Diego Valdes

IST 736

Home Work 4

11/2/2019

**Introduction**

You can’t forget something you never went through. You can’t understand something you never experienced. You don’t recognize how valuable something is that you never had to earn. All of these statements summarize American’s stand on their democracy. It is human nature not to give adequate appreciation for a struggle that one did not participate and to neglect a thing that has simply been handed down through the generations.

The political climate in the country has shifted over the past few years. Many experts will pinpoint a different stat to that change. While many of them agree that their has always been some level of disfunction in Washington, they can all recognize that it has not been as bad as where it is now. Congress seems unable to accomplish anything, and the people seem less engaged and willing to stand for anything of substance.

During these troubling moments in American democracy, when the challenges of government begin to weigh heavy on a politician’s shoulders, many make reference about returning to the Federalist Papers for guidance. It is a sensible approach; the constitution may be the foundation for American democracy, but it is a one page document with many statements open for interpretation and no examples. And since none of the founding fathers bothered to stick around, it is one of the few documents that serves into a dive of the possible meaning behind the constitution.

**Analysis**

**About the Data**

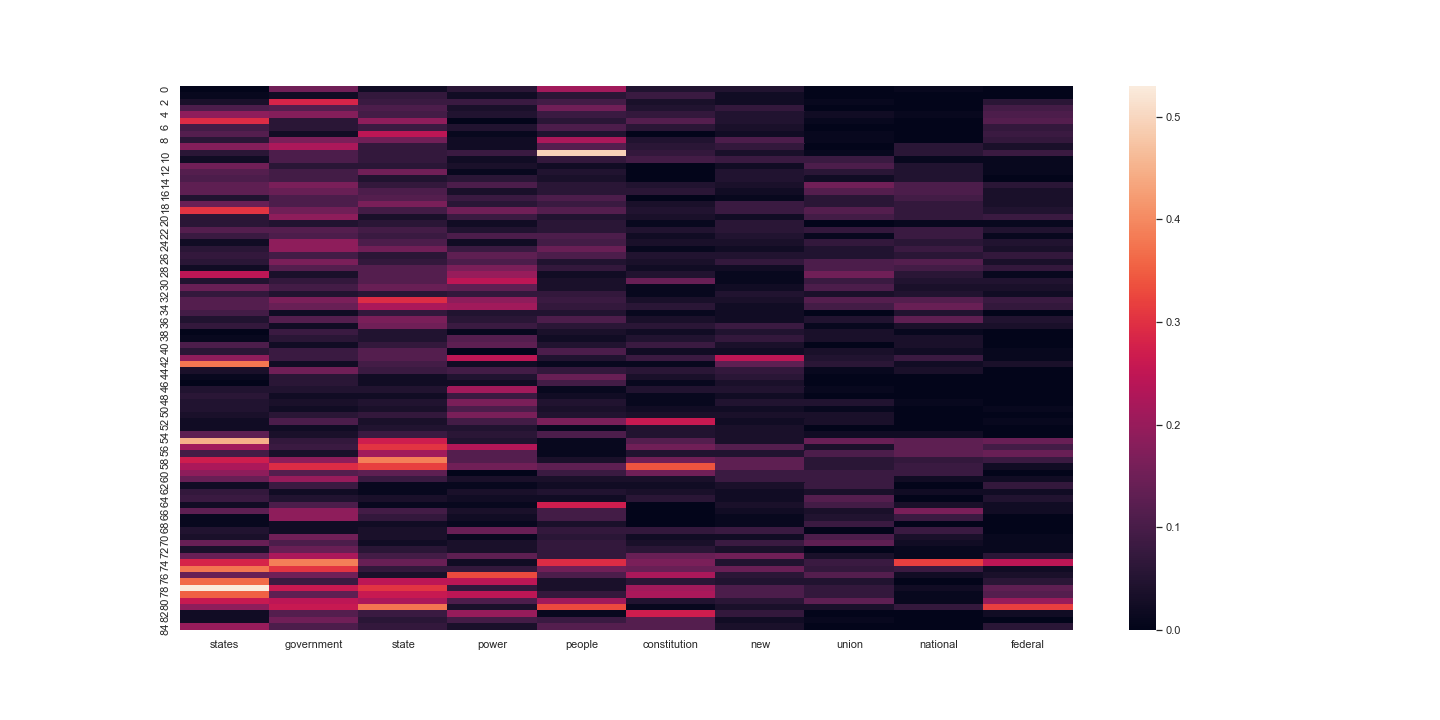
The data is a collection of The Federalist papers, where each of the 85 essays is a text file. Each essay is titled ‘Federalist No. <essay number>, contains the author (or believed author), the edition of publication, and the full text essay.

In preparation to vectorize the corpus, a list of file path and the file name were recorded. The file name was used in the generation of the vectorized dataframe to refer back to the essay.

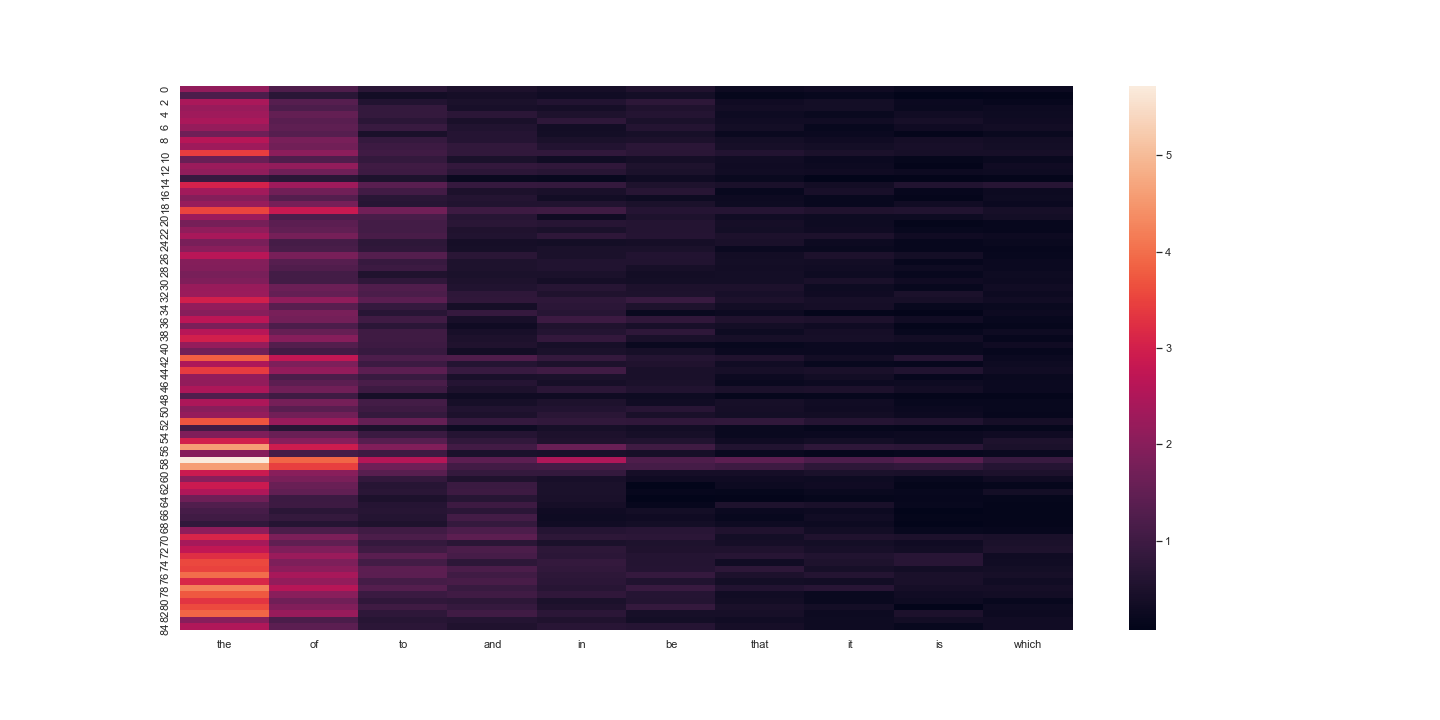
Two vectorized datasets were created. The first removed stop words as defined in the nltk.corpus package. The second left The Federalist Papers untouched. This was to compare the impact of word importance when these words were left in. It was also an exercise in curiosity, as corpus vectorization of The Federalist Papers had been done before in a previous analysis using R and a comparison of the process seemed appropriate.

The vectorized data sets contained 85 rows (each row an essay), but the data frame containing stop words had over 250 more columns, 8449 to 8720. An additional step of preprocessing was to establish a threshold for each word, the number of times it appears in the data set to trim them down. Using a threshold of 100 for the first data (no stop words) set brought the most common words to 92 and 200 brought the second (stop words) to 110. These two datasets represent the most frequent words that appear in The Federalist papers based on their specific filtering criteria.

An interesting outcome of this, states is the most used word in the first dataset, states, is the 27th most frequent word in the second dataset.



Ten most frequent words in each dataset 1.



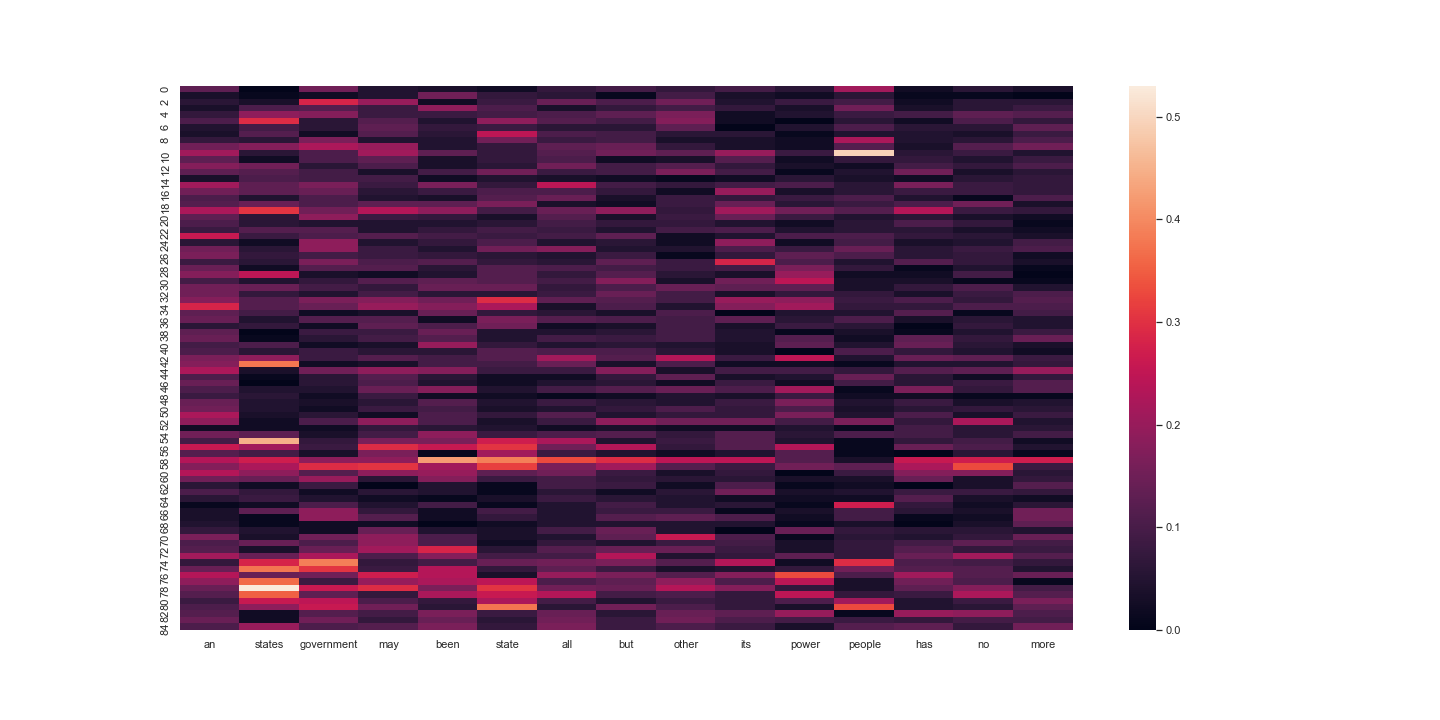
Ten most frequent words in each dataset 2.

**Models**

The CountVectorizer from sklearn was used to vectorize the corpus. Implementation was a small matter of passing the list of documents to a function. The package vectorized each essay, then all of those vectors were converted to a dataframe. The resulting data frames were 85 x 8449 (no stop words) and 85 x 8720 (with stop words).

**Results**

Comparing the differences between the two datasets, the impact of word choice becomes very apparent. In dataset 1 (no stop words) the most frequent word ‘states’ is the 27th most frequent in dataset 2 (stop words). Examining the top 10 words in dataset 1, and where they fall in dataset 2, shows that these important words begin to lose their weight. The 10th most frequent word ‘federal’ makes it in to the top 100 of dataset 2. In any analysis, these two sets would produce greatly differing results, highlighting the importance of word inclusion/exclusion.



Most frequent 25 – 40 words dataset 2 (includes top 5 words dataset 1).

Further examination of the data sets shows that 65% of the words in dataset 2 are stop words. The remaining 35% do appear in dataset 1.

**Conclusion**

The Federalist Papers are often referenced in legal arguments and by politicians on Capitol Hill. They are a mainstay in our democracy as a source for clarification on some of the more general and debatable points in the constitution. It is easy for a human to read anyone of those essays and derive the purpose for the writer. But if someone analyzing the work wanted to engage the text, what would be the best approach?

While text mining can be simplified to a few lines of code, the art of selection plays a heavy toll. What are the best words to include? There is simply no good answer, or at least one that has been provided. Remove the wrong stop word and the significant variable to prove authorship may be lost, filter out the short words and context/sentiment may be misclassified.

In the end, the results of an analysis derived from text may not depend so much on what machine learning algorithm is used to process the data, but what choices were made to prepare it.