Visualizing Text

MSPA PREDICT 455-DL-SEC55

Darryl Buswell

1 Introduction

This assignment explores text from Abraham Lincoln's four Addresses (Addresses), with the aim to identify and compare trends in each Address using visualization techniques. Full text for each speech is available from the GitHub repository, Web and Network Data Science (WNDS 2016). Table A1 shows a summary of properties for each of the four files used. Each file is in text format, and each represents an Address made by Abraham Lincoln over the years 1861, 1862, 1863 and 1864.

2 Data Pre-Processing

Data was pre-processed to remove punctuation, digits, stop words and excess white space between words. This transformation resulted in retention of words with meaning which can be aggregated and analyzed. The package 'tm' was leveraged to carry out transformations and to convert the text into a 'corpus'. A separate corpus was created for each of the original datasets and for an aggregated dataset which contains text from all four or the original datasets. Table A2 summarizes each of the applied transformations, as well as the order in which they were applied.

A Term Document Matrix (TDM) was created for each corpus, which consists of a matrix of word documents against the frequency by which they appear. A separate TDM was created for single N-gram's (e.g. 'United'), two joining N-grams (e.g. 'United States'), and three joining N-grams (e.g. 'United States America'). Tokenizing over N-gram sequences facilitated the exploration of not only Lincoln's use of single words, but also the use of two and three word combinations.

Table A3 shows the number of unique N-gram occurrences within each TDM. Note that the number of unique unigrams which occur over all four Addresses is much less than the sum of unique unigrams which occur over each. This suggests that many unigram occurrences were repeated over each Address.

3 Data Exploration

Frequently appearing N-grams over all four (combined) Addresses are represented as word clouds and bar plots in Figures B1 to B6. Figure B1 and B2 show that the most common unigrams over all four Addresses were 'the', 'congress', and 'states', which occured 43, 29, and 28 times, respectively. The most common bigram over all four Addresses was 'united states', followed by 'i recommend' and 'circuit courts'. The amount of occurrences of 'united states' comes in at 9, greatly outweighing 'i recommend' and 'circuit courts' which showed 5 and 4 occurrences, respectively. The most common trigrams were 'circuit judges provided' and 'confiscate property used'.

Figure B7 presents a comparison word cloud which shows the year of each Address in a 2x2 grid. The unigrams occured at a high frequency across all Addresses, but occurred most frequently in the year specified in the word cloud. Some results were suprising, for example, the Battle of Fort Sumter occurred over 1861, however the highest use of the unigram 'naval' occurred during the 1863 Address. Also, the American Civil War started in 1861, yet the unigram 'war' was mentioned most often in 1864.

We can relax the focus on bigrams and trigrams within each Address, and instead draw attention to correlations between each unigram. For this, we can leverage the 'biocLite' package. A visual representation of unigram correlations over all four (combined) Addresses is shown in Figure B8. A correlation threshold of 0.2 was applied to the unigram correlations and a limit was set for the number of unigrams shown. Not surprisingly, unigrams previously reported to have high frequencies are shown to have high correlations.

Finally, we can use the R package 'topicmodels' to implement a Latent Dirichlet Allocation (LDA) model. The LDA can be used to classify themes within each Address. For the most part, default parameter options were used for the LDA implementation, however both the number of topics and the number of unigrams within each topic was set at four and ten respectively. The list of topics and included unigrams is shown in Table A4. It is difficult to define any of the topics based on their included unigrams, and results could not

be improved by modifying either the topic count or number of included unigrams. However, applying this routine to a larger text corpus may improve the results.

A bar plot was then used to represent how often the unigrams within each topic were referred to for each Address. Note that the proportion of use of each unigram (proportion of topic reference) was calculated by ignoring the frequency of any unigrams which are not within one of the four topics. As Figure B9 shows, the proportion of topic reference does not change radically between each Address. Do note that there was a slight increase in occurrences of unigrams within Topic 3 between the 1861 and 1862 Address, which coincided with a decrease in occurrences of unigrams within Topic 2. There was also a slight increase in occurrences of unigrams within Topic 1 between 1862 and 1863, which coincided with a decrease in occurrences of unigrams within Topic 4.

4 Conclusion

We were able to process the dataset in order to find common unigram, bigram, and trigram frequencies, find correlations between unigrams and classify the text into topics. The powerful visualization packages within R also allowed these text transformations to be presented using a combination of wordclouds, comparison clouds, barplots, stacked barplots and correlation diagrams. For future work, it may be worth applying the same process to a larger corpus of text, potentially capturing all Addresses made over a number of administrations. Doing so may expose new trends within the underlying data and allow a wider set of inferences to be made.

Appendix A Table Output

Table A1 Dataset Summary

File (name)	Size (mb)	Lines (no.)	Longest Line (chars)
R_W_Lincoln_1861.txt	0.04	161	1887
R_W_Lincoln_1862.txt	0.05	215	3149
R_W_Lincoln_1863.txt	0.04	113	2810
R_W_Lincoln_1864.txt	0.03	167	2376

Table A2 Pre-processing Routine

Order	Transformation
1	Convert from UTF-8 to ASCII
2	Remove punctuation
3	Remove digits
4	Remove English stopwords (e.g. to, a, is)
5	Convert text to lowercase
6	Strip all whitespace

Table A3 N-gram Audit

Union Address	Unigram	Bigram	Trigram
Lincoln 1861	1651	3308	3369
Lincoln 1862	1782	3987	4103
Lincoln 1863	1511	2966	3047
Lincoln 1864	1480	2823	2889
Lincoln 1861-1864	3686	12372	13238

Table A4 Topic Terms

Topic 1	Topic 2	Topic 3	Topic 4
the	the	the	the
will	states	states	will
congress	year	will	states
can	united	labor	can
government	department	people	national
upon	receipts	time	war
country	upon	union	now
general	great	congress	may
states	treasury	shall	new
may	last	upon	great

Appendix B Figure Output

Figure B1 Lincoln SoUA 1861-1864 Unigram Wordcloud

Figure B2 Lincoln SoUA 1861-1864 Unigram (Frequency >=60)

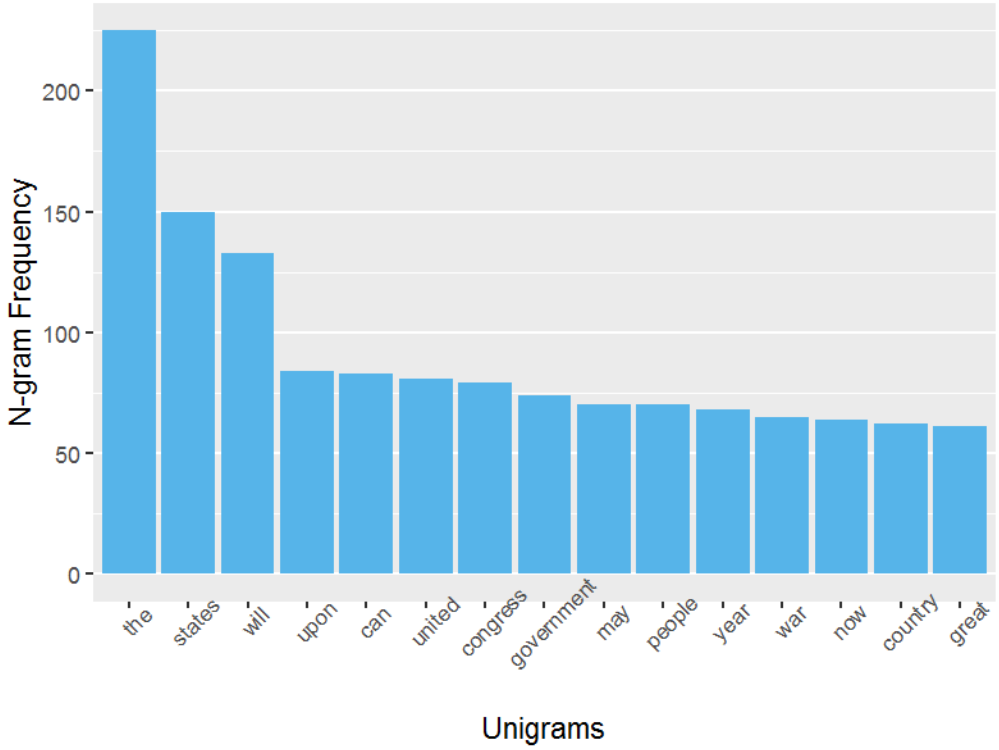


Figure B3 Lincoln SoUA 1861-1864 Bigram Wordcloud

Figure B4 Lincoln SoUA 1861-1864 Bigram (Frequency >=9)

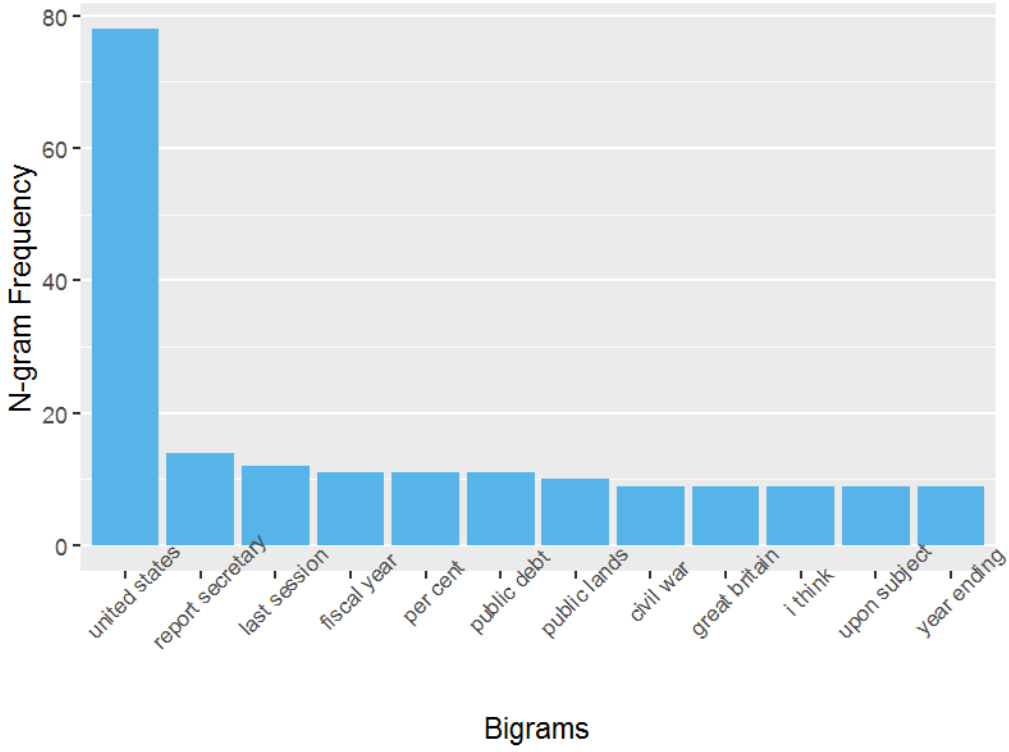


Figure B5 Lincoln SoUA 1861-1864 Trigram Wordcloud

Figure B6 Lincoln SoUA 1861-1864 Trigram (Frequency >=4)

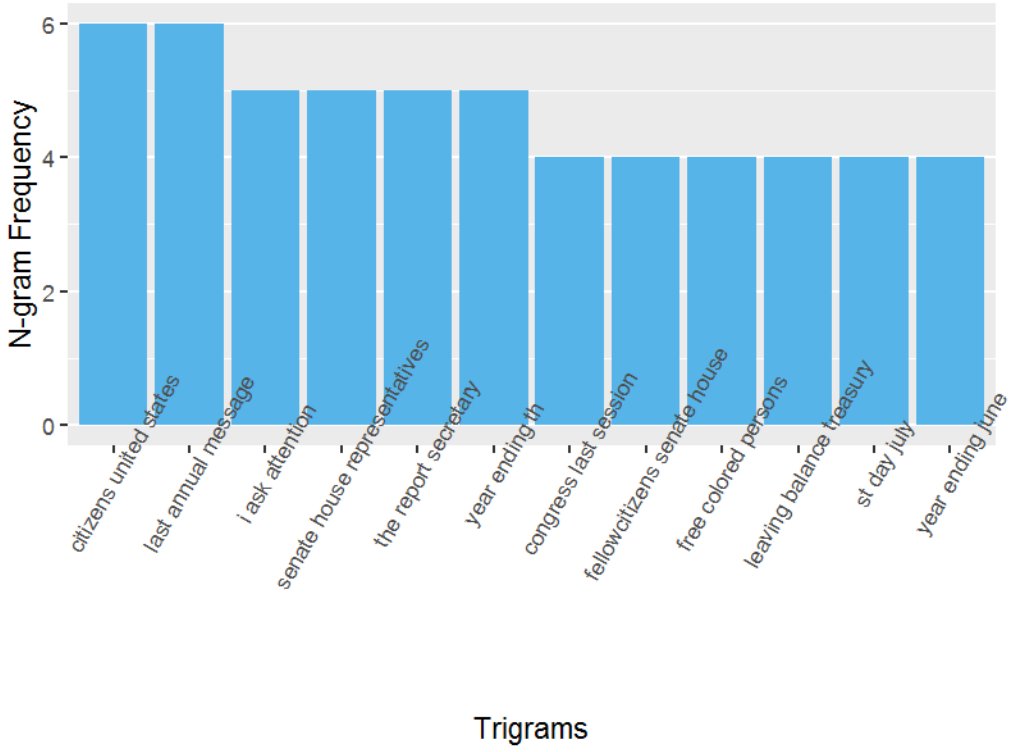


Figure B7 Lincoln SoUA 1861-1864 Unigram Comparison Cloud

Figure B8 Lincoln SoUA 1861-1864 Correlation Cloud

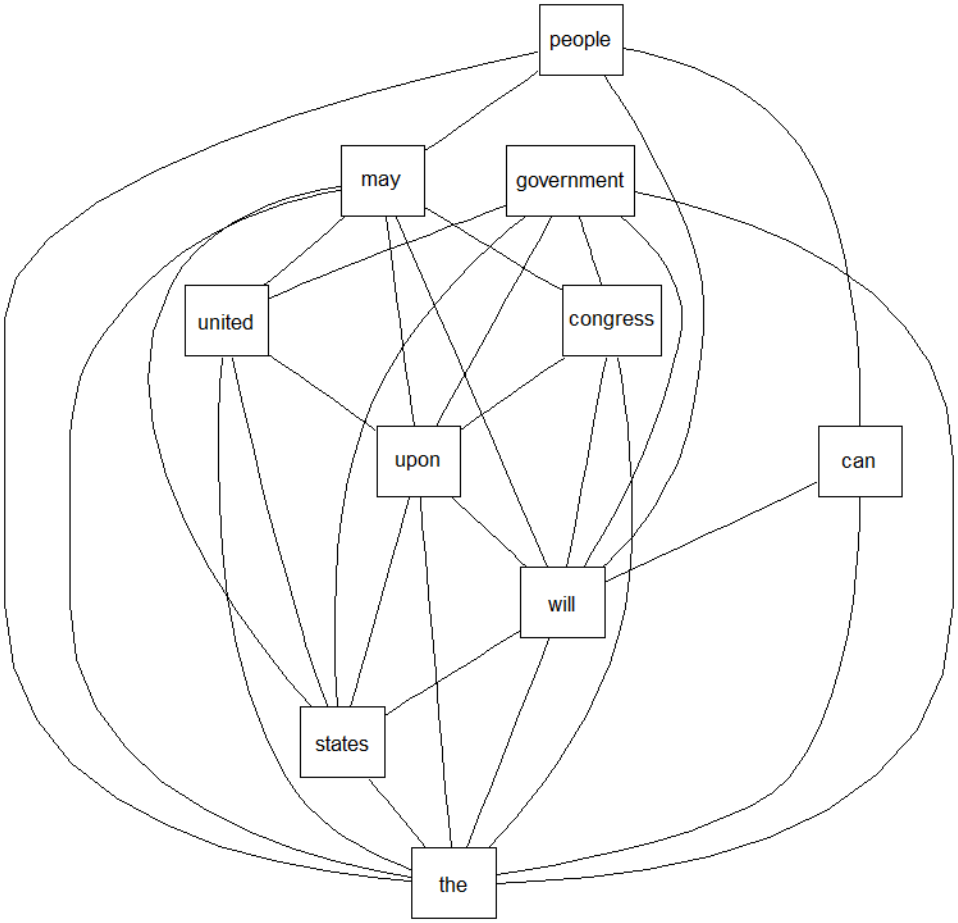
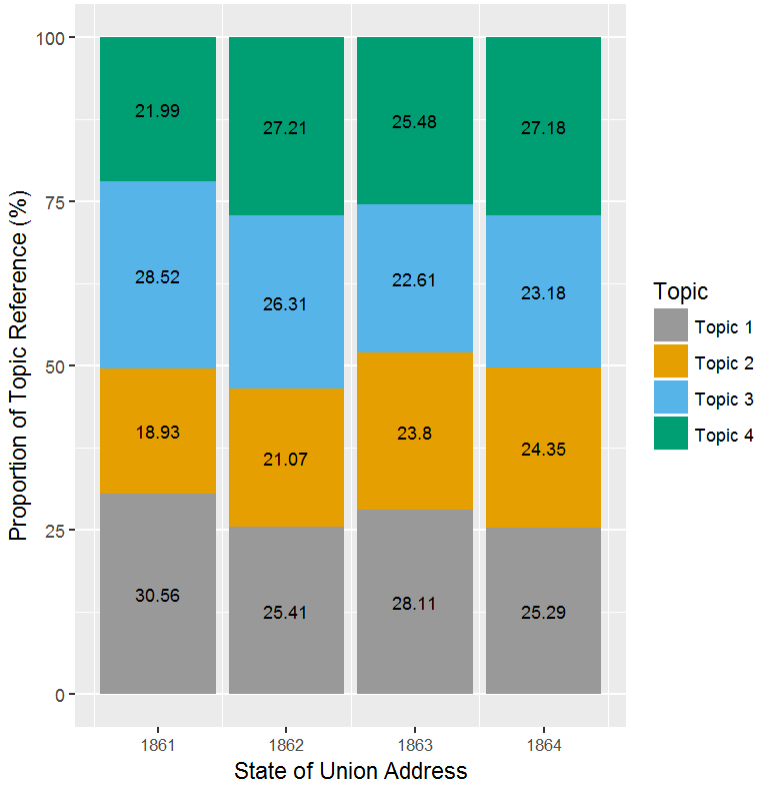


Figure B9 Lincoln SoUA 1861-1864 Topic References



References

WNDS. 2016. “GitHub: Web and Network Data Science.” https://github.com/mtpa/wnds.