Emily Dickinson Poetry – DS 5001 Final Report

For this final project, I focused my analysis on the poetry of Emily Dickinson taken from Project Gutenburg. The text file that I analyzed were three different series (i.e. collections) of over 400 poems consolidated into one file. Each of the three series were divided into the following four categories: Life, Love, Nature, and Time & Eternity. The goal of this analysis was to determine if the exploratory text analysis (or ETA) could identify these themes and identify the type of sentiment in Dickinson' poem reflected in the four categories. In addition, I wanted to get a sense of the type of author Dickinson was and the types of stories she liked to write about.

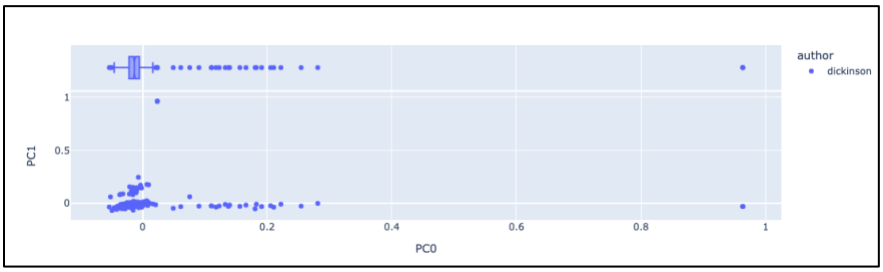
After preprocessing the data, I wanted to determine what words were identified as being the most significant in the corpus. Since we are looking for themes related to life, love, nature, and time & eternity, my goal was to find words that reflected these categories within the poems themselves. Using TFIDF, I was able to determine summary statistics like sum, mean, max, etc. for each term in the VOCAB table. Using a heatmap on the VOCAB table, I was able to determine what words were most significant within the poetry. Using the heatmap and sorting on the "tfidf_max" column, I ascertained that "love" and "nature" were some of the most significant words in the collection of poems. These two terms are perfectly reasonable given the fact that they were also placed as two of the four categories in the original data. We would expect to see those terms as significant because they accurately reflect the themes that we expect Dickinson to write about. Below, I have only showed these top two terms because they have the strongest significance in the table (the values for "tfidf_max" drop significantly after these two terms).

[992]:

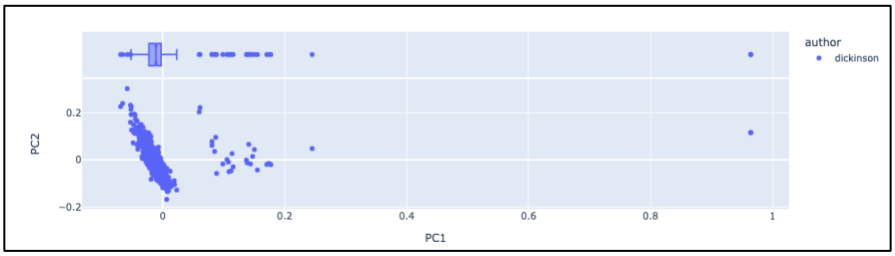
term_id	term_str	n	num	stopwords	port_stem	pos_max	doc_freq	inverse_doc_freq	tfidf_mean	tfidf_sum	tfidf_median	tfidf_max
2672	love	30	0	0	love	NN	24	1.261263	0.186658	4.479785	0.031532	1.261263
2914	nature	30	0	0	natur	NN	28	1.194316	0.150187	4.205230	0.023934	1.194316

Next, I used the dimensionality reduction technique known as Principal Components Analysis (PCA) to ascertain what similarities and differences existed between the poems. Since we typically see most of the variance captured in the first couple PCA, our analysis will only look at comparisons between PC0 & PC1, PC1 & PC2, and PC2 & PC3. For our first plot, between PC0 and PC1, we see that that the data points are highly clustered in the bottom left corner of the plot (bottom of PC0, left end of PC1). The data points seem to mainly lie across PC0 with some spread as we move to the right of PC0. We can also identify two outliers on the plot near the top of PC0 (Poem #364) and near end of PC1 (Poem #341). It is unclear to us what exactly the principal components are detailing about the collection of Dickinson's poems – a deeper analysis would be required to understand this. One thing, though, that we can ascertain from this initial plot is that most Dickinson's poems seem to have some sort of similarity between them. We see that most of these data points are clustered around each other (except for the small spread seen to the right of the main cluster and the two outliers). Thinking about the four categories referenced in the introduction (Life, Love, Nature, Time & Eternity), one might

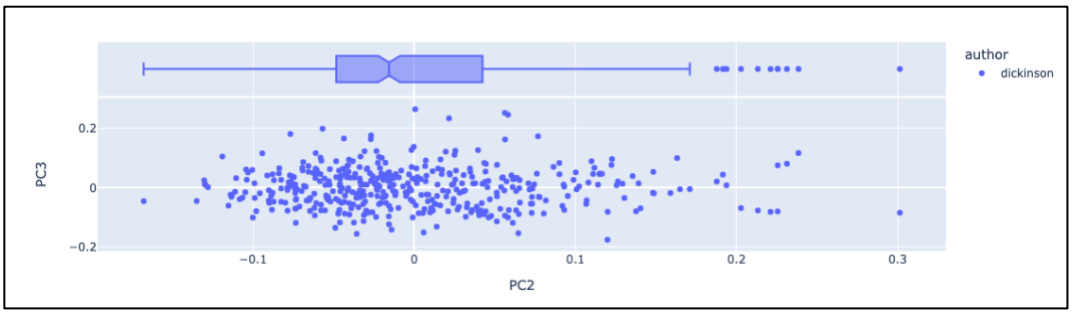
expect that there would be overlapping topics and terms that stretch across the four categories. Also, since we are only looking at poetry by Dickinson, it would make sense to us that we wouldn't expect to see a great spread in the plots (unless Dickinson was known for writing about a wide variety of topics). Being unfamiliar with Dickinson's poetry, I am curious to better understand our two outliers in the plot below. If further analysis was to be performed, I would want to get greater context on the overall poetry of Dickinson and compare it to these two outliers.



Looking at the plot below for PC1 and PC2, we see a different trend in the plot where the data points seem to lie across PC2 rather than PC1 (this makes sense since in the first plot the data points seemed to be relatively in the same position for PC1). This highlights that PC1 may better be able to tell us when a poem is not by Dickinson, if comparing her work to other poets. We also see, when looking at the PC2 axis, that there seems to be a main cluster with some slight spread as we move along the PC1 axis. In this plot, also, we seem to have only one outlier that belongs to Poem #364. Again, this might be a poem that is drastically different from typical Dickinson poetry.



Finally, looking at the PC2 and PC3 plot, we observe a different pattern in the plot not seen in the other previous plots. In this plot, we see that the data points are spread out across PC2. Again, we cannot attribute what exactly is causing this phenomenon, but one assumption is that the explained variance between P2 and PC3 is drastically different. In the plot, we see little spread on PC3 but a widespread throughout PC2. This seems to confirm the general assumption that most of the information gained can be found in the first couple principle components (i.e. maximum variance achieved).



Overall, from the PCA plots, we identify that Dickinson's poetry, mostly, seem to be similar in relation to one another except for some poems that differ from the norm. An improvement in this analysis (or a next step) would be to gain more knowledge about the poems to identify why exactly these PCA plots looked the way they did.

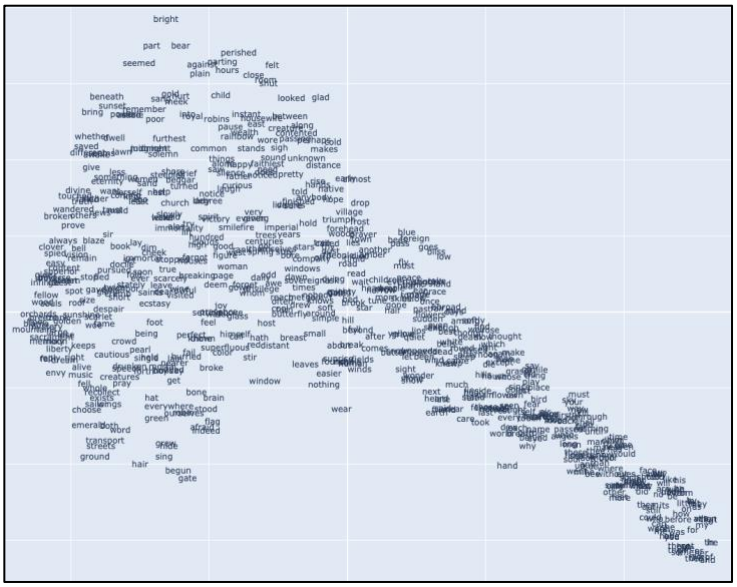
Next, looking at Topic Models, I was able to design my code directly with the prior knowledge that there were four main topics categorized within the poetry (life, love, nature, and time & eternity). We can this knowledge in our development of Topic Models (through LDA) to identify how the poems correlate to these four categories. For Topic Models, we develop two separate tables to gather further insight into the text. First, we develop a THETA table that provides insight into which poems correlate best amongst the four topics. To clearly understand which topic correlated highest with a certain poem, a heatmap was shown to depict this relationship. Looking at the THETA table and the first set of poems, we can see that the highest correlated topic is not consistent between each poem. This goes against what I would expect to see because, in the original data set, each series is organized by the four categories in order (i.e. poems about "Life" are sectioned off together, etc.). One potential reason by this unintuitive result (seen in the heatmap table below) is that the generality of the four categories may cause our analysis to see them as similar and that some of these poems could be sorted under another more than one category. We can look at the THETA table as a potential way to assign these poems to other categories than what the original organizer had intended.

	topic_id	0	1	2	3
series_num	poem_num				
	1	0.022912	0.025169	0.927944	0.023975
	2	0.944634	0.018268	0.018491	0.018607
	3	0.025463	0.025776	0.025316	0.923444
	4	0.012747	0.013113	0.012751	0.961389
	5	0.010151	0.010204	0.969486	0.010158
	6	0.032116	0.031551	0.033120	0.903213
	7	0.032602	0.032051	0.032123	0.903224
	8	0.012946	0.012606	0.012632	0.961815
	9	0.880613	0.038277	0.037513	0.043597
	10	0.008469	0.008549	0.008519	0.974463
	11	0.924236	0.025237	0.025360	0.025167
	12	0.032735	0.032185	0.901761	0.033320
	13	0.018744	0.018251	0.018384	0.944621
	14	0.025617	0.026343	0.025369	0.922670
	15	0.981249	0.006201	0.006285	0.006265
	16	0.954797	0.015170	0.014971	0.015063

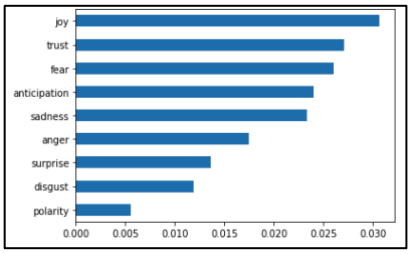
Using the PHI table, we can determine the top terms that are captured for each topic. We get some interesting results from this table – many words appear in multiple topics. This seems to fall in line with the previous assumption made based off the THETA table. Again, we might attribute this to the fact that the four categories in question have similarities to one another where similar terms are used in more than one topic. We see words like “day”, “life”, “heart”, “sun”, etc. were part of the top 10 terms for more than one topic. These results gained from both the THETA and PHI table also help support the fact that Dickinson had a specific focus in the themes of her poetry.

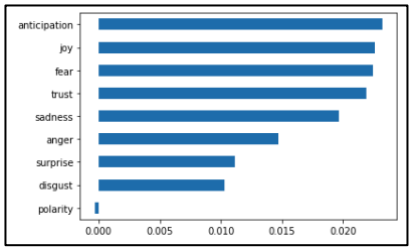
term_str	0	1	2	3	4	5	6	7	8	9
topic_id										
0	morning	life	day	thee	time	feet	face	eyes	heart	star
1	day	soul	summer	sun	life	man	grave	prayer	time	air
2	air	friend	bird	way	sea	eyes	face	heart	noon	door
3	sun	night	life	time	summer	day	sky	sea	bee	fingers

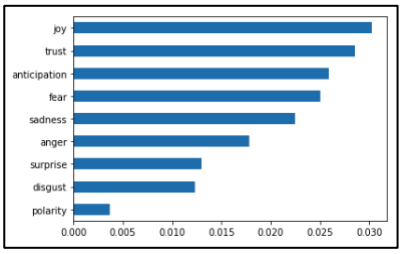
Next, we look at Word2Vec and develop a t-SNE plot to determine what words cluster closely together and see if we can derive any interesting insights into the text through this plot. My original expectation for this analysis was that the clustering would clearly define the four categories of life, love, nature, and time & eternality. After trying some variations of the t-SNE plot, it became obvious that it would be difficult to decipher similarity between terms. The t-SNE plot does not seem to be able to decipher any meaningful clusters from the text. Looking at specific sections of the plot, I found that there were no distinct clusters that provided meaning to our analysis question and nothing very close to showcasing the four categories of interest. However, the fact that we are not able to determine any clear clustering within the plot might be an indicator of the earlier observation that the categories of interest are similar, and that Dickinson’s poetry may overlap between these categories of life, love, nature, time & eternity.



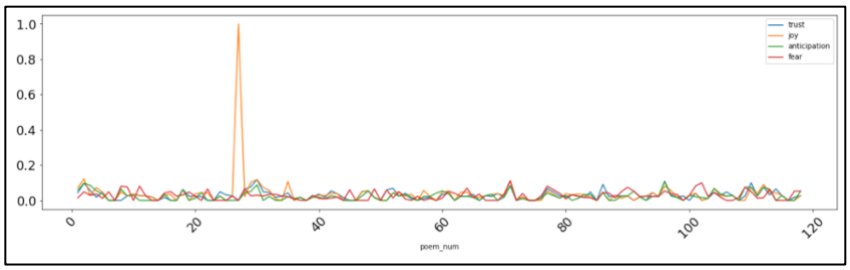
For my sensitivity analysis, I used the Salex lexicon from our previous Code Walkthrough to determine the polarity for each of the terms in the vocabulary. From the Salex lexicon, the emotions examined were trust, joy, anticipation, fear, disgust, anger, surprise, sadness, and polarity (derived from positive minus negative emotions). Each term in the vocabular was assigned either a 0 or 1 for each emotion and we saw that many terms ended up having a complete sparse vector. Looking at each series in the collection of poems, we can deduce that Dickinson's poetry seems to incorporate themes that were mainly positive. From the plots, we see that the emotions of joy, anticipation, trust, and fear are reflected in all three series of poetry. I think that this makes a lot of sense when reflecting back to the categories that Dickinson's poetry was organized in (life, love, nature, time & eternity). At first glance, these topics most likely give readers a positive connotation but, when digging deeper, we know that discussion of these topics are not always good (hence, the emotion of fear being prevalent in her poetry). On top of that, seeing that the top 4 emotions are consistent in each series within the full collection, I think that this further cements Dickinson as an author who has a particular niche with her poetry (at least within the context of the text we are analyzing). For the bar plots below, we can see the top emotions for Series 1 to 3 (from left to right) and how these bar plots are similar (in order and magnitude of impact within the poetry).



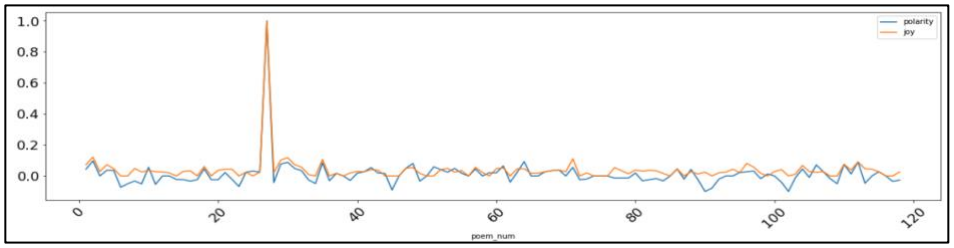


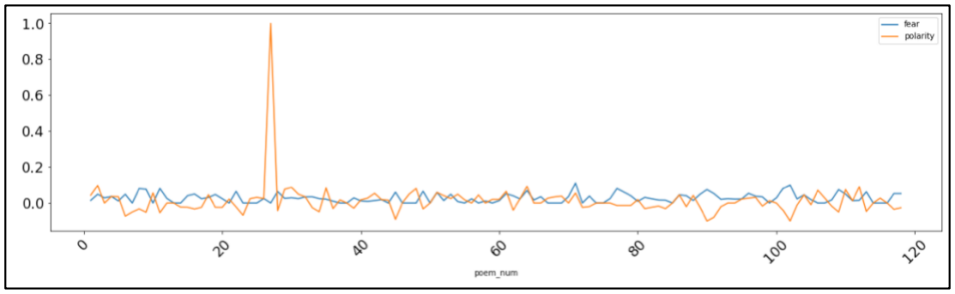


To dig deeper into our sentiment analysis, we look at how different sentiments translate throughout the poems of each series. these visualizations, I used "poems" as the bag over a series to see the change in sentiment over the poems in a series. Based off our previous analysis, we saw that each of the three series in our data produce similar results, so we only use the poems from Series 1 for the below sentiment visualizations. To develop these visualizations, I used the top 4 emotions that were displayed in the bar plots above (joy, trust, fear, and anticipation). Looking at the plot for these four sentiments, we see that, for the most part, there seems to be similar consistency in the prevalence of these sentiments throughout the poems in the series. Again, this shows that there is a consistent mixture of these emotions in Dickinson's poetry which points to her having a similar theme and tone within her poems. The one big observation that we see is that the sentiment "joy" spikes up at a certain point in the plot. There could be two potential takeaways here: 1) this is an error in the text or code that is causing this sensation or 2) there is a specific poem in Series 1 that reflects a lot more joy than generally reflected in Dickinson's' poetry. Despite this spike, the major takeaway that we can gather from this plot is that Dickinson seems to keep to consistent themes and stories within her poetry and are most likely integrated throughout most of the poetry under analysis.



In the next three sentiment plots, I focused comparing polarity to the emotions of fear and jo. From these comparisons, we can deduce that the plots are reasonable because, looking at polarity and joy, we see that their trends are similar (polarity and joy increase/decrease together). Looking at polarity vs fear, though, we expect (and see) that an increase in polarity leads to a decrease in fear (and vice versa).





Overall, the analysis of Emily Dickinson's poetry provided potential insights into the type of author she was and the stories that she wrote. From the four categories under observation (Life, Love, Nature, and Time & Eternity), we saw that there might be overlap in these categories throughout her poetry and that it may actually be difficult to categorize her poems under one umbrella. Despite that, using TFIDF statistics, we were able to see that some of the most influential words in the text were "love" and "nature", which reflected our topics of interest. The analysis also demonstrated that Emily Dickinson's poetry is consistent in its story telling and does not stray far from her niche. Through the three different series of texts that came with our text file, our sentiment analysis showed that the emotions that she emphasized remained the same throughout her poetry (joy, trust, anticipation, and fear). Looking at the top terms of our four topics, we saw that many of the top terms were found in the top terms of other topics as well. Our sentiment analysis also detailed that her poetry, though mainly positive, integrated negative sentiments as well.