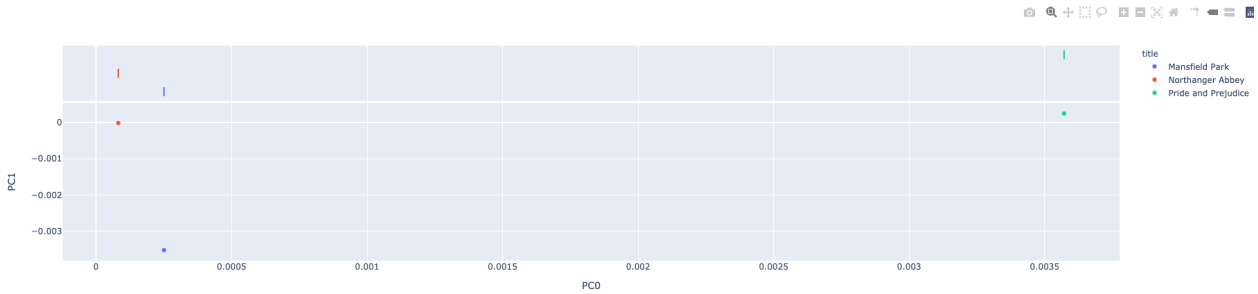
Text Analysis on Select Jane Austen Works

Drake Wagner
Dbw2tn

For my final project, we decided to further analyze three of Jane Austen's most famous novels. These works include *Mansfield Park, Northanger Abbey, and Pride and Prejudice*. All three of these files were obtained from Project Gutenberg, like the majority of the files we used in class. We chose this website because of its reliability and easy access to archived ebooks. We chose these specific books because, having read two of them, they seemed very happy and upbeat for the majority of the time. Therefore, I wanted to see if exploratory text analysis could statistically show me whether or not this was true, in terms of sentience. We began by running a principal component analysis, then ran linear discriminant analysis, ran some word embedding, and lastly looked at the sentiment analysis of the three stories.

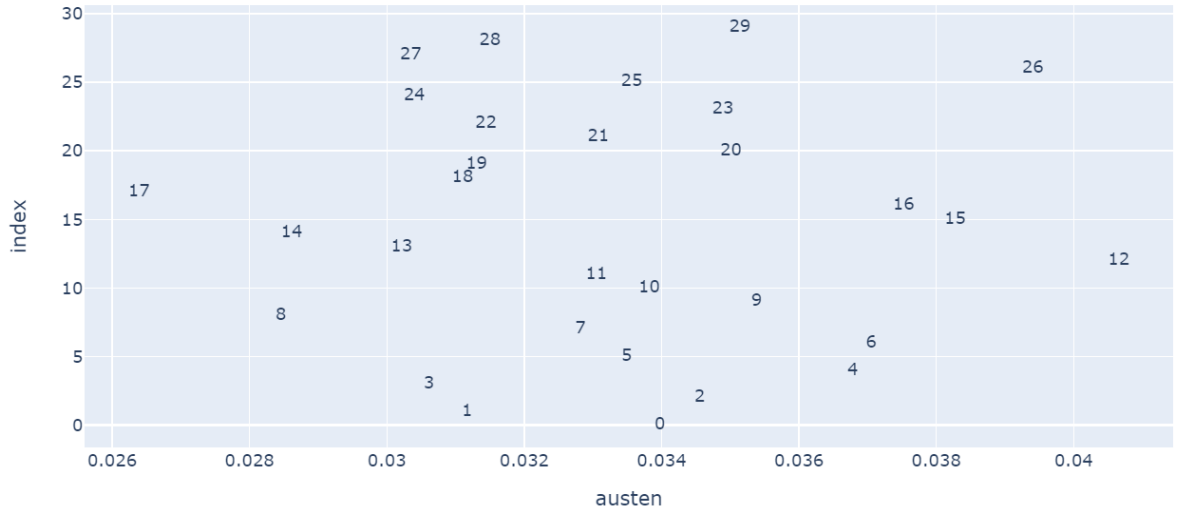
Each file used includes a csv file called "library" which includes the metadata for source files, a csv file called "token" which includes statistical and linguistic data of the tokenized texts, and "vocab", which includes more detailed data on the terms used and their statistics. All three of these core tables work together in our analysis.

Principal Component Analysis was the first statistical analysis we ran on the data, as a means to analyze the features and components of the texts. Sometimes simple clustering is not enough information, since texts with similar contexts use different words, while simple clustering also does not always capture the subject matter of the texts. Therefore, using bagging through PCA shows us that words in the same bags that don't ever appear together are likely synonyms. For example, you won't say "The fox ran quickly fast." But if there were two different sentences, one that said "quickly" and one that said "fast" in the same context, then bagging can deduce that they are likely synonyms. In addition to this, PCA provides us with a covariance matrix, similar to pairwise matrices, showing the distance between terms in a linear format. This lets us add both docs and components, as well as words and components, to our database. Projecting this matrix, running PCA, and identifying the axes of most variance/information, we get this graph showing the 3 texts that we plotted, graphed by the first and second principal components. This visualization shows us the similarity in language between each of the books, specifically that *Mansfield Park* and *Northanger Abbey* are more similarly phrased than when compared to *Pride and Prejudice*, which is in a much further away spatial location .

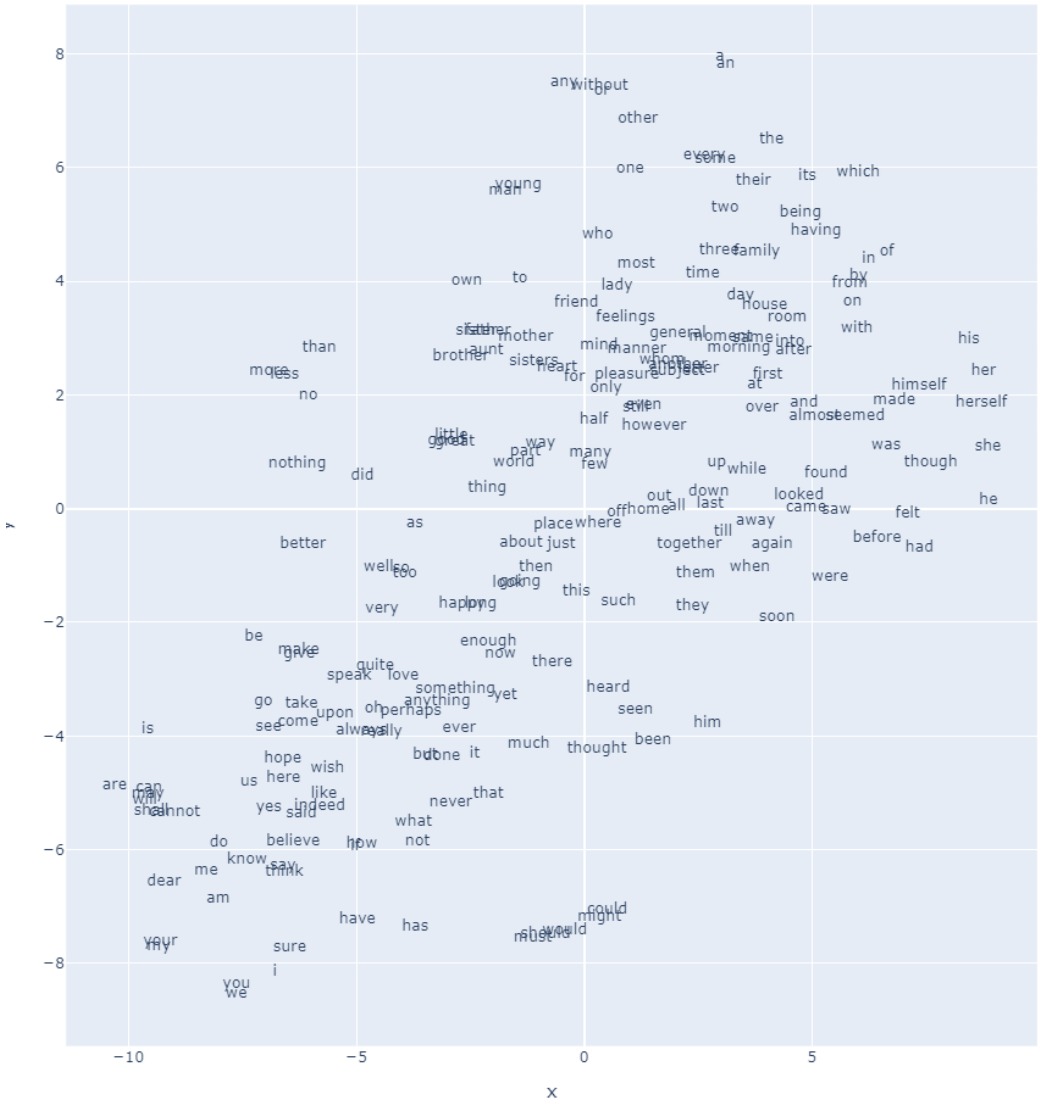


Our next analysis, Linear Discriminant Analysis, is a way for us to analyze topic models within the texts. In other words, what are some of the general themes of each story? In more detail, we use LDA to look at the distribution of words, which give us information about the culture and language of the text. We also look at which topics are written about and prioritized, using a topic table and graph. Doing this allows us to deduce that both of the tested books here prefer topics 12 and 26 (see figure below). If we look at our table of top terms, we see that both of these topics prioritize family and happiness members and happiness, as expected.

author	austen	topterms
topic_id		
12	0.040655	yes thing sister good time months oh moments pleasure away
26	0.039403	time friend day letter sister uncle father brother morning character
15	0.038274	answer time day pleasure feelings way mother oh mind word
16	0.037521	sir sisters care oh sister town power men wonder cousin
6	0.037054	room time countenance moment man mother letter day beauty way



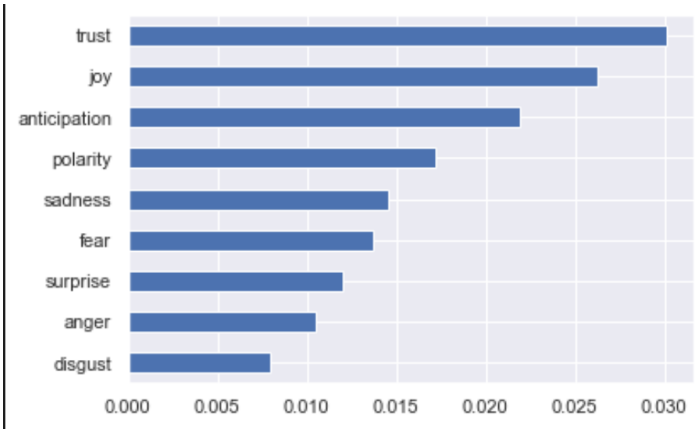
Next, we used Word2Vec as a way of word embedding the texts. Word2Vec works by using one-hot encoding to label each word, then attributes their context as either positive or negative. We first had to group our token corpus into paragraphs, assigning that as our new bag. After passing this new corpus into the Word2Vec function via Python, the results were stored in a dataframe, which can then be visualized by running the TSNE function. This plots the different points as coordinates on a word plot, shown below.

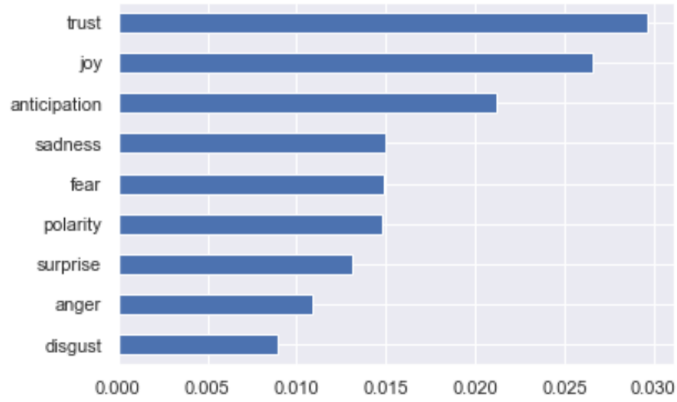


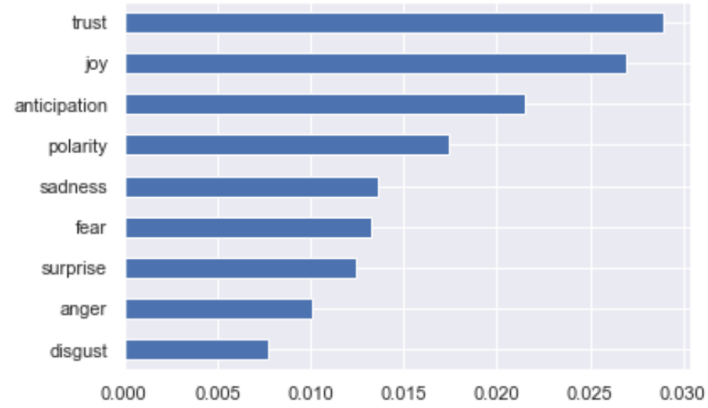
One again, we are able to see many terms for family members, but also an overarching theme of happiness, through terms such as “love”, “together”, “hope”, “dear”, and “happy”, just to name a few. Many of these emotional words are nearby other more generic words, leading us to believe that happy-iness is a core theme throughout the texts.

The last piece of our analysis was done by running Sentiment analysis on the corpus. After importing the lexicon csv and joining the token column to the salex table, our new token table was ready to be run through some visualizations. Below, in order, are the emotion graphs listed for *Mansfield Park, Northanger Abbey,* and *Pride and Prejudice*. It is interesting to see that trust, joy, and sadness are the top three emotions for each text, in the same order. Going into the analysis, I had guessed that joy would be at the top, but trust (another positive emotion) is up there instead. While sadness is also included in each book, it is only used about half as much as

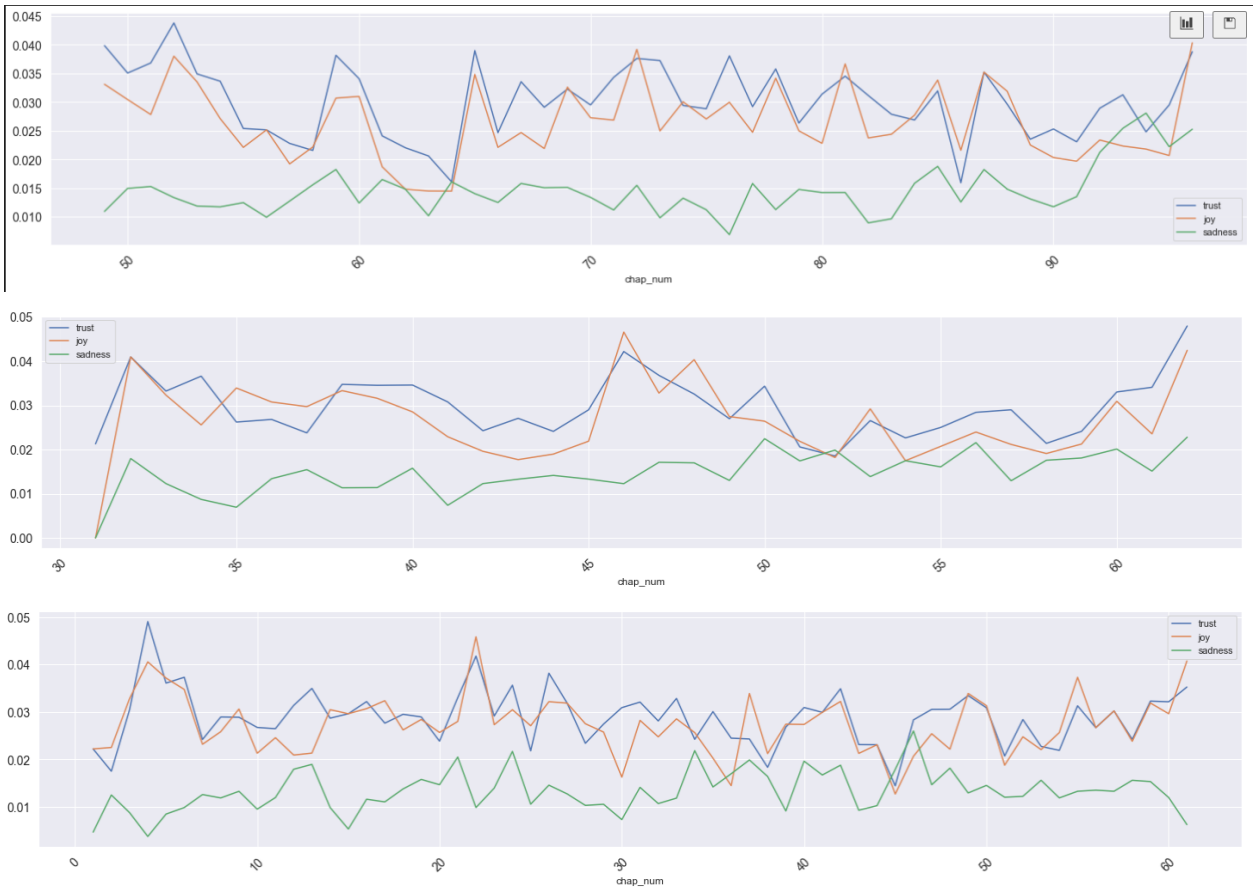
joy and trust are. It is easy to see that negative emotions are used less frequently than positive emotions in these three books.







We also look at the fluctuations of these emotions over time, via these graphs below, in the same book order as above.



These graphs of sentiment over time reinforce earlier conclusions that sadness is not nearly as prevalent in these texts as joy and trust are.

Having done our analysis, we can conclude that these three Austin books do, in fact, have a positive, happy tone in the text. There are many similarities between all three of these texts, hinting that Austen's other works likely follow this same theme of joy and trust reigning above sadness. While PCA showed a slight difference between *Pride and Prejudice* and the other texts, LDA, Word2Vec, and sentiment analysis all showed the vast similarities between the three, including similar topics (through LDA), similarities in words and vocabulary (Word2Vec), and the sentiment within the text over the length of the book. This final sentiment analysis is arguably the most important in proving our hypothesis, although all of these methods were valuable resources in this overall analysis.