Xi Wang (xw2ez@virginia.edu)
DS 5001 Exploratory Text Analytics
May 6, 2024

Final Report

I. Introduction

Throughout this final project, I have delved into the literature available through Project Gutenberg, a repository of public domain texts, to conduct an in-depth text analysis of nine selected literary works from three different authors. The nine books I selected were *Emma*, *Mansfield Park*, *Pride and Prejudice*, *Adam Bede*, *Israel Potter*, *The Apple - Tree Table and Other Sketches*, *The Piazza Tales*, *Middlemarch*, and *The Mill on the Floss*. By leveraging the techniques and tools learned through this course, I aim to uncover underlying, cultural patterns, topics, and stylistic elements within these long-form texts. By exploring the writings of Jane Austen, George Eliot, and Herman Melville, I want to gain insights into their respective narrative styles and the socio-historical contexts that shaped their literary output.

II. Methods

In this project, I followed the instructions to acquire a collection of long-form texts and perform ETA operations from beginning to end. To begin, I converted the source formats of the texts into structured tables conforming to F2 and F1. I then annotated these tables with statistical and linguistic features, including stopwords, parts-of-speech, stems and lemmas, etc. Next, I generated a vector representation of the corpus using TFIDF values, which were incorporated into the TOKEN and VOCAB tables. Building upon this foundation, I applied methods learned in class, including Principal Component Analysis (PCA), Latent Dirichlet Allocation (LDA), and Word2Vec, to further extend the annotated and vectorized model. By leveraging these methods, I intend to explore the results regarding cultural patterns.

III. Results and Interpretation

A. PCA

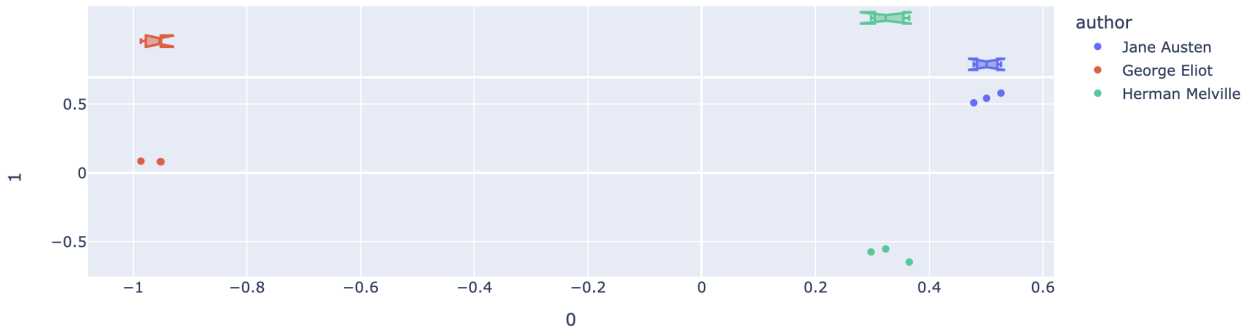
Based on the image below, the high negative loadings for PC0 for *Middlemarch*, *Adam Bede*, and *The Mill on the Floss* (book id = 145, 507, and 6688, respectively) shows that the three books by George Eliot are all strongly inversely correlated with PC0. Thus, literary works by George Eliot are similar. However, works by Jane Austen and Herman Melville, though still have similarities, are also different in some patterns.

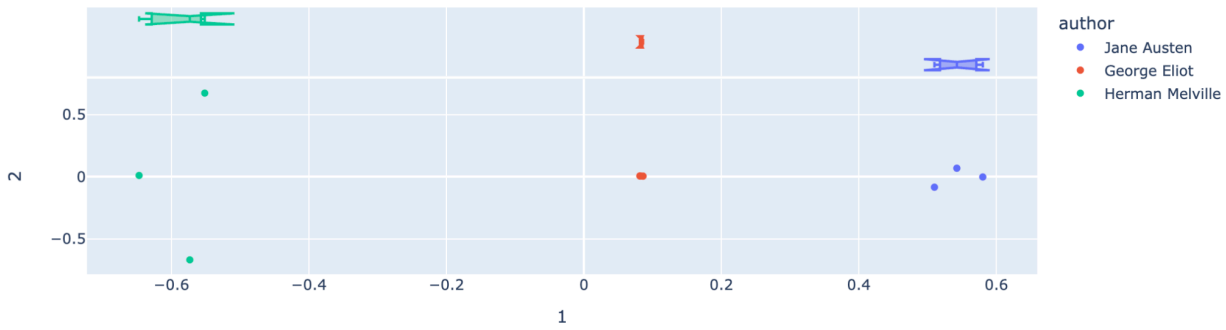
Regarding Jane Austen, *Mansfield Park* (book id = 141) is positively associated with PC0 and PC1, is negatively associated with PC5, and has very weak association with other components. *Emma* is also positively correlated with PC0 and PC1; however, it is positively correlated with PC5 and negatively correlated with PC3. As for *Pride and Prejudice*, it is positively associated with PC0, PC1, and PC3.

In terms of Herman Melville, *Israel Potter* is negatively associated with PC1 and positively correlated with PC4. *The Piazza Tales* is negatively associated with PC1 and PC2. *The Apple - Tree Tables and Other Sketches* is positively associated with PC2, while it is negatively associated with PC1. All the three books by Herman Melville have a negative relationship with PC1.

book_id	0	1	2	3	4	5	6	7	8	9	author	title
141	0.526236	0.580221	-0.002076	-0.138653	0.053246	-0.603617	-0.000557	-0.000305	-2.306640e-15	2.223915e-15	Jane Austen	Mansfield Park
145	-0.952238	0.082215	0.005230	-0.002085	0.014584	-0.002804	-0.284386	0.073181	1.222655e-15	-6.752140e-16	George Eliot	Middlemarch
158	0.500842	0.542498	0.068471	-0.511071	0.035240	0.433277	0.000227	-0.000679	4.956430e-16	-3.116431e-15	Jane Austen	Emma
507	-0.950940	0.081385	0.006205	-0.002233	0.014823	-0.004483	0.289836	0.069293	1.116295e-15	-1.238593e-15	George Eliot	Adam Bede
1342	0.478346	0.509895	-0.084514	0.678966	-0.050830	0.201172	0.000149	-0.000946	2.087523e-15	1.311478e-15	Jane Austen	Pride and Prejudice
6688	-0.986358	0.086030	0.004595	-0.004093	0.014706	-0.005731	-0.005226	-0.139223	-2.794206e-15	-3.560520e-16	George Eliot	The Mill on the Floss
15422	0.364926	-0.647098	0.010123	0.050363	0.667210	0.016876	-0.000260	-0.000849	1.767250e-17	4.720616e-16	Herman Melville	Israel Potter
15859	0.297757	-0.573304	-0.669519	-0.106779	-0.350531	-0.011589	0.000041	-0.000557	-2.455718e-16	4.993835e-16	Herman Melville	The Piazza Tales
53861	0.323501	-0.551576	0.673200	0.029893	-0.368741	-0.032370	-0.000389	-0.000850	2.519957e-16	5.271391e-16	Herman Melville	The Apple - Tree Tables and Other Sketches

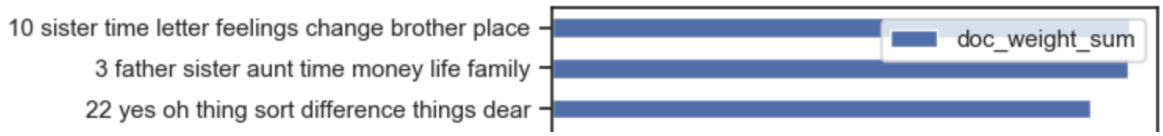
Then, I explored the literary style of different authors by utilizing PCA. As the images below demonstrate, PC0 and PC1, and PC1 and PC2, to a large extent, distinguish the three authors. We can see that works by the same authors are clustered together, indicating that they are self-similar. However, this distinct separation may be attributed to the small sample size.





B. LDA

Overall, if all nine books are considered, the top three topics sorted by doc weight are topic 10, 3, and 22 as indicated below. These topics are related to family and human relationships, such as sister, brother, father, and family.



OHCO by Paragraphs

The three images below reveal the top three topics by author when OHCO is paragraphs. Specifically, the top three topics for works by George Eliot are topic 25, 20, and 13. Looking at the corresponding terms, George Eliot often uses words such as “man”, “woman”, “time”, etc. For Herman Melville, the top three topics are 4, 24, and 21, and top terms are nautical-related, such as “sea”, “ship”, “man”, etc. Regarding Jane Austen, the top three topics are 10, 3, and 6, and top terms are “sister”, “brother”, “family”, “time”, etc. The top topics are all different for these three authors; thus there is a significant difference among the three authors.

author	George Eliot	Herman Melville	Jane Austen	topterms
topic_id				
25	0.045651	0.036510	0.024387	ll time ah house head day glass sure way box
20	0.043516	0.024251	0.025398	mind voice man moment words hands silence moments woman ye
13	0.041597	0.031416	0.031248	don children father man mother world boy people things care

author	George Eliot	Herman Melville	Jane Austen	topterms
topic_id				
4	0.022967	0.106349	0.014360	sea ship sight face air glance officer time ships eyes
24	0.036653	0.078700	0.019776	chimney door house way wife head half night room morning
21	0.032376	0.043985	0.018856	sir man anybody house time reason em way mind money

author	George Eliot	Herman Melville	Jane Austen	topterms
topic_id				
10	0.026849	0.015427	0.083828	sister time letter feelings change brother place father business family
3	0.034806	0.017011	0.069653	father sister aunt time money life family sense man love
6	0.025793	0.028063	0.061001	time thing body friend family daughter things day party idea

OHCO by Books

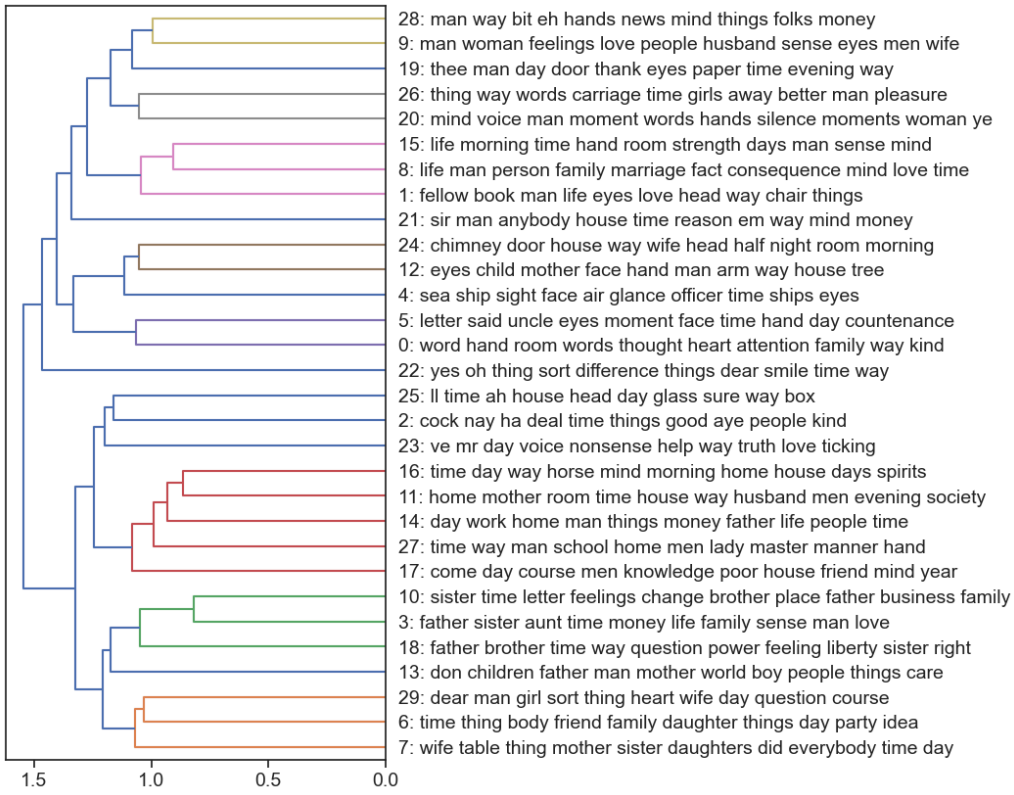
The three images below reveal the top three topics by author when OHCO is books. After changing the OHCO from paragraphs to books, topics changed. Although the topic id changes, the top terms overlap. In addition, the top terms for the three authors also overlap. They all seem to focus on human relationships within society.

author	George Eliot	Herman Melville	Jane Austen	topterms
topic_id				
14	0.780543	0.012489	0.108776	man way time life mind things father day eyes money
16	0.214340	0.000005	0.000405	man way day thee time face eyes work mother woman
24	0.003743	0.654125	0.003360	man time chimney wife way day sort house sir men

author	George Eliot	Herman Melville	Jane Austen	topterms
topic_id				
24	0.003743	0.654125	0.003360	man time chimney wife way day sort house sir men
27	0.000001	0.333262	0.000002	sea time isle ship man tower isles boat day tortoises
14	0.780543	0.012489	0.108776	man way time life mind things father day eyes money

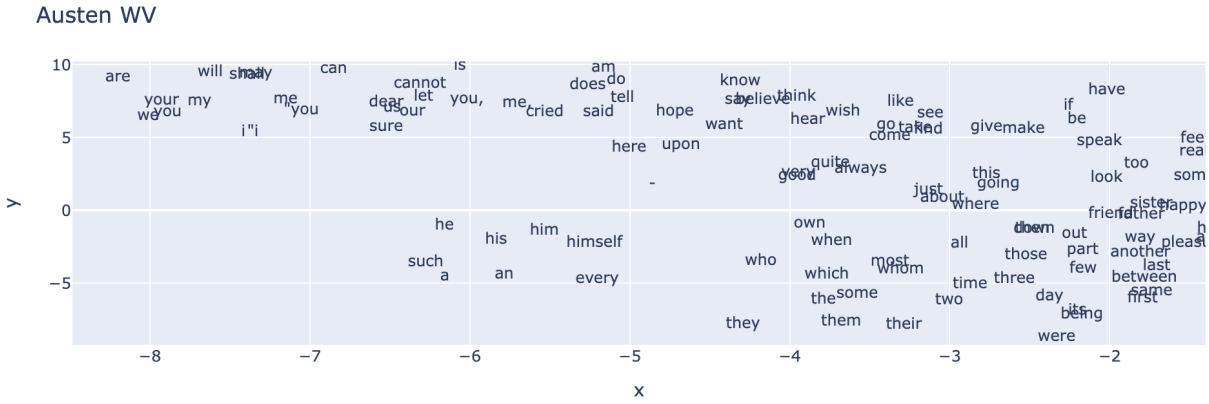
author	George Eliot	Herman Melville	Jane Austen	topterms
topic_id				
3	0.001347	0.000005	0.605062	time sister room day family house man mother feelings father
29	0.000005	0.000005	0.282355	thing time body man father friend oh day way home
14	0.780543	0.012489	0.108776	man way time life mind things father day eyes money

Based on the image below, we can see how topics are clustered and related to each other. The top three topics for Herman Melville (4, 24, and 21) are clustered together. Two of the top three topics for Jane Austen (10 and 3) are very closely related.

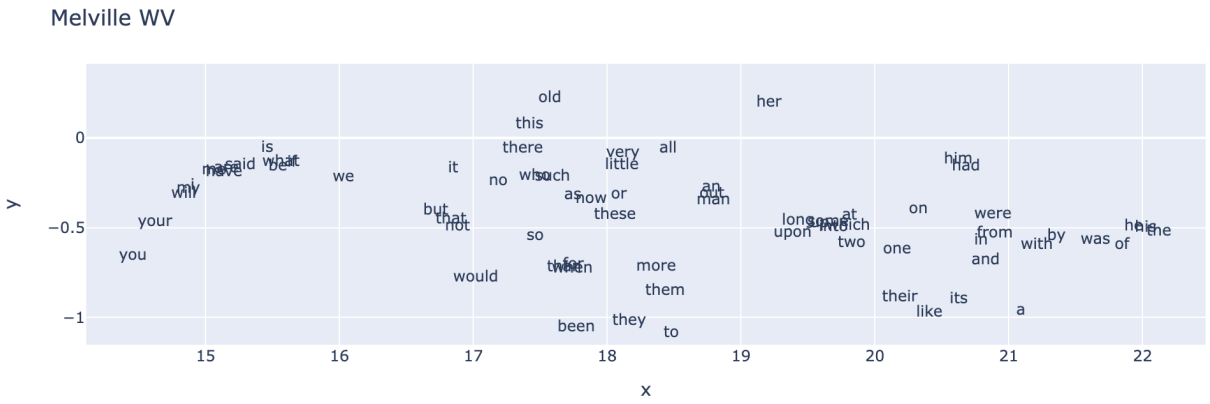


C. Word2Vec

As we have visualized in the image below, which are works by Jane Austen only, pronouns form a cluster and relative pronouns form a cluster, and action words (e.g., cried, said, does, tell, hope, and want) form a cluster.



The image below shows works by Herman Melville. From the clusters, we can see that prepositions (e.g., with, by, of, and from) form a cluster, and words such as you, your, and my form a cluster.

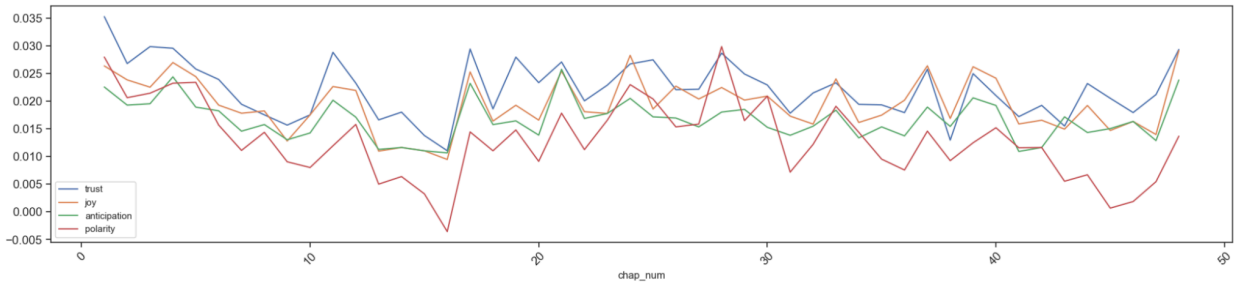


The image below shows works by George Eliot, and I cannot see distinct clusters. They are all clustered together.

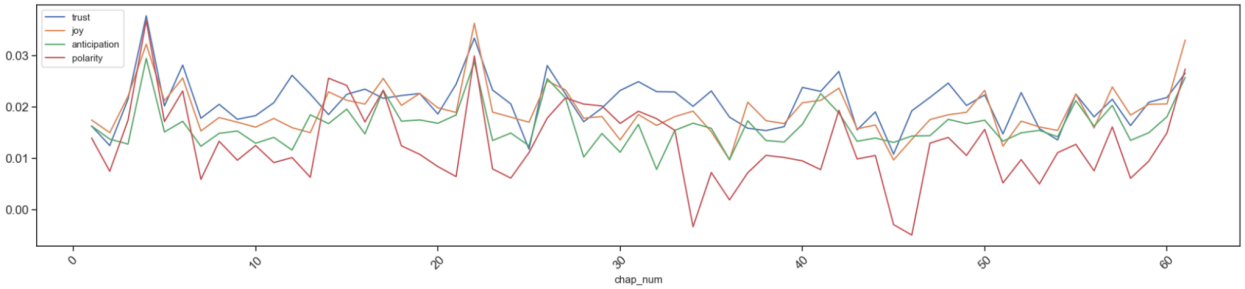
D. Sentiment Analysis

For sentiment analysis, I ran an analysis to explore the sentiment of all nine books by chapters. Due to space constraints, I will just include a few examples that I found interesting. Regarding emotions, the top three emotions for works by Jane Austen and George Eliot are trust, joy, and anticipation, while the top three emotions for Herman Melville are more diverse, including trust, joy, fear, sadness, and anticipation.

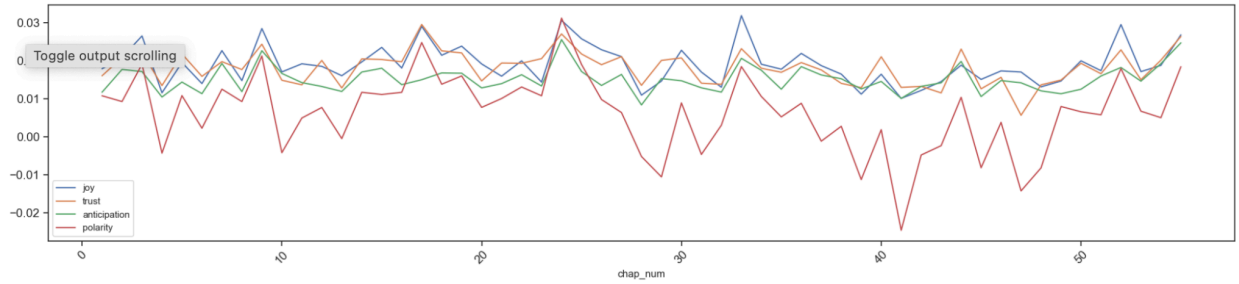
The plot below for *Mansfield Park* by Jane Austen indicates that the book starts high on all emotions and polarity. It then gradually decreases and reaches the first low point at around chapter 16. It then fluctuates and increases for joy, trust, and anticipation.



The plot below is for *Pride and Prejudice* by Jane Austen, and it has two obvious bumps (one at the beginning and around chapter 21).



The plot for *Adam Bede* by Geroge Eliot has constantly low polarity, most of the time below 0.



IV. Conclusion and Limitations

In conclusion, this final project has provided a comprehensive exploration of nine selected literary works from three distinct authors, namely Jane Austen, George Eliot, and Herman Melville. Through the application of various text analytics techniques and methodologies learned throughout the course, I have gained a deeper understanding of the cultural contents, narrative styles, and thematic preoccupations within these long-form texts. PCA provides insights into the similarities and differences among the books. The clustering of works by the same author indicates consistent narrative styles and elements, while the distinct separation between authors suggests unique writing styles and thematic focuses. Further analysis using LDA and Word2Vec has provided additional understanding of the cultural contents and themes present in the corpus. By exploring topics derived from LDA and word clusters from Word2Vec, common themes such as family relationships, societal norms, and nautical imagery have been identified across the works of each author. Specifically, Jane Austen tends to focus on human relationships within society, while Herman Melville tends to emphasize adventure and the natural world. Lastly, sentiment analysis has offered insights into the emotional nuances and tonal shifts within the texts. While works by Jane Austen and George Eliot predominantly evoke sentiments of trust, joy, and anticipation, those by Herman Melville exhibit a broader spectrum of emotions, including fear, sadness, and anticipation.

However, the limited selection of three books from each author may constrain the generalizability of the patterns or trends identified in this analysis. The small sample size

limits the scope of our findings and may not fully capture the breadth and diversity of each author's literary output. Additionally, the selection bias inherent in choosing specific works for analysis may skew our interpretations and conclusions.