Nick Clifford
29 April, 2020

Exploratory Text Analysis: The Valley of the Shadow

Conclusions:

Principal Component Analysis

Hoping to discern some kind of latent properties of the documents, I first turned to principal component analysis (PCA). After first being introduced to this large corpus of documents, I hoped to use PCA to more efficiently navigate the context of information within each letter. Specifically, the similarities in the context of the more prominent authors, what was being written throughout the war, and what does this also say about the two counties together.

What I found was a pretty decent separation between author clusters from the first and fourth principal components (Figure (1)). Using the loadings of these components, I inspected the top terms associated with each and found them both to be a confusing mixture of contexts. Both make mention of words related to education (teacher, school, scholars, etc.) combined with words found in a sentimental letter (respectfully, love, come, home). The other components also throw in war-related words (wounded, killed, enemy, artillery, prisoners). I also found a less clear separation between county documents during the war using components with loadings similar to author contexts (Figure (2)). In summary however, I believe these components help shed some light on how these historical figures wrote letters during the civil war. While some were quick and to the point, others would have put an extra effort in the endearment aspect when writing to their loved ones.

Figure 1

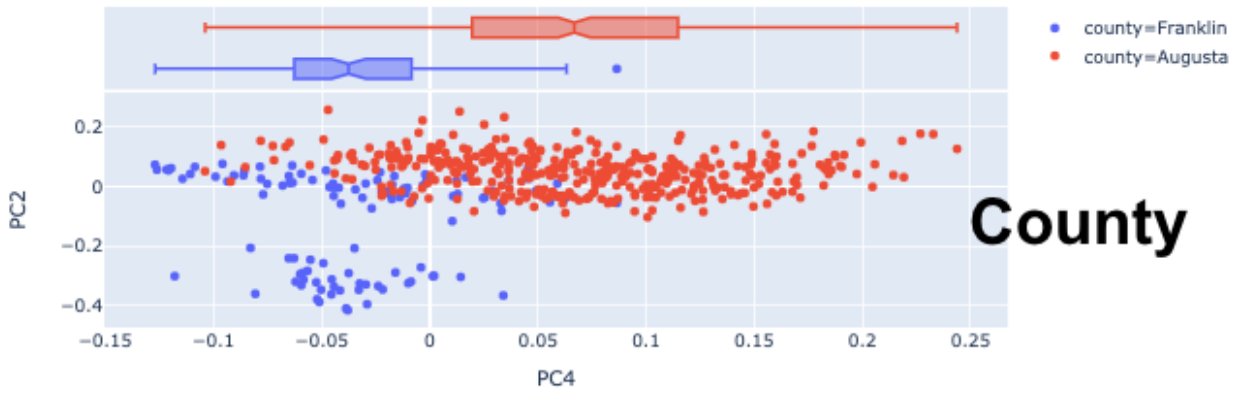

Figure 2

Word Embeddings

I hoped to use word embeddings to help structure the context of this corpus through a T-SNE graph, but also to glean some cultural insight through the use of semantic algebra analogies.

Using word2vec, I wanted to compare the differences in the T-SNE graphs when either bagging by paragraph or sentence. I found that the sentence bagging left more random pronouns scattered around the semantic space, which is most likely since some sentences may simply be the beginning salutations or farewell of a letter that includes a person's name. Instead of removing all pronouns from the Token table before bagging, I felt as though important pronouns such as geographic locations and army generals would be necessary for exploring the semantic space. As such, I decided to bag by paragraph, which was able to drop out more random names. Figure (3) shows the resulting T-SNE graph and the interesting clusters that I found within. I found it interesting how two clusters related to letter writing was split into: either greetings/terms of endearment (love, respectfully, miss, pleasure) or synonyms for lettered mail and 'requesting' types of words (send, please, money, pay). Additionally, one of the largest groups is related to war-like words (as expected from a civil war domain corpus). But more interesting is how war words (battle, army, cavalry) is heavily used in conjunction with geographic locations (river, field, Richmond, Staunton). Perhaps this is since most authors are writing to loved ones and friends, and in describing scenes of war must pair the action to locations to give it additional meaning. I find this similar to how many gruesome battles in the Civil War have specific locations that are famous for the bloodshed there (cornfield, sunken road, Dunker's church).

Lastly, I relied on the war/location cluster to come up with analogies that would help explore the semantic space concerning the war. I found the most interesting analogies to be as follows: (South : Richmond :: North : **Washington)**, (Horse : Cavalry :: Men : **Regiment)**, (Cold : Sick :: Battle : **Wounded),** and (Husband : Wife :: Men : **Horses)**

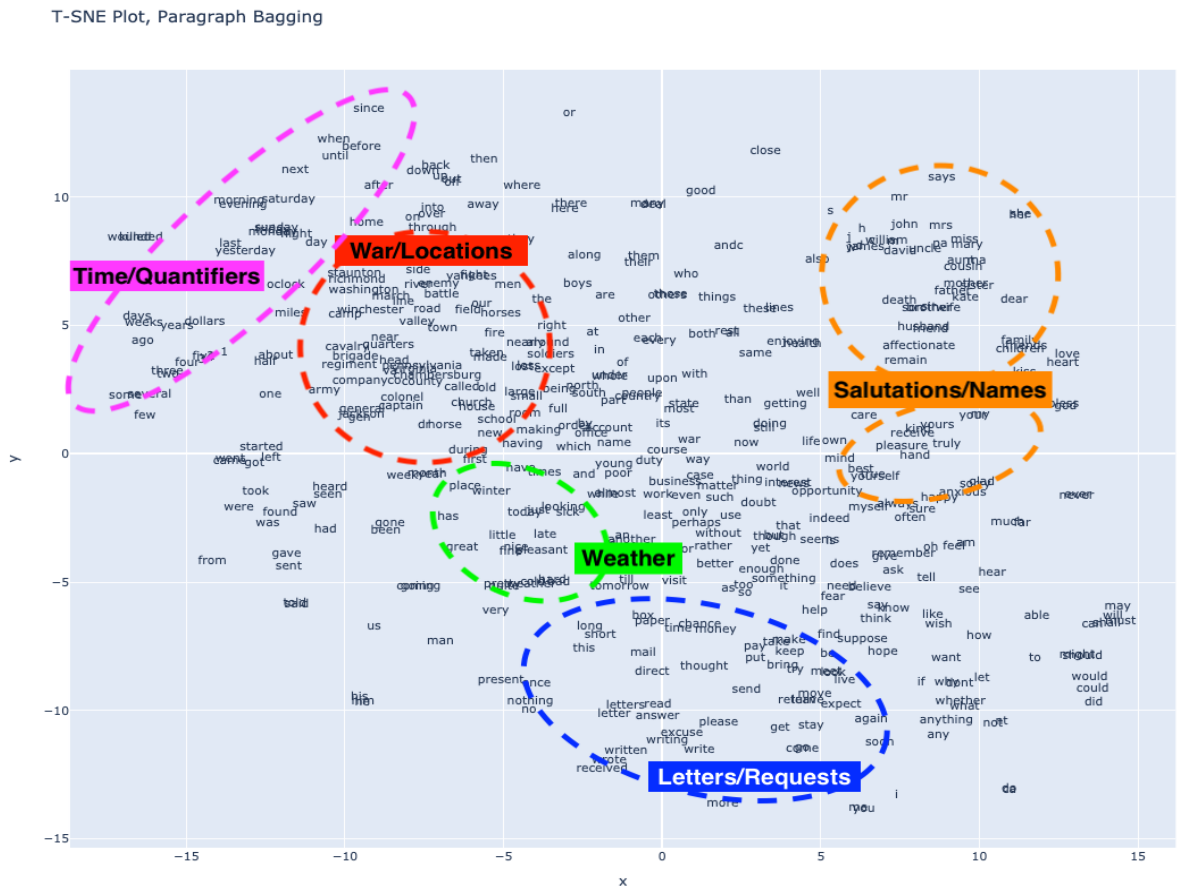

Figure 3

Sentiment Analysis

I was quite interested in applying sentiment analysis to this corpus to answer two questions: (1) What could it teach us about the effects of the war on during that time? And (2) Could you use sentiment/polarity to map and outline events throughout the war, similar to mapping the plot of a story? For all of my sentiment analysis, I relied on the NRC lexicon-based approach.

First, I filtered out documents that were not written during, before, or after the civil war. Then I compared the average emotion for each year between 1858-1870 by bagging at the pages level, multiplying the TFIDF score to each sentiment value and averaging scores across documents of similar years. Looking at the Figure (4), it looks like the overall polarity drops once the beginning of the war begins in 1860, then drops to its lowest value in ’62 and rises again during the last year of the war and after. To produce a better visualization, I took the difference in each emotion score into the following years and plotted the differences as a heatmap (Figure 5). Unexpectedly, during the war, it looks like most emotion values drop only slightly, but the most variation comes during before and after the war. This may just be a side effect of more of

the corpus originating during war years, thus bringing down the averages. However, I did find it interesting to see how anticipation and fear increase before the war breaks out.

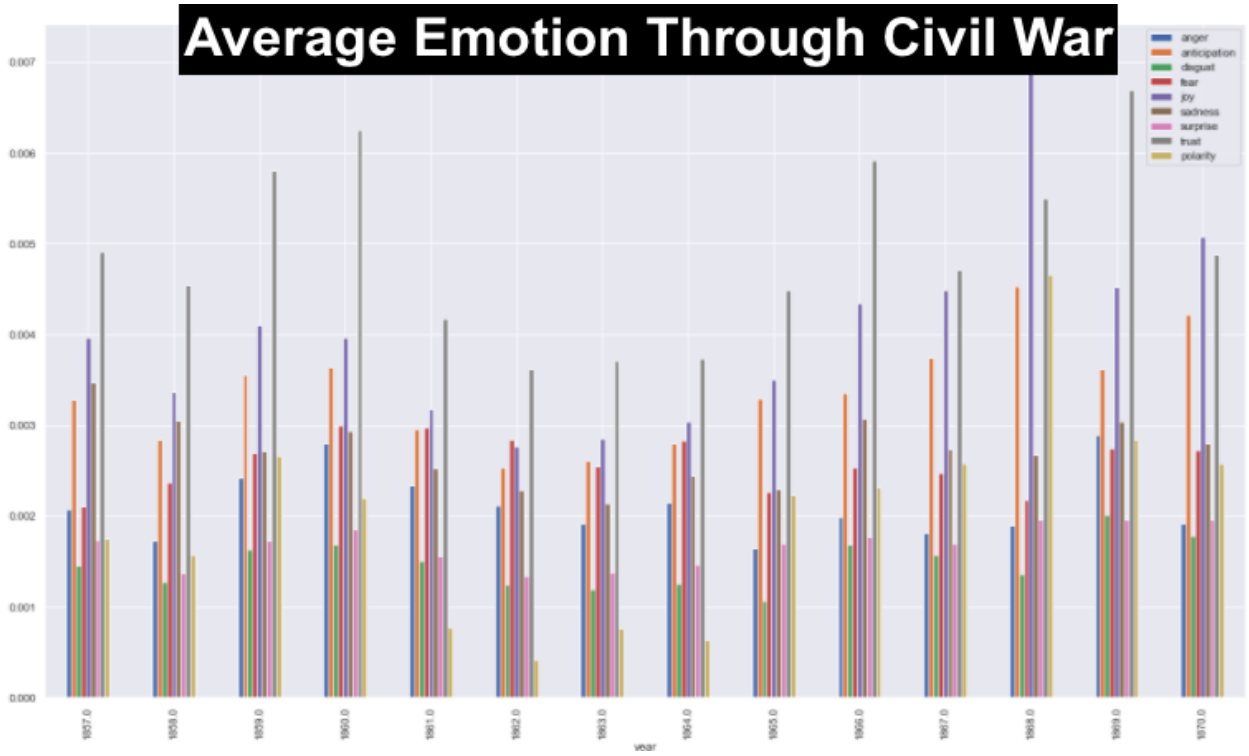

Figure 4

year	anger	anticipation	disgust	fear	joy	sadness	surprise	trust	polarity
1858.0	-0.000340	-0.000444	-0.000180	0.000270	-0.000595	-0.000416	-0.000362	-0.000366	-0.000176
1859.0	0.000688	0.000715	0.000359	0.000321	0.000735	-0.000331	0.000355	0.001262	0.001086
1860.0	0.000382	0.000090	0.000055	0.000307	-0.000144	0.000220	0.000124	0.000449	-0.000455
1861.0	-0.000465	-0.000686	-0.000179	-0.000024	-0.000782	-0.000409	-0.000295	-0.002086	-0.001433
1862.0	-0.000218	-0.000426	-0.000268	-0.000137	-0.000406	-0.000249	-0.000218	-0.000551	-0.000352
1863.0	-0.000199	0.000077	-0.000053	-0.000295	0.000085	-0.000143	0.000040	0.000091	0.000347
1864.0	0.000234	0.000191	0.000066	0.000286	0.000188	0.000298	0.000086	0.000022	-0.000128
1865.0	-0.000507	0.000490	-0.000193	-0.000571	0.000460	-0.000137	0.000236	0.000756	0.001597
1866.0	0.000342	0.000065	0.000622	0.000280	0.000840	0.000776	0.000067	0.001431	0.000085
1867.0	-0.000175	0.000385	-0.000117	-0.000071	0.000143	-0.000338	-0.000068	-0.001210	0.000264
1868.0	0.000085	0.000789	-0.000206	-0.000290	0.002583	-0.000064	0.000264	0.000791	0.002077
1869.0	0.000992	-0.000907	0.000653	0.000569	-0.002551	0.000369	-0.000008	0.001193	-0.001816
1870.0	-0.000975	0.000591	-0.000230	-0.000020	0.000556	-0.000245	0.000006	-0.001815	-0.000266

Figure 5

To further study emotions in the time of war, I essentially separated these plots according to the two counties. Looking at the two bar plots reinforces my belief that the magnitude of the scores is highly dependent on the number of observations (documents) within the year. Even still, perhaps the rank order of the emotion types over time is informative. The heatmaps show interesting differences between the counties. Specifically, the southern count, Augusta, increases in anticipation and anger during the war, while the northern county, Franklin, increases in disgust before the war and increases in sadness during the war.

county	year	anger	anticipation	disgust	fear	joy	sadness	surprise	trust	polarity
Augusta	1858.0	-0.001075	0.000138	-0.000443	0.000315	-0.000469	-0.000252	-0.000196	-0.000656	0.000067
	1859.0	0.000879	0.000825	0.000419	-0.000054	0.001319	-0.000672	0.000273	0.002093	0.002156
	1860.0	0.000282	0.000196	-0.000034	0.000453	-0.000497	0.000405	0.000140	-0.000402	-0.001498
	1861.0	-0.000539	-0.000818	-0.000151	-0.000200	-0.000606	-0.000414	-0.000268	-0.001598	-0.000678
	1862.0	-0.000121	-0.000367	-0.000094	0.000019	-0.000352	-0.000149	-0.000229	-0.000425	-0.000324
	1863.0	-0.000239	0.000002	-0.000081	-0.000360	0.000150	-0.000218	0.000039	0.000034	0.000489
	1864.0	0.000480	0.000270	0.000170	0.000620	0.000287	0.000598	0.000141	0.000369	-0.000344
	1865.0	-0.000871	0.000187	-0.000242	-0.000971	0.000047	-0.000514	-0.000141	0.000607	0.001956
	1866.0	0.000546	0.000314	0.000566	0.000429	0.001094	0.000929	0.000416	0.001532	-0.000029
	1867.0	-0.000177	0.000395	-0.000107	-0.000063	0.000158	-0.000346	-0.000069	-0.001218	0.000275
	1868.0	0.000095	0.000739	-0.000218	-0.000286	0.002534	-0.000035	0.000282	0.000787	0.002005
	1869.0	0.000986	-0.000861	0.000638	0.000560	-0.002410	0.000377	0.000003	0.001139	-0.001661
	1870.0	-0.000977	0.000584	-0.000212	-0.000023	0.000449	-0.000274	-0.000022	-0.001749	-0.000360

Figure 6: Yearly difference in average sentiment for Augusta County

county	year	anger	anticipation	disgust	fear	joy	sadness	surprise	trust	polarity
Franklin	1858.0	0.000314	-0.000824	0.000108	0.000095	-0.000403	-0.000577	-0.000501	0.000368	0.000248
	1859.0	0.000426	0.000564	0.000275	0.000829	-0.000062	0.000128	0.000465	0.000123	-0.000377
	1860.0	0.000492	-0.000012	0.000115	0.000112	0.000251	-0.000061	0.000088	0.001303	0.000620
	1861.0	0.000201	-0.000514	0.000393	0.000879	-0.000717	0.000018	-0.000255	-0.001044	-0.001352
	1862.0	-0.000848	-0.000588	-0.001080	-0.000996	-0.000697	-0.000770	-0.000239	-0.001891	-0.000993
	1863.0	-0.000125	0.000287	0.000032	-0.000154	0.000028	0.000029	0.000054	0.000119	0.000023
	1864.0	-0.000395	-0.000050	-0.000256	-0.000592	-0.000179	-0.000495	-0.000056	-0.000680	0.000494
	1865.0	0.000477	0.001290	-0.000035	0.000515	0.001539	0.000883	0.001219	0.001133	0.000598
	1867.0	-0.000091	-0.002005	-0.000913	-0.001526	-0.002229	0.001503	-0.000720	0.001294	-0.001212
	1868.0	-0.000626	0.004625	0.001518	0.000412	0.006866	-0.002619	-0.000813	0.000059	0.007059
	1869.0	0.001357	-0.003136	0.001025	0.001119	-0.007423	0.000608	0.000061	0.002505	-0.007685

Figure 7: Yearly difference in average sentiment for Franklin County

Next, to outline events throughout the war, I decided to focus on the individuals who were more prevalent in their writings to produce sentiment plots of better time resolution. Looking first at Jedediah Hotchkiss of Augusta, who has 187 letters to his name, I initially grouped the sentiments by exact dates. However, upon inspection of the sentiment time-series, large peaks would arise for letters that were very brief in length. However, grouping by month returns very interesting results. One can see a dip in polarity and increase in fear during and after May 1863; the same month of General "Stonewall" Jackson's death. This is interesting since it was his brigade that Hotchkiss was a topographical engineer, and whom he traveled with since 1861. It seems like the beginning months where the polarity drops also correspond to early confederate victories like Jackson's famous win at Bull Run, which Hotchkiss describes in his letters.

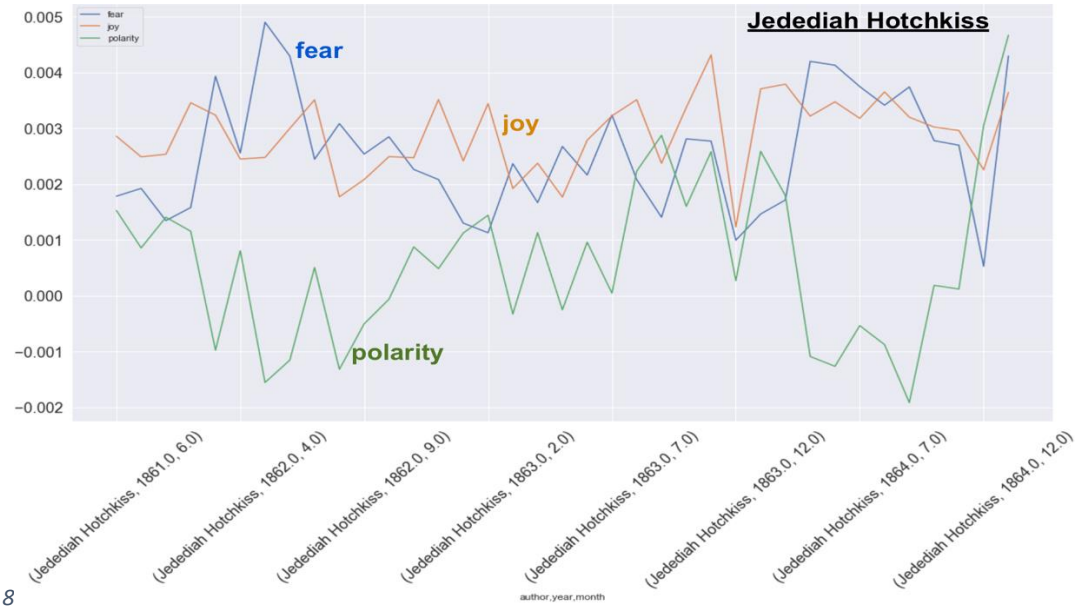

Figure 8

To capture the experience of somebody from the North, I looked at Alexander McClure, a Pennsylvanian politician who has 53 letters to his name. Grouping again by month, I attempted to learn some of McClure's backstory to help explain certain polarity trends. For example, the large increase in sadness likely a result of McClure bemoaning the Union's loss at Manassas in July 1861, while the increase in trust is a result of McClure expressing the merits of soldiers that he meets in July 1862. Finally, his trust rises again towards the end of the war, as he expresses his great confidence in winning an upcoming election for the Pennsylvania state legislature in 1864. I believe tracking the polarity of events in this manner helped speed up my understanding and focus when searching for documents on these figures. In the future, I would consider implementing a heuristics sentiment approach, such as VADER, to get more accurate results.

Figure 9