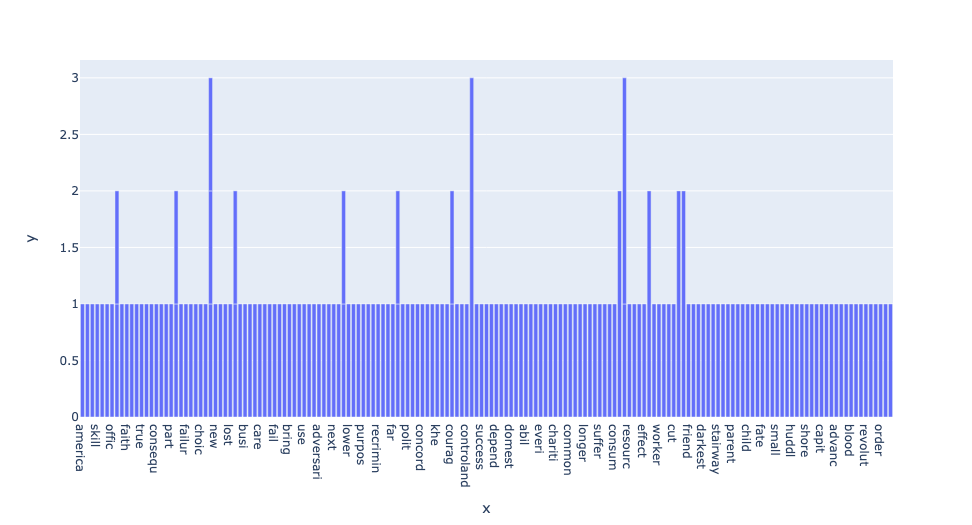
Clinton 1993



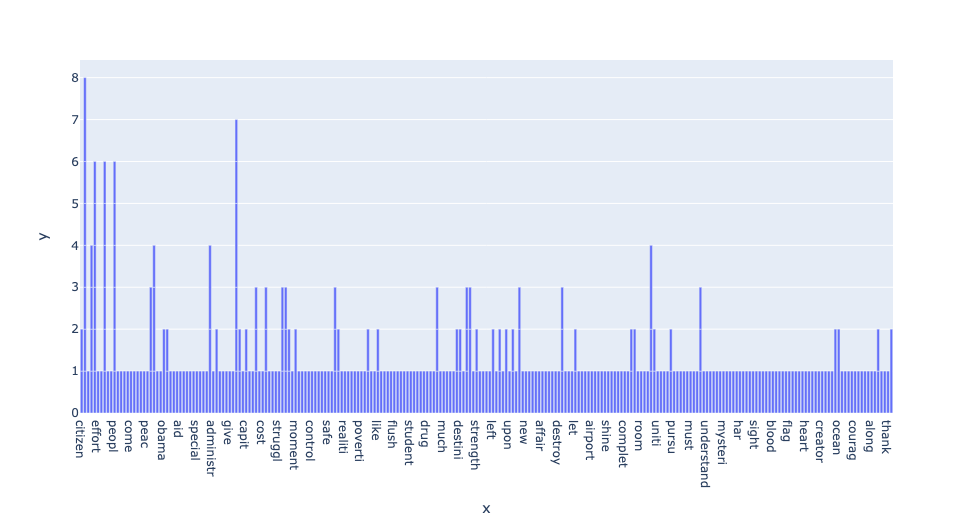
G.W. Bush 2001



Obama 2006



Trump 2017

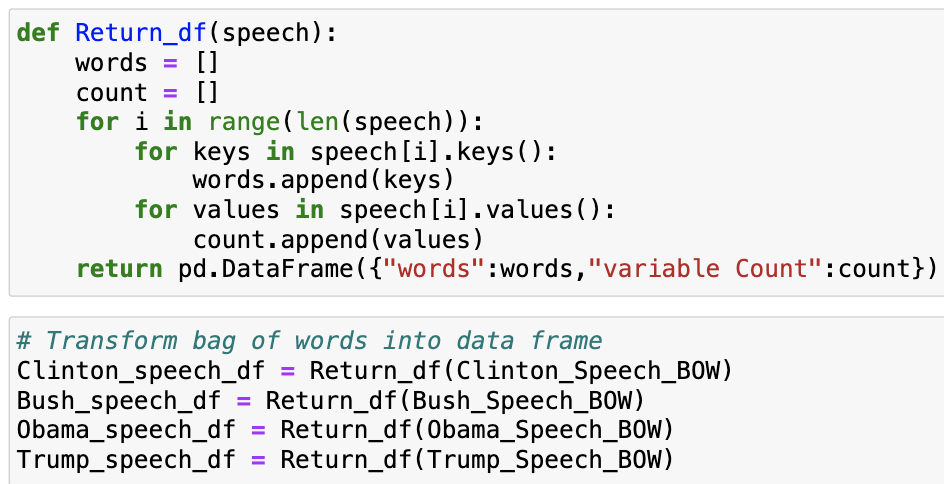
Looking at the word frequencies in more detail, it shows further that Bill Clinton’s speech also has the highest word counts per word (with multiple words counted 4 times upwards to 6 times) compared to the other speeches. Trump, in comparison, never uses a word more than twice in his speech.

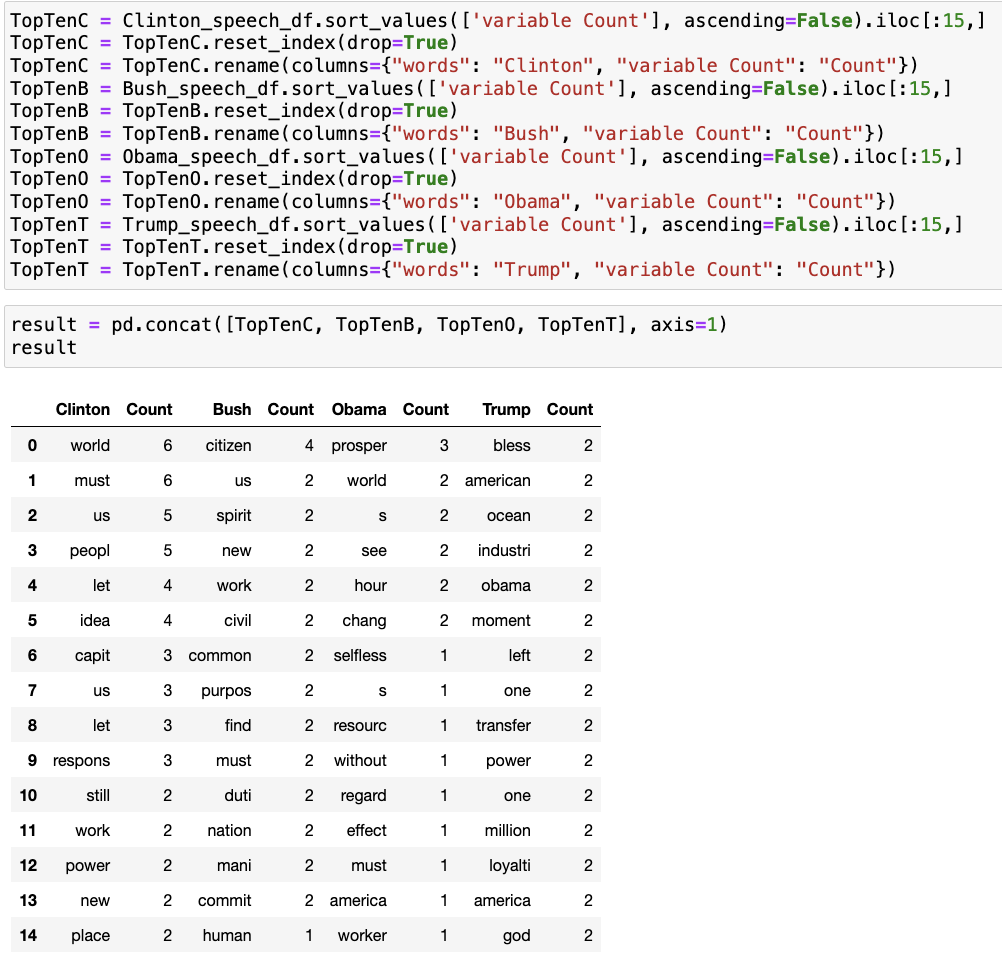
I think in order to draw any conclusions in regard to the context of a speech (i.e., key topic) it is worth to look at each frequency table individually: For Bill Clinton’s speech for example, we already know through the bar charts that he uses many words in his speech at least twice. I would therefore only pay attention to the words that he uses at least four times during his speech. These words are world, must, us, peopl, let, idea.

G.W. Bush’s speech is a little harder to contextualize because even though he uses many words more than twice in his speech, that is essentially it. The only word he actually uses very prominently is the word “citizen”.

Even though the least diverse one when it comes to word frequency of the whole speech, Barack Obama’s speech is now again a bit easier to contextualize because there are such few words that he uses more than once: prosper, world, see, hour, chang.

Lastly, Donald Trump’s speech is probably the hardest to conceptualize out of the four because, as mentioned before, he does have some diversity in word frequency in his speech but he doesn’t use any word more than twice. What helps a little bit in his speech though is that he uses words that are very easy to conceptualize in and of itself, for example “America”, “god”, “bless”, “power” and “loyalty”.





*Observing and interpreting differences in word usage, connecting patterns to political context and party affiliation*

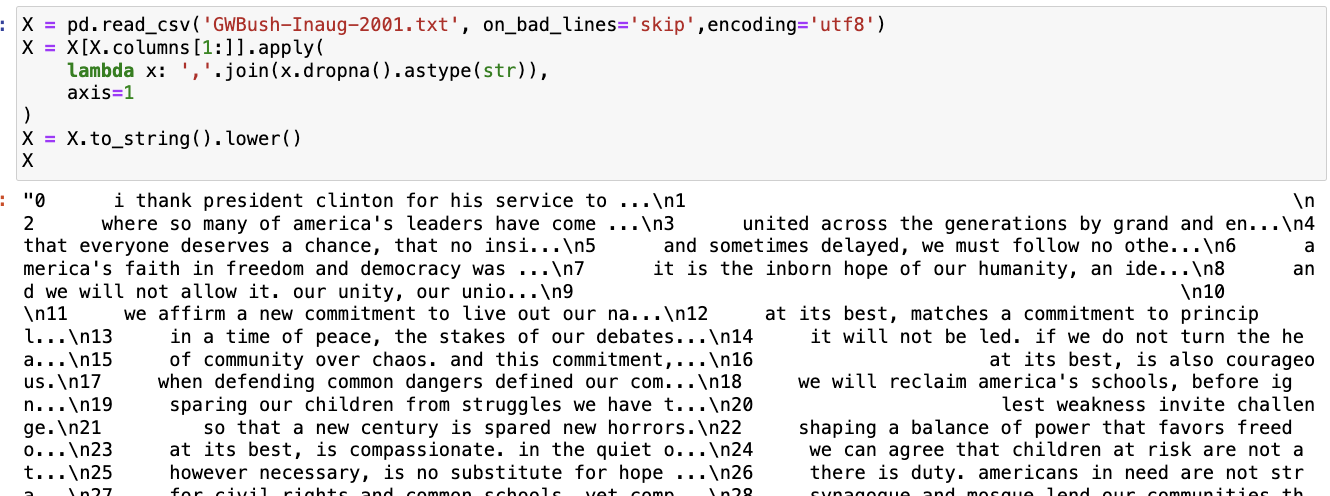
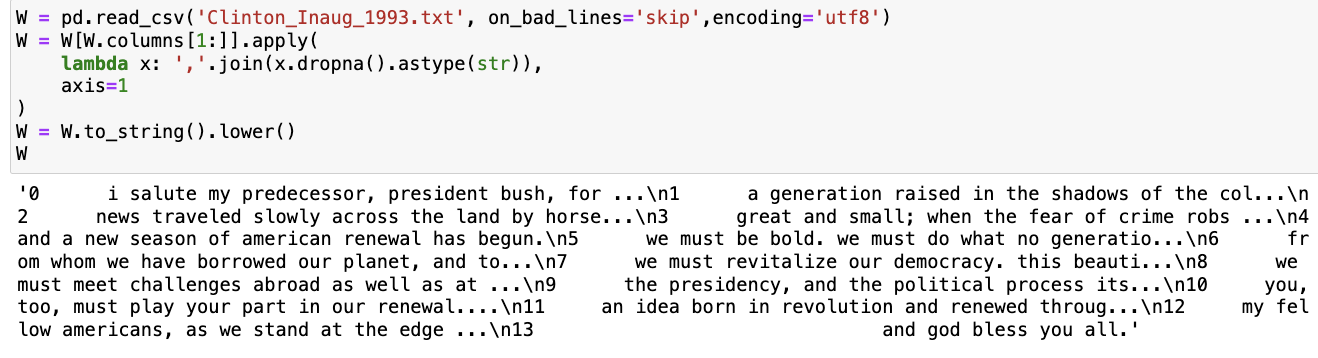
I expected to see that the bag of words of Bill Clinton and Barack Obama would be similar as well as the ones of G.W. Bush and Donald Trump. My expectations were based off of the theory that being elected by the same party voters indicates that you care about the same topics as a president. Consequently, I thought that the speeches between the different parties will differ a lot.

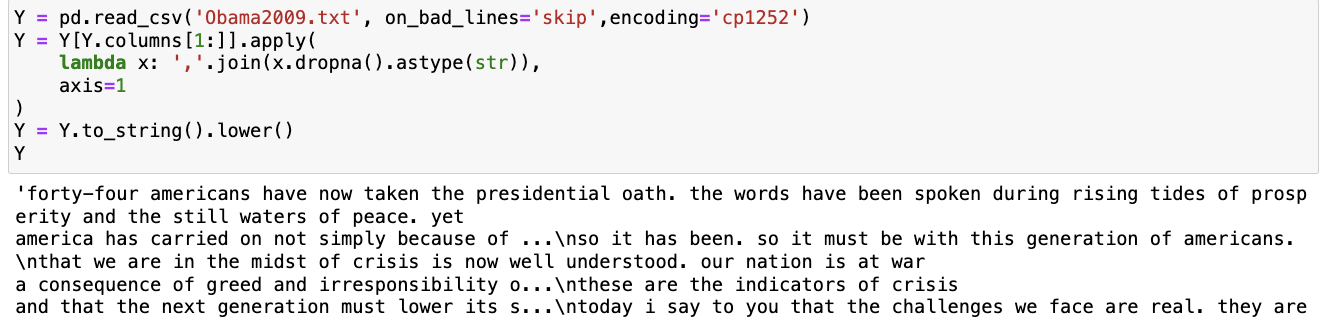
What actually came out during the analysis of the word of bags frequency is that in general all four speeches differ a lot from each other, no matter if it comes to the diversity of relative word frequencies or frequently used words in detail. There is really not much similarity at all which is a little shocking to me.

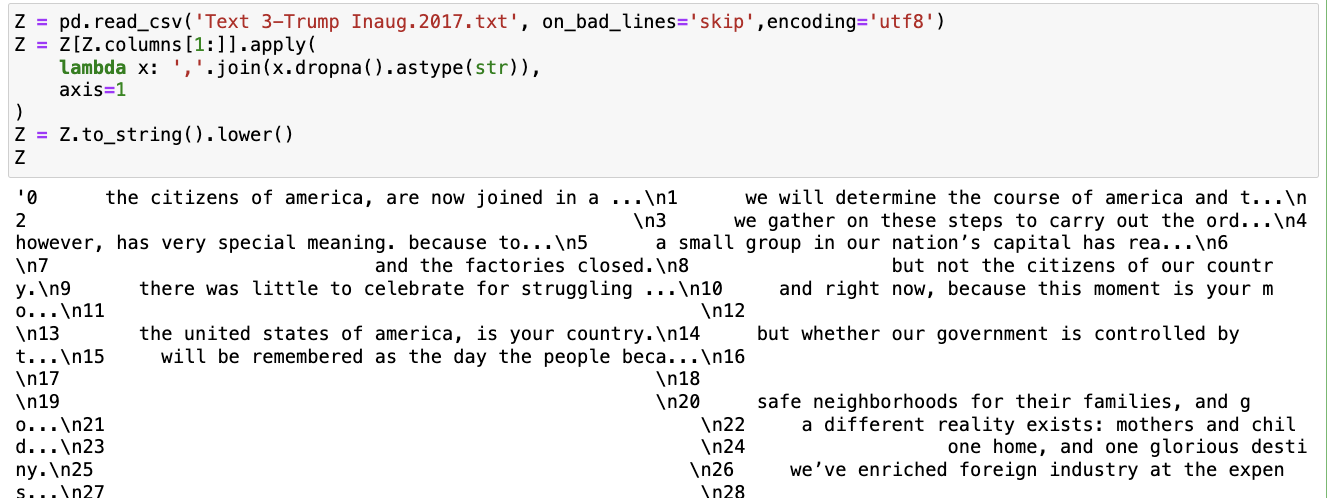
*Comparing speeches using statistical metrics like cosine similarity to measure linguistic overlap and divergence*

I will now compare the texts through the very popular cosine similarity. Cosine similarity is used to compare texts semantically in size.

I had problems with running my code using my bag of words from the analysis so I decided to recreate the bag of words again in a different way that included the function for cosine similarity in one code line which ended up working. I did it as follows:



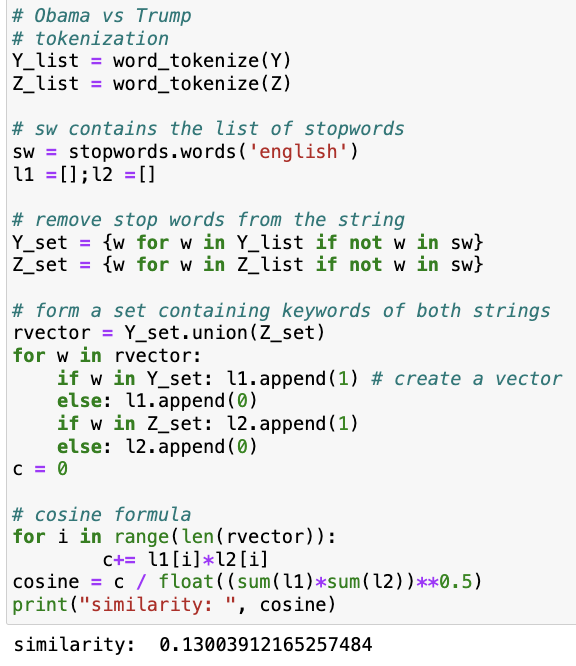
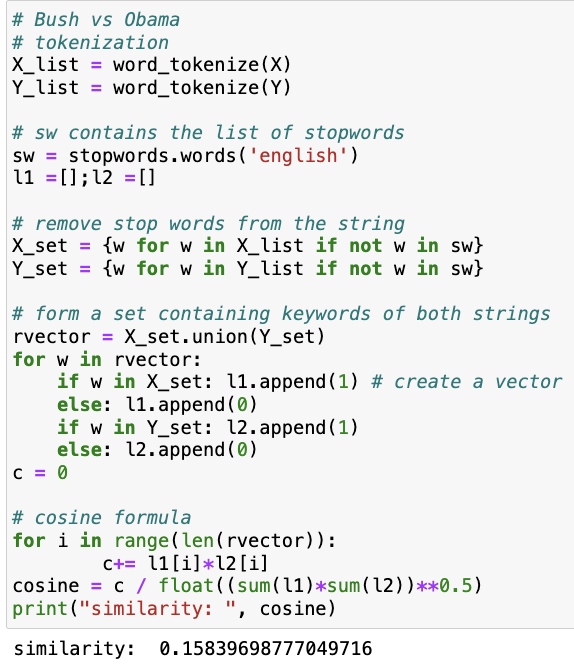
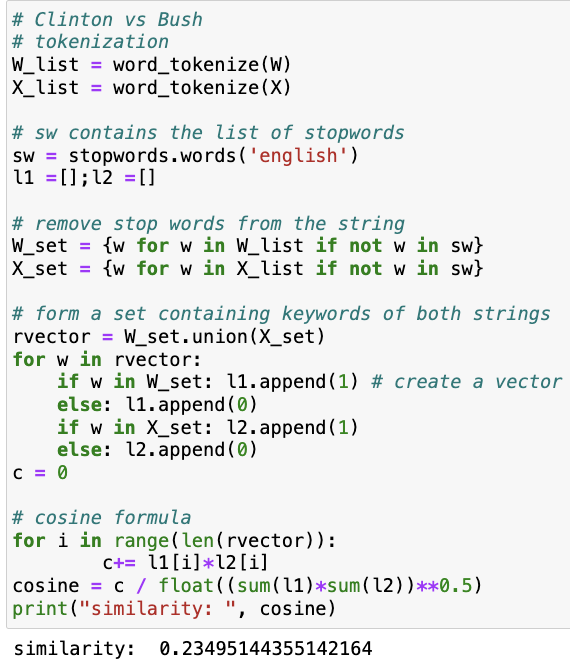


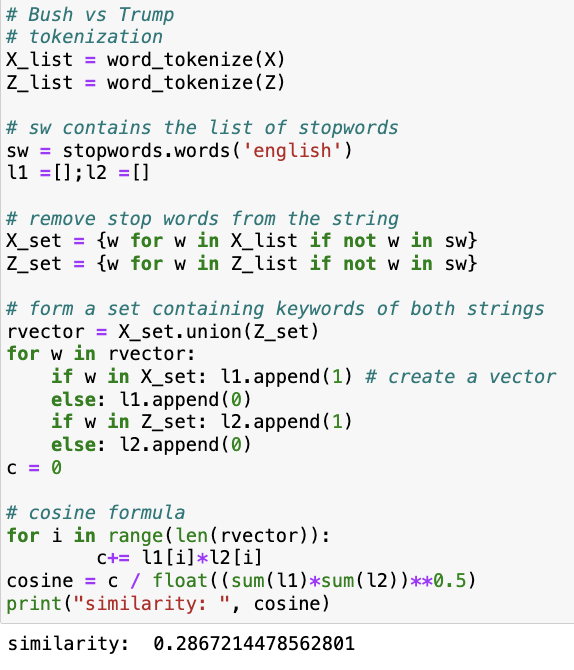
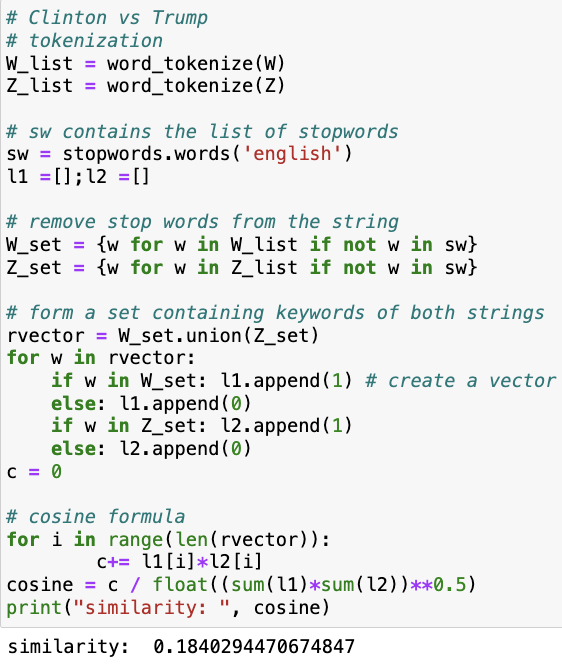
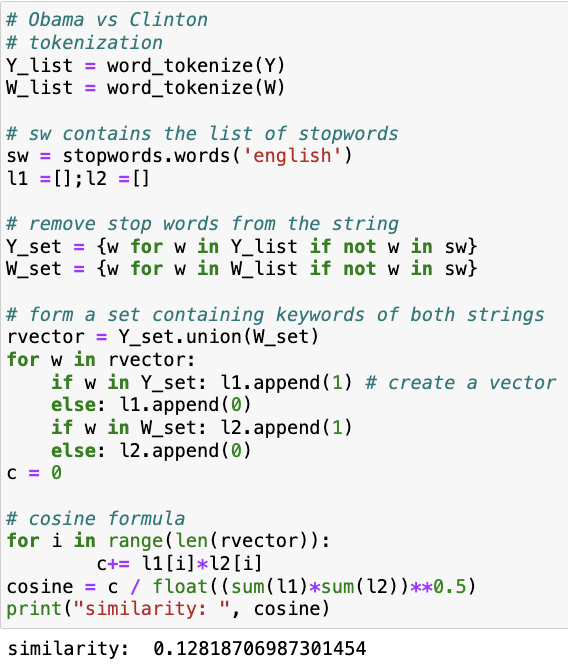


I compared the speeches all one by one below. A cosine similarity can be between 0 and 1 meaning 0 that there is no similarity at all and 1 means entirely similar. We can see that the speeches are not very similar to another all throughout which confirms our findings from questions three and four.

The speeches that were most similar to each other were the speeches of G.W. Bush and Donald Trump (0.287) and Bill Clinton and G.W. Bush (0.235). The speeches that were the least similar were the ones of Clinton and Barack Obama (0.128) and Barack Obama and Donald Trump (0.13).

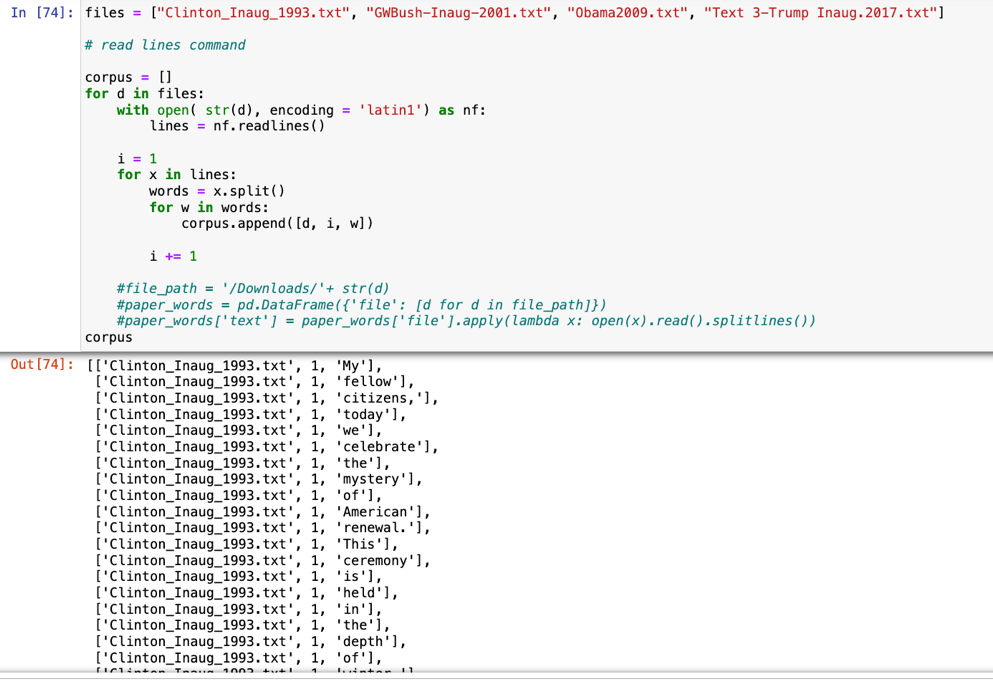
And here again, I find it very interesting that, despite all of the speeches being generally pretty dissimilar from one another, the speeches of Bill Clinton and Barack Obama are the most dissimilar of them all since they are both Democrats. I would have expected the speeches of politicians from the same party to be very similar.

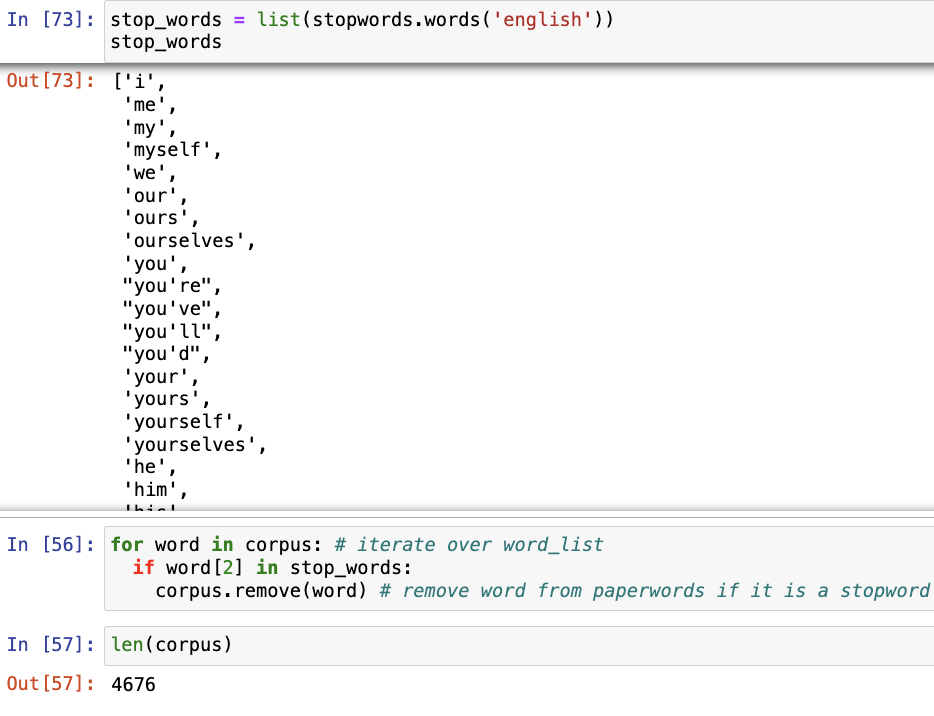


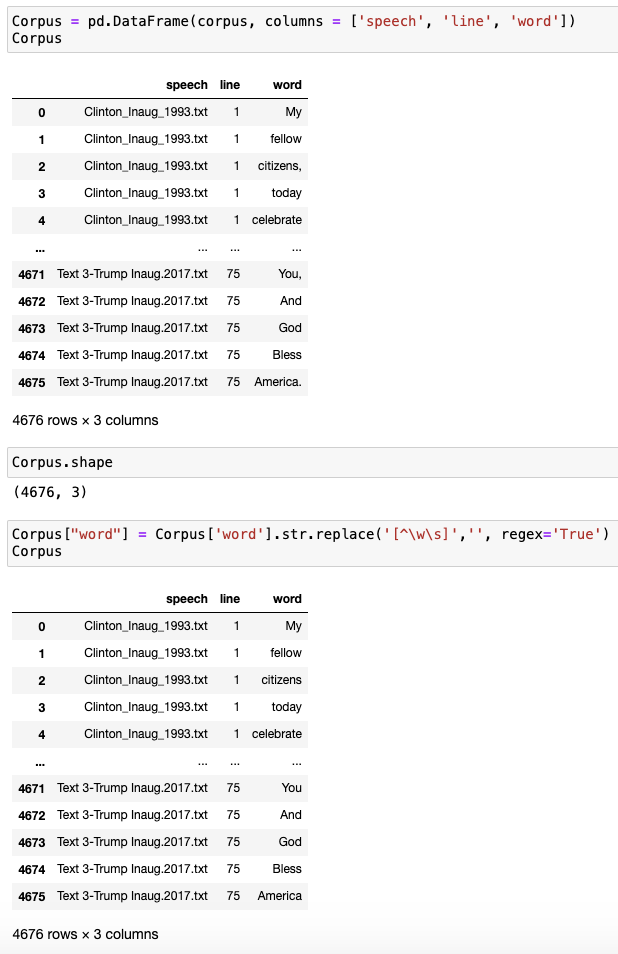


*Sentiment analysis*

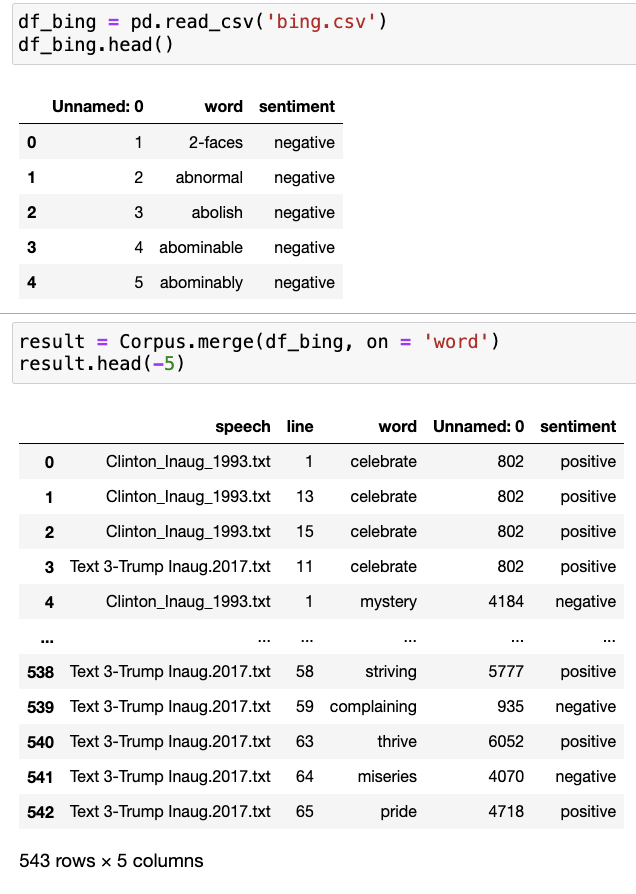
As a next step I am going to conduct a sentiment analysis of the four speeches. I think this is perfect for my case because sentiment analysis is supposed to show if the emotional tone and therefore message is positive or negative. In the context of a political speech, the tone can tell us how that person feels about a certain topic. In terms of an inaugural speech, the tone of the speech can tell us a lot about what topics will become very important moving forward to the new president, how they feel about it and also what they think about the current state of the country. I also generally expect the tones of speeches to differ between parties (with Republicans potentially more “extreme” in the positives and negatives and Democrats more “balanced”).



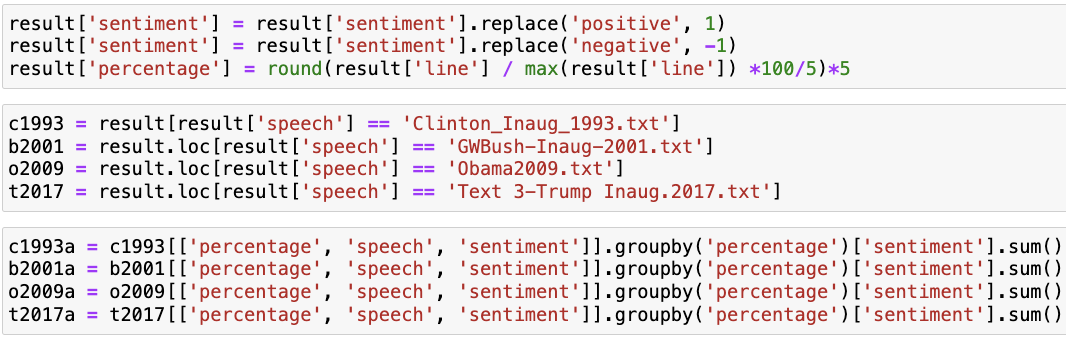


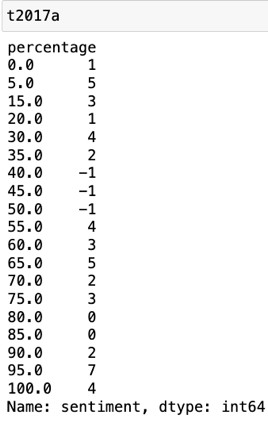
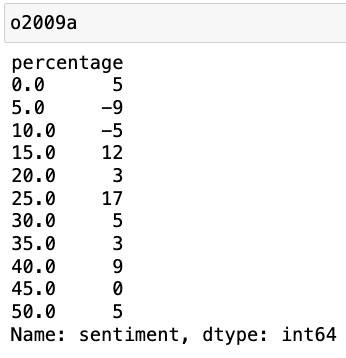
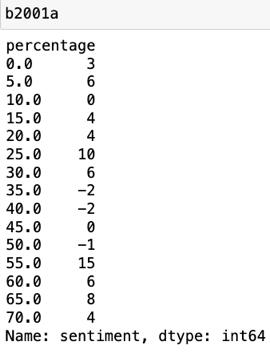
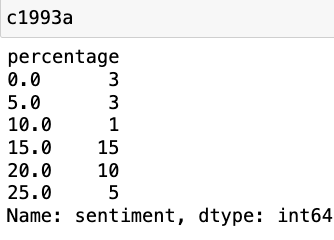


After cleaning the text data again and now starting with the actual sentiment analysis, I first obtained the sentiment list from our NLTK module (that our TAs provided) and turned it into a data frame. I then merged the sentiment list with the text data in order to separate positive from negative words.

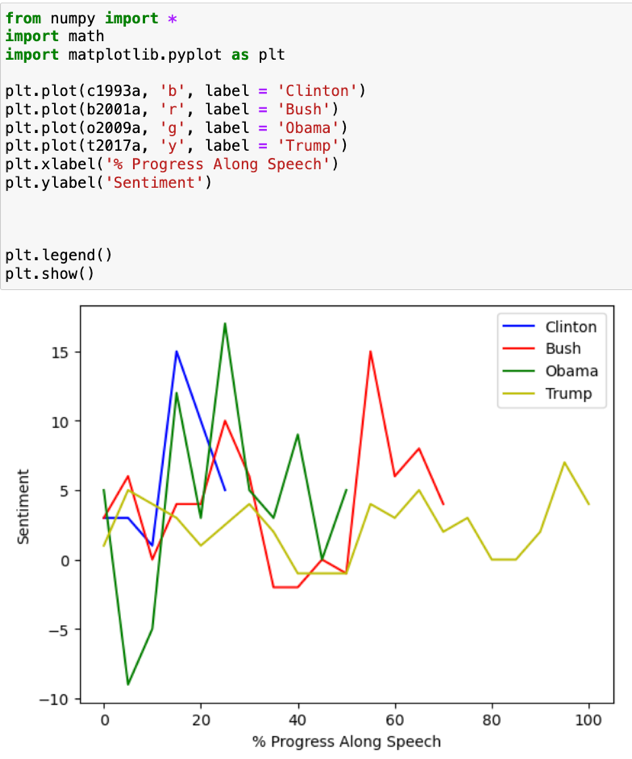


Here I separated the length of the speech into % parts to see how the sentiment scores are being distributed over it by speech individually. Keep in mind that the speeches are of different length.





Now that the machine has all the data and knows which words to count as negative words and which as positive ones, it shows us a really nice sentiment analysis that is further visualized by the graph below:



And again, my theory was entirely wrong: Barack Obama has the most emotionally diverse speech, Bill Clinton has the most positive speech and Donald Trump has the least emotionally diverse speech. G.W. Bush’s speech is somewhere between Bill Clinton and Donald Trump. I’m very surprised by this, especially by Donald Trump because I always see him as an extremely emotional speaker. There also again seems to be no similarity between people of the same party which continues to baffle me. Sure, the speeches differ in terms of sentiment, but there are absolute no polar opposites. What makes sense to me though is that because these are all inaugural speeches, they are all generally uplifting/ positive. The only one that starts out with some negative sentiment is Barack Obama’s speech but that quickly changes throughout his speech and most of it is very positive.

*Wordclouds*

Word clouds are a helpful way to visualize how often certain words appear in each speech. The more a word is used, the bigger it shows up in the cloud. Looking at the four presidents, there isn’t a ton of overlap in the language they use, especially when comparing across party lines. Clinton and Obama tend to focus more on ideas like service, community, and work, while Bush and Trump lean heavily into national values and strength. That said, some themes do repeat, like the emphasis on “america” and “nation,” which stand out clearly in both Bush’s and Trump’s speeches.

Bill Clinton 1993 G. W. Bush 2001



Obama 2009 Donald Trump 2017

