Exploratory Text Analytics Final Project

Analysis of Classical Greek and Roman Works

Author: Alexandra Ferentinos

DS5001

Date: 12-06-2024

Introduction:

The stories of the pantheon, the gods, heroes, monsters, and love weave literary narratives that paint the stories of the beliefs and life perspectives of the past. This report explores the corpus of Classical Greek and Roman literature, as they are often considered foundational texts that still inspire modern literature, media, and pop culture today. The corpus includes nine classic works spanning epic poetry, drama, and mythological narratives, from Homer's well-known epics to Ovid's tale of metamorphosis.

doc_id	book_id	title
0	31	Sophocles-31: Plays of Sophocles: Oedipus the ...
1	228	Virgil-228: The Aeneid
2	348	Hesiod-348: Hesiod, the Homeric Hymns, and Hom...
3	830	Apollonius Rhodius-830: The Argonautica
4	1727	Homer-1727: The Odyssey
5	6130	Homer-6130: The Iliad
6	7700	Aristophanes-7700: Lysistrata
7	21765	Ovid-21765: The Metamorphoses of Ovid, Books I...
8	35451	Euripides-35451: Medea of Euripides

Figure 1: DOC table displaying the Book ID and Title of the works in the corpus.

Figure 1 shows the overview of the titles of the works in the corpus. The works range from Sophocles' Oedipus trilogy that explore themes of fate vs free will; Homer's Iliad and Odyssey which describe the events of the Trojan war and Odysseus homecoming; and Ovid's Metamorphoses, which show the perspectives of the gods vs man, as well as what is morally right vs wrong. This corpus covers a robust source of universal themes that will be explored in the exploratory text analysis of these works, such as divine intervention, heroism, and free will.

Using natural language processing techniques learned over this semester mainly: topic modeling, and sentiment analysis, this report aims to analyze the patterns present in the authors approach of classical themes. The themes of divine intervention, heroism, free will, and many others.

Data Collection:

The data was collected from the Project Gutenberg website that has a collection of over 70,000 free eBooks for use and download. [i] The texts were downloaded as plain text UTF-8 files that contain Gutenberg metadata, as well as the classical text. Each text file was approximately over 10,000 lines long and were stored in a storage directory. The text files were then

individually analyzed by hand to determine the proper regex format required per work to discern the chapter formatting, as well as unnecessary metadata removal.

The nine texts stored in the directory were then accessed by the acquire_epubs function used and implemented throughout the semester. These texts were then tokenized, reduced, and a document table and a vocabulary table were created (F1-F5 Text data forms). When the Token table was made it still contained stop words; this would have added noise to the later analysis of the works. A dictionary of common stop words was made and used to filter it down to a reduced token table. This reduced table was added to the csv file read in pipeline to streamline the analysis of the works.

Analysis:

As mentioned previously, the corpus is comprised of nine classical Greco-Roman texts with the following categories and distributions of tokens: Epic Poetry category; Homer's Iliad and Odyssey (approx. 147,000 tokens), Virgil's Aeneid (approx. 56,000 tokens), Ovid's Metamorphoses (approx. 58,000 tokens), Apollonius' Argonautica (approx. 21,000 tokens). Drama category; Sophocles' plays (approx. 13,500 tokens), Euripides' Medea (approx. 7,500 tokens), Aristophanes' Lysistrata (approx. 6,300 tokens). Finally, the Historical category (approx. 13,000 tokens). The categories are useful for analysis and when interpreting trends in the Zipf distribution, as shown in Figure 2. When interpreting the graph, there is a concentrated density of Epic Poetry category tokens in the upper ranges of the plot, while the Drama category is densely comprised that of the mid-range of the graph.

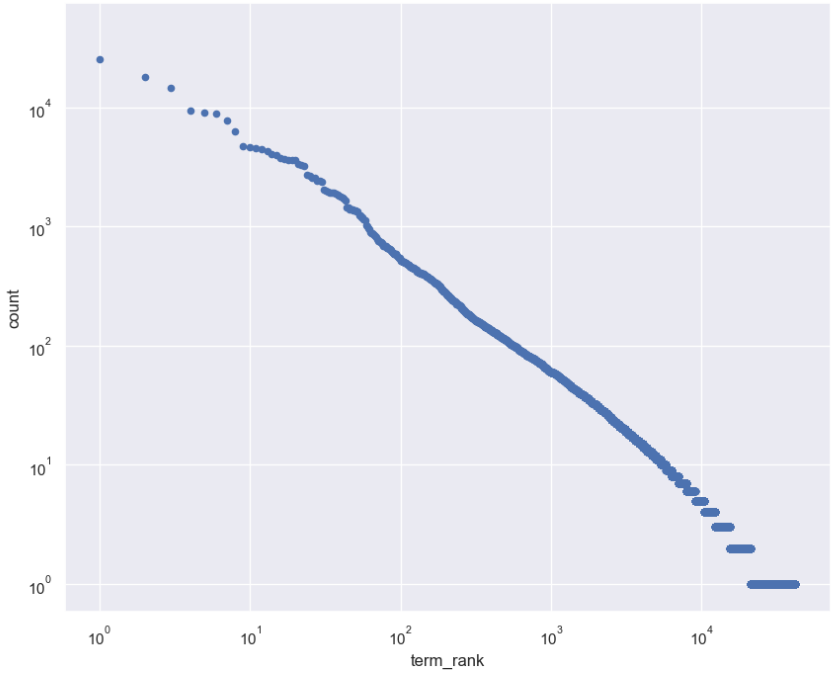
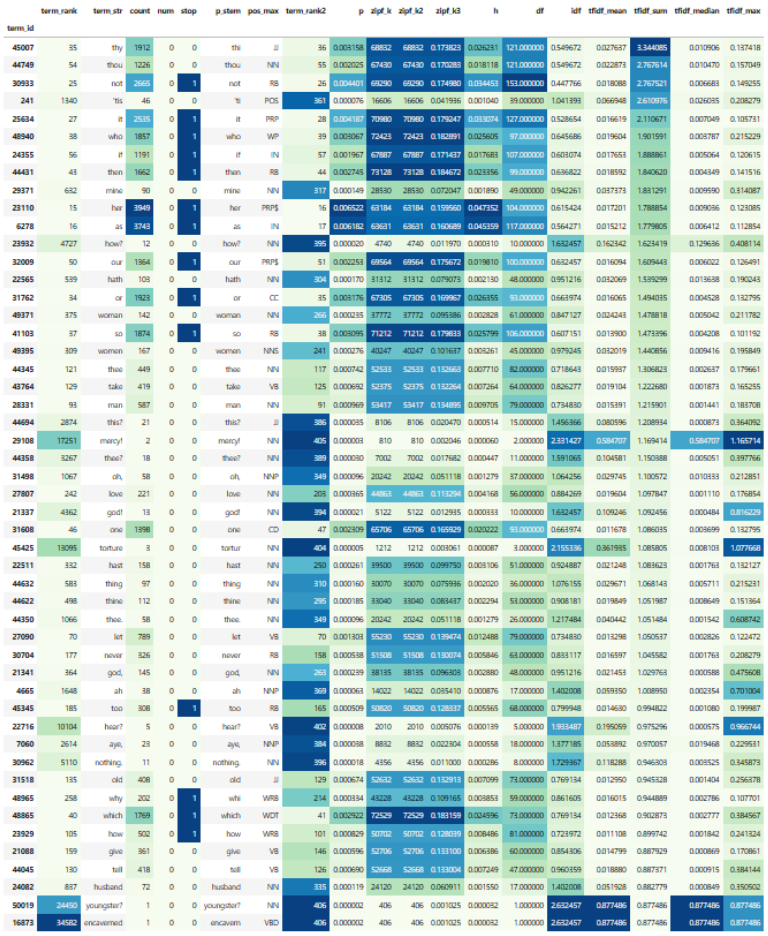


Figure 2: Work Frequency Zipf distribution of the Classical Corpus.

The TFIDF significance matrix of the classical text corpus was effective in illuminating the patterns of term usage across the various texts of the corpus. Figure 3 is a graph of the heatmap of the TFIDF weights. It incorporates term rank, polarity, and significance scores across the texts, where the darker blue color indicates a higher significance score.



term_id	term_rank	term_str	count	num	stop	p_stem	pos_max	term_rank2	p	zipf_k	zipf_k2	zipf_k3	h	df	idf	tfidf_mean	tfidf_sum	tfidf_median	tfidf_max
45007	35	thy	1912	0	0	thi	JJ	36	0.003158	68852	68832	0.173823	0.026231	121.000000	0.549672	0.027637	3.344085	0.010906	0.137418
44749	54	thou	1226	0	0	thou	NN	55	0.002029	67430	67430	0.170283	0.018118	121.000000	0.549672	0.022873	2.767614	0.010470	0.157049
30933	25	not	2665	0	1	not	RB	26	0.004401	69290	69290	0.174980	0.034453	153.000000	0.447766	0.018088	2.767521	0.006683	0.149255
241	1340	'tis	46	0	0	'ti	POS	361	0.000076	16606	16606	0.041936	0.001040	39.000000	1.041393	0.066948	2.610976	0.026035	0.208279
25634	27	it	2535	0	1	it	PRP	28	0.004187	70980	70980	0.179247	0.033074	127.000000	0.528654	0.016619	2.110671	0.007049	0.105731
48940	38	who	1857	0	1	who	WP	39	0.003067	72423	72423	0.182891	0.025605	97.000000	0.645686	0.019604	1.901591	0.005787	0.215229
24355	56	if	1191	0	1	if	IN	57	0.001967	67887	67887	0.171437	0.017683	107.000000	0.603074	0.017653	1.888861	0.005064	0.120615
44431	43	thin	1662	0	1	thin	RB	44	0.002748	73128	73128	0.184672	0.023356	99.000000	0.636822	0.018592	1.840620	0.004349	0.141516
29371	652	mine	90	0	0	mine	NN	317	0.000149	28580	28580	0.072047	0.001890	49.000000	0.942261	0.037373	1.881291	0.009590	0.314087
23110	15	her	3949	0	1	her	PRP$	16	0.006522	63184	63184	0.159560	0.047352	104.000000	0.615424	0.017201	1.788854	0.009096	0.123085
6278	16	as	3743	0	1	as	IN	17	0.006182	63631	63631	0.160689	0.045359	117.000000	0.564271	0.015212	1.779805	0.006412	0.112854
23932	4727	how?	12	0	0	how?	NN	395	0.000020	4740	4740	0.011970	0.000310	10.000000	1.632457	0.162342	1.623419	0.129636	0.408114
32009	50	our	1364	0	1	our	PRP$	51	0.002253	69564	69564	0.175672	0.019810	100.000000	0.632457	0.016094	1.609443	0.006022	0.126491
22565	539	hath	103	0	0	hath	NN	304	0.000170	31312	31312	0.079073	0.002130	48.000000	0.951216	0.032069	1.539299	0.013638	0.190243
31762	34	or	1923	0	1	or	CC	35	0.003176	67305	67305	0.169967	0.026355	93.000000	0.663974	0.016065	1.494035	0.004528	0.132795
49371	375	woman	142	0	0	woman	NN	266	0.000235	37772	37772	0.095386	0.002828	61.000000	0.847127	0.024243	1.478818	0.005042	0.211782
41103	37	so	1874	0	1	so	RB	38	0.003095	71212	71212	0.179833	0.025799	106.000000	0.607151	0.013900	1.473896	0.004208	0.101192
49395	309	women	167	0	0	women	NNS	241	0.000276	40247	40247	0.101637	0.003261	45.000000	0.979245	0.032019	1.440856	0.009416	0.195849
44345	121	thee	449	0	0	thee	NN	117	0.000742	52533	52533	0.132663	0.007710	82.000000	0.718643	0.015997	1.306823	0.002687	0.179661
43764	129	take	419	0	0	take	VB	125	0.000692	52375	52375	0.132264	0.007264	64.000000	0.826277	0.019104	1.222680	0.001873	0.165255
28331	93	man	587	0	0	man	NN	91	0.000969	53417	53417	0.134895	0.009705	79.000000	0.734830	0.015391	1.215901	0.001441	0.183708
44694	2874	this?	21	0	0	this?	JJ	386	0.000035	8106	8106	0.020470	0.000514	15.000000	1.456366	0.080596	1.208934	0.000873	0.364092
29108	17251	mercy!	2	0	0	mercy!	NN	405	0.000003	810	810	0.002046	0.000060	2.000000	2.331427	0.584707	1.169414	0.584707	1.165714
44358	3267	thee?	18	0	0	thee?	NN	389	0.000030	7002	7002	0.017682	0.000447	11.000000	1.591065	0.104581	1.150888	0.005051	0.397766
31498	1067	oh,	58	0	0	oh,	NNP	349	0.000096	20242	20242	0.051118	0.001279	37.000000	1.064256	0.029745	1.100572	0.010333	0.212851
27807	242	love	221	0	0	love	NN	203	0.000369	44863	44863	0.113294	0.004168	56.000000	0.884269	0.019604	1.097847	0.001110	0.176854
21337	4362	god!	13	0	0	god!	NN	394	0.000021	5122	5122	0.012935	0.000333	10.000000	1.632457	0.109246	1.092456	0.000484	0.816229
31608	46	one	1398	0	0	one	CD	47	0.002309	65706	65706	0.165929	0.020222	93.000000	0.663974	0.011678	1.086035	0.003699	0.132795
45425	13095	torture	3	0	0	tortur	NN	404	0.000005	1212	1212	0.003061	0.000087	3.000000	2.155336	0.861985	1.085805	0.008108	1.077668
22511	332	hast	158	0	0	hast	NN	250	0.000261	39500	39500	0.099750	0.003106	51.000000	0.924887	0.021248	1.083623	0.001763	0.132127
44632	583	thing	97	0	0	thing	NN	310	0.000160	30070	30070	0.075936	0.002020	36.000000	1.076155	0.029671	1.068143	0.005711	0.215231
44622	498	thine	112	0	0	thine	NN	295	0.000185	33040	33040	0.083437	0.002294	53.000000	0.908181	0.019849	1.051987	0.008649	0.151364
44350	1066	thee.	58	0	0	thee.	NN	349	0.000096	20242	20242	0.051118	0.001279	26.000000	1.217484	0.040442	1.051484	0.001542	0.608742
27090	70	let	789	0	0	let	VB	70	0.001808	55230	55230	0.139474	0.012488	79.000000	0.734830	0.013298	1.050537	0.002826	0.122472
30704	177	never	326	0	0	never	RB	158	0.000538	51508	51508	0.130074	0.005846	63.000000	0.833117	0.016597	1.045582	0.001763	0.208279
21341	364	god,	145	0	0	god,	NN	263	0.000239	38135	38135	0.096308	0.002880	48.000000	0.951216	0.021453	1.029763	0.000588	0.475608
4665	1648	ah	38	0	0	ah	NNP	369	0.000061	14022	14022	0.035410	0.000876	17.000000	1.402008	0.059350	1.008950	0.002354	0.701004
45345	185	too	308	0	1	too	RB	165	0.000509	50620	50620	0.128337	0.005565	68.000000	0.799948	0.014830	0.994822	0.001080	0.199987
22716	10104	hear?	5	0	0	hear?	VB	402	0.000008	2010	2010	0.005076	0.000139	5.000000	1.933487	0.195059	0.975296	0.000575	0.966744
7060	2614	aye,	23	0	0	aye,	NNP	384	0.000038	8832	8832	0.022304	0.000558	18.000000	1.377185	0.053892	0.970057	0.019468	0.229531
30962	5110	nothing.	11	0	0	nothing.	NN	396	0.000018	4356	4356	0.011000	0.000286	8.000000	1.729967	0.118288	0.946303	0.005525	0.345873
31518	135	old	408	0	0	old	JJ	129	0.000674	52632	52632	0.132913	0.007099	73.000000	0.769134	0.012950	0.945328	0.001404	0.256378
48965	258	why	202	0	1	whi	WRB	214	0.000334	43228	43228	0.109165	0.003853	59.000000	0.861605	0.016015	0.944889	0.002786	0.107701
48865	40	which	1769	0	1	which	WDT	41	0.002922	72529	72529	0.183159	0.024596	73.000000	0.769134	0.012368	0.902873	0.002777	0.384567
23929	105	how	502	0	1	how	WRB	101	0.000829	50702	50702	0.128035	0.008486	81.000000	0.723972	0.011108	0.899742	0.001842	0.241324
21088	159	give	361	0	0	give	VB	146	0.000596	52706	52706	0.133100	0.006186	60.000000	0.854806	0.014799	0.887929	0.000869	0.170861
44045	130	tell	418	0	0	tell	VB	126	0.000690	52668	52668	0.133004	0.007249	47.000000	0.960359	0.018880	0.887371	0.000915	0.384144
24082	887	husband	72	0	0	husband	NN	335	0.000119	24120	24120	0.060911	0.001550	17.000000	1.402008	0.051928	0.882779	0.000849	0.350502
50019	24450	youngster?	1	0	0	youngster?	NN	406	0.000002	406	406	0.001025	0.000052	1.000000	2.632457	0.877486	0.877486	0.877486	0.877486
16873	34582	encavemed	1	0	0	encavem	VBD	406	0.000002	406	406	0.001025	0.000052	1.000000	2.632457	0.877486	0.877486	0.877486	0.877486

Figure 3: TFIDF Matrix Heatmap of the Classical Texts.

The main takeaways from this heatmap are that the Epic Poetry category has high frequency of terms associated with “gods”, “fate”, and “heroes.” This makes sense for this genre as it is comprised of themes such as being propositioned by the gods, and the hero’s journey. The Drama category has a high frequency of terms like: “chorus”, “revenge”, and “women.” This coincides with the classics drama genre of the chorus announcing moments of tension and women being primarily associated with acts of revenge and drama in this time. Finally, the History category has moderate frequency usage of words like “hero”, “travel”, and “journey”. While these terms can be more ambiguous, it makes sense as this is a smaller category. Overall, the heatmap demonstrates the patterns per category parallel to their category themes.

Figure 4 gives insight into the LDA analysis of the theme clusters across the Greco-Roman corpus. There appears to be three key topic clusters in terms of the content groupings. First, the brown colored cluster grouping centers around daily life, with the words being related to domesticity. However, if one knows Odysseus from the Odyssey the terms of domesticity have a much darker context in relation to his retaliation towards the suitors pursuing his wife, Penelope. Second, the green and blue areas relate to the interactions of the gods and man. Finally, the orange section relates to mentions of war and strategy.

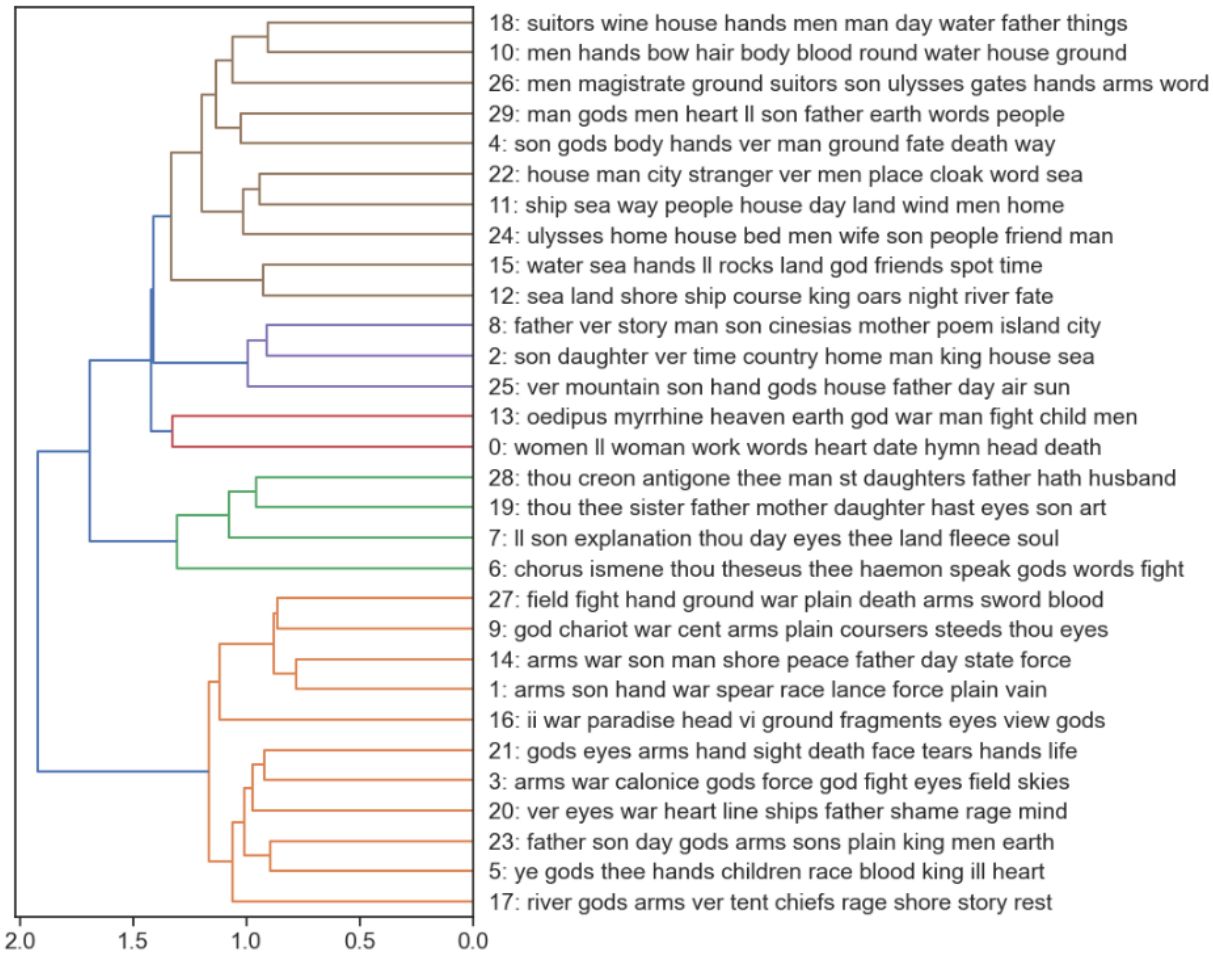


Figure 4: Hierarchical clustering dendrogram of the Classical texts.

Figure 5 shows the clustering of the principal components across the corpus. This visualization further demonstrates the thematic divide amongst the works. Texts like “The Iliad”, “The Odyssey”, and “The Aeneid”, which are all Epics, are divided from the Dramatic category works of “Medea” and “The works of Sophocles.” The others’ work has a slight overlap amongst the themes, while still preserving some of their independence. A good example of this is “The Argonautica”, as it is in between the separation of the Epic works and the Dramatic ones. This provides additional context to how there is thematic independence amongst some of the works but shared themes between others.

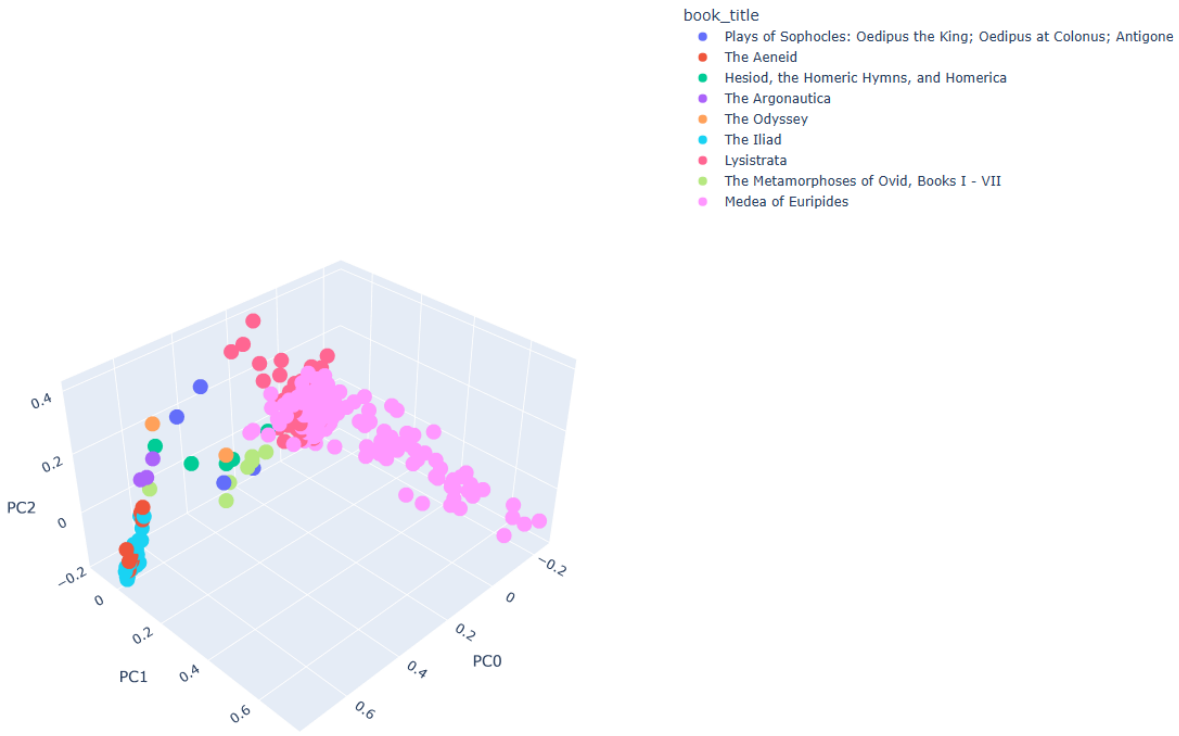


Figure 5: PCA plot of the clustering of principal components of the classical corpus, Three-dimensional format.

An overview of the word embedding semantics across the corpus text at each category level gives more insight into the patterns between the authors and category genre. Figure 6 of the Epic Poet category shows an apparent correlation between vocabulary that has to do with the divine, “gods”, and marital themes. This correlation is fitting as the works of Homer, Virgil, Ovid, and Apollonius rely heavily on narratives of the will of the gods and the effects it has on the mortals’ lives, specifically their marriages. The most notable example is Odysseus’s relationship with the will of the gods and his wife Penelope in the Odyssey and Iliad by Homer.

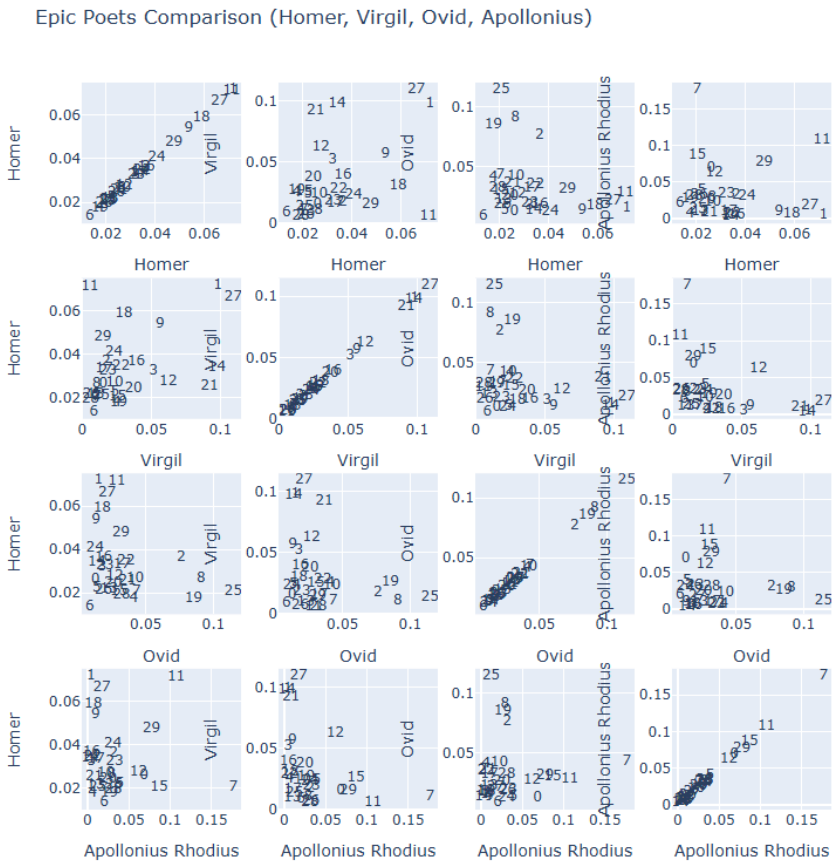


Figure 6: Word embedding pairwise comparison between Epic Poets.

Figure 7 shows the word embeddings comparison between the Dramatic Poets. The clustering behavior seen in the works of Euripides and Sophocles to the tragic vocabulary aligns with their works motifs of the darker tales of irony. The tale of Oedipus with its tragic ironic fate of love and lust is one example. The work of Aristophanes leans away from the tragic vocabulary into the extreme of comic humor. This behavior in the Dramatic Poets category is rational because Drama encompasses aspects of tragedy, irony, humor, and comedy. Finally, Hesiod's work moves amongst the scales of tragedy and comedy, as this text is that of a more historical perspective of the period.

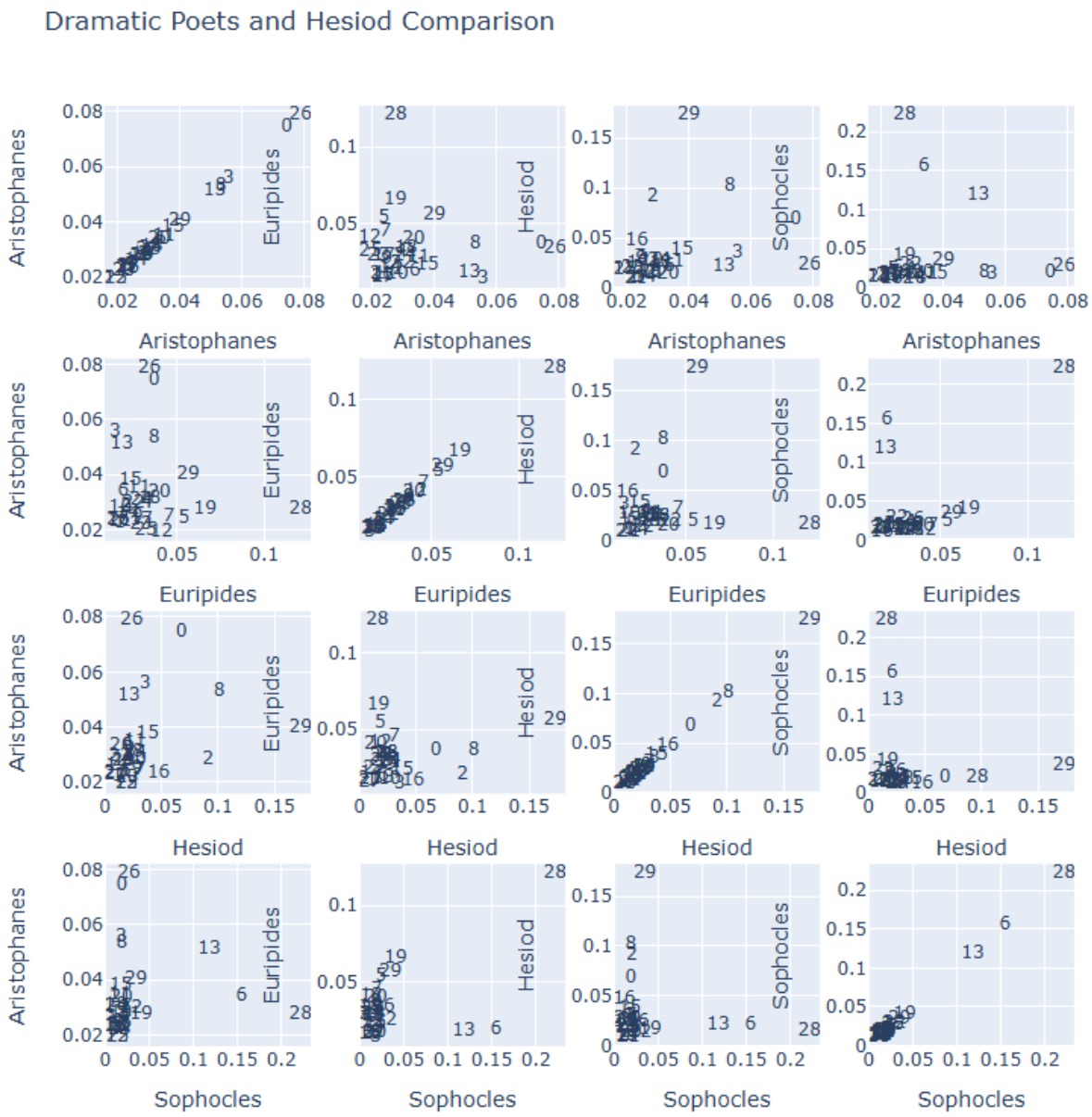


Figure 7: Word embedding pairwise comparison between Dramatic Poets.

Figure 8 is a visual representation of the word embedding pairwise comparison between Epic and Dramatic Poets. This graph shows distinct separation between the Epic and Dramatic corpus texts vocabularies. This visualization also highlights the shared vocabulary related to language of mythos, like “divine”, “nymph,” “chorus,” etc.

Figure 8: Word embedding pairwise comparison Epic vs Dramatic Poets.

With the word embedding relationships established, analysis of emotional patterns across the corpus through sentiment analysis provides additional context for understanding emotional ties to the vocabulary usage within each category. First, the VADER sentiment analysis of Homer's Iliad from the Epic category in Figure 9 shows the behavior of emotional progression throughout the work. The blue line represents the positive sentiment, and the orange line represents the negative sentiment. Overall, the positive sentiment dominates the beginning and some of the middle of the corpus, then there is a stark decline in the positive sentiment towards the end of the story. This behavior makes sense as Epic tales most often have tragic endings to round out the hero's journey, in this case with the brutal and bloody death of the strongest Trojan hero, Hector.

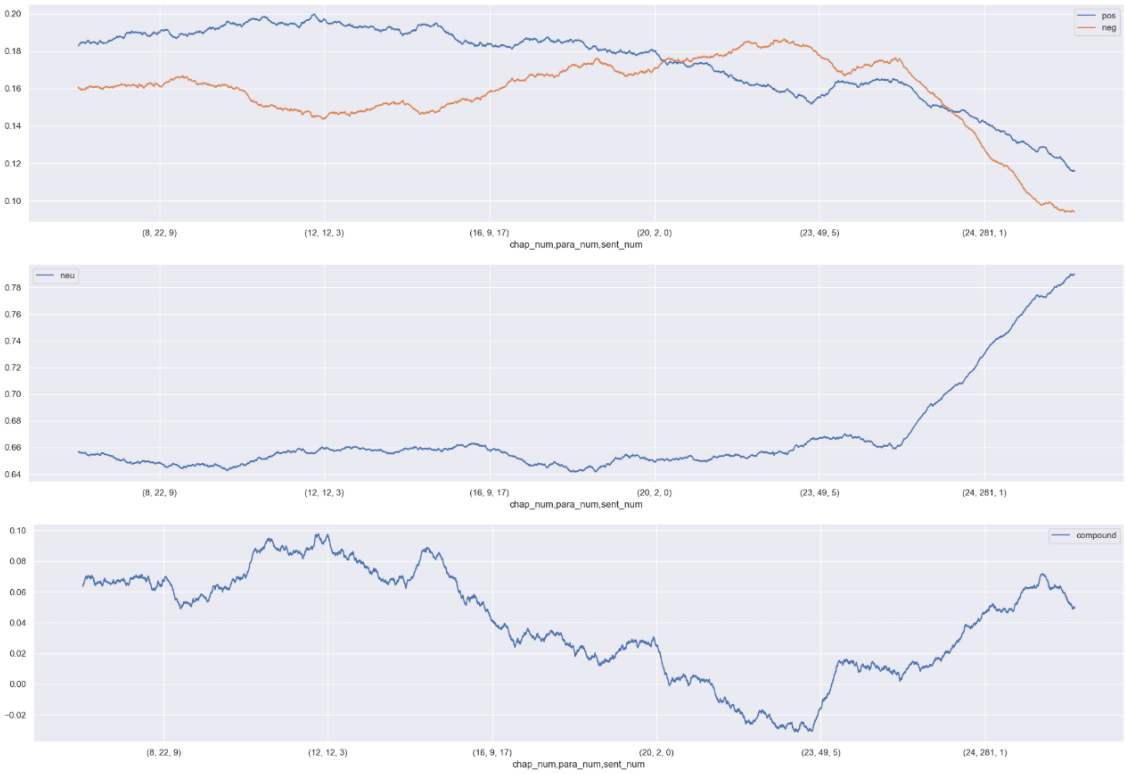


Figure 9: VADER sentiment analysis of Homer's Iliad.

Figure 10 is the Vader sentiment analysis of Euripides's Medea, part of the Dramatic Poetry category. The behavior of the positive and negative sentiment lines contrasts with the consistent progression of that in Figure 9. The positive sentiment in the story of Medea begins at a low point while the negative sentiment is at a high. From left to right there are two instances of positive and negative crossover and throughout the middle arch of the work the emotional theme crosses into positivity. Finally, the story closes with the second crossover back into negative sentiment. This behavior is consistent with Dramatic works as they start in dreary, bleak situations that the characters must preserver through. As the characters seemingly succeed in beating out the odds, there is a stark turn that befalls them, and the story ends in tragedy. The large separation in crossover points from negative to positive and the initial sentiment that starts the works are the key points of contrast between Epic and Dramatic Poetry works.

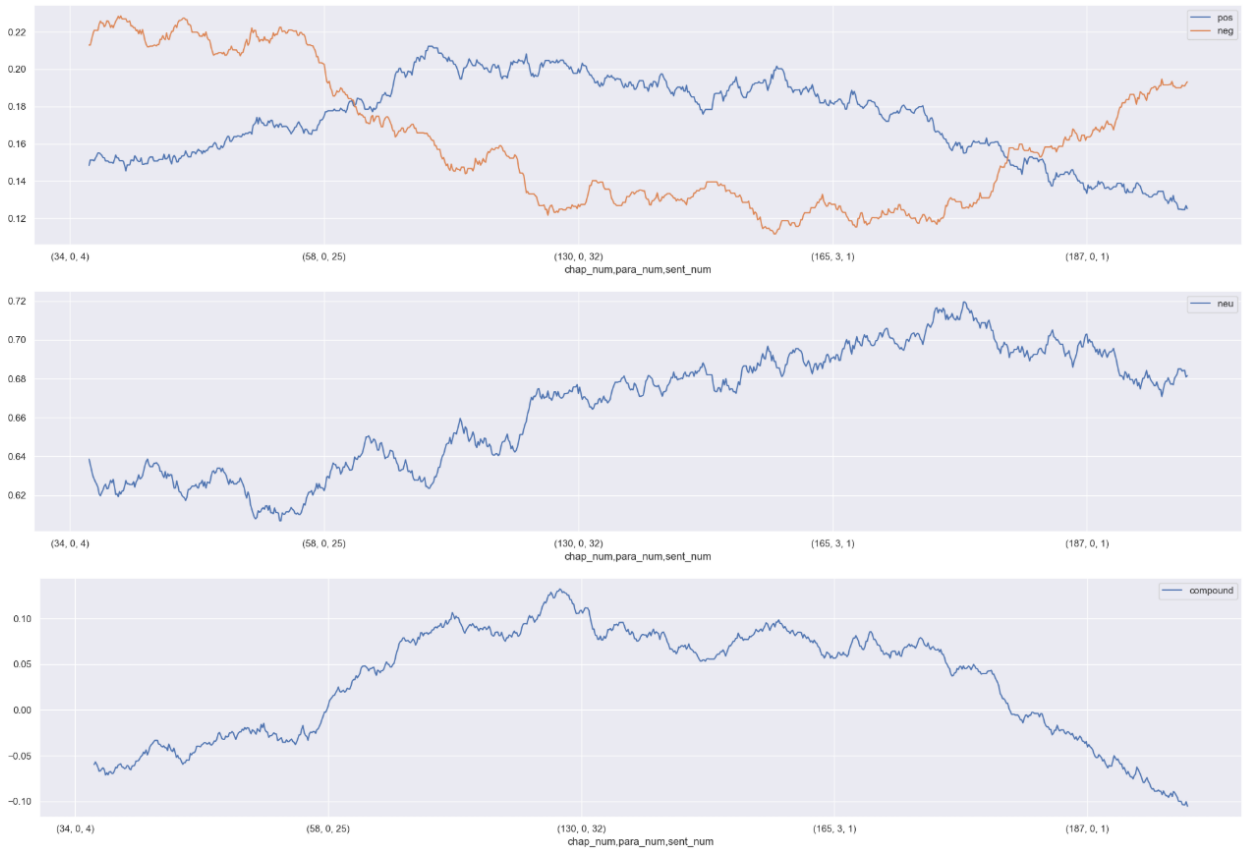


Figure 10: VADER sentiment analysis of Euripides' Medea.

Figure 11 is the Vader sentiment analysis of Hesiod's Theogony. The inclusion of this graph is to round out the category comparisons, as the emotional patterns of historical based texts are quite different than those of the classical Greco-Roman Epics or Dramatic Poetry. The vertical points of separation between the positive and negative sentiment lines are quite small in comparison to that of Figures 9 and 10. Additionally, the behavior of the sentiment lines here is much smoother when compared with the Figures 9 and 10, whose lines appear jagged from the incremental changes of highs and lows from left to right. This is likely because the mode of text expression is that of a historical narrative and not a dramatic one.

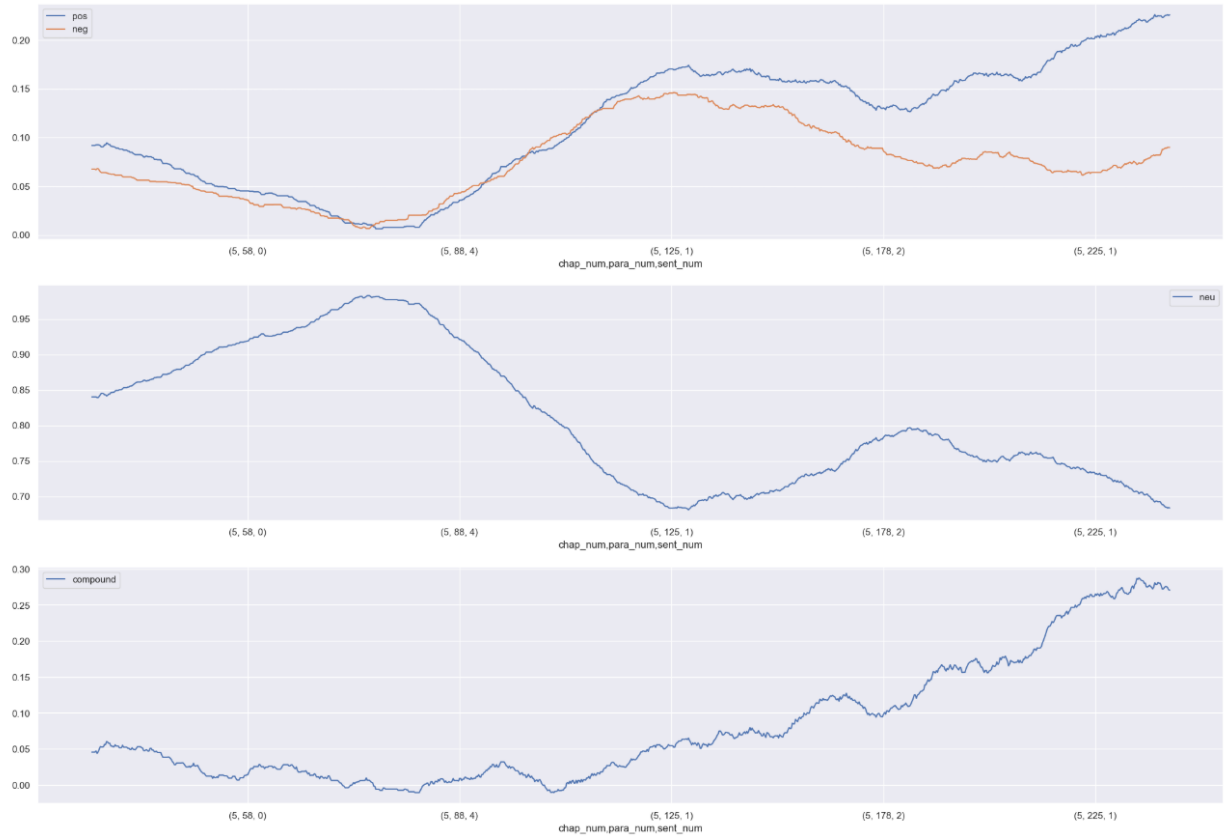


Figure 11: VADER sentiment analysis of the Hesiod's Theogony.

Overall, the contrasting patterns in the VADER graphs of positive and negative sentiment lines in the sentiment analysis provide quantitative evidence for distinguishing works into categories such as Epic, Dramatic Poetry, and History. These patterns highlight how distinct narrative purposes influence the shape and behavior of sentiment trajectories. For example, Epic works often align with themes of a hero's journey and feature steady emotional arcs. Dramatic Poetry exhibits sharp fluctuations reflecting rapid change in plot developments to evoke tension, and Historical works maintain consistent tones. These variations demonstrate how emotive patterns differ across categories.

Conclusion:

The analysis of the nine classical Greo-Roman corpus demonstrated patterns across the classic works categories of Epic, Dramatic Poetry, and Historical texts throughout varied ETA methods of analysis. The findings from the TFIDF analysis revealed the initial thematic/genre division of the sets of works in the corpus into the stated categories above. Where Epic works were strongly related to the vocabulary frequency of terms such as "gods", "fate", and "heroes", Dramatic Poetry had terms like "chorus", "revenge", and "women" and historical text had more generic term usages. The thematic divisions were further supported with the LDA topic modeling and PCA analysis, which displayed distinct genre clustering and notable overlaps amongst

categories. For example, Apollonius's "The Argonautica," had genre elements of both the Epic and Dramatic Poetry category, which demonstrated that certain works could encompass elements of multiple genres. Lastly, the hierarchical clustering analysis brought another angle of support to the thematic division as there appeared to be three distinct clusters. The three clusters were comprised of vocabulary related to domestic life, interactions with the gods, and warfare, ultimately showing how these 3 clusters contain elements that are present across the corpus.

Furthermore, the sentiment analysis of the works revealed distinct emotive patterns related to each category/thematic genre. Epic works showed a behavior of an initial positive sentiment progression followed by a decline, alluding to the tragic ends of Epic works. Dramatic works, like Euripides Medea, exhibited sporadic changes in sentiment behavior, jagged lines, and several crossovers from positive to negative and negative to positive sentiments. Historical texts maintained a consistent emotional tone, as seen in the line behavior of the positive and negative sentiment, which were near smooth and consistent.

Finally, examining all the various analyses of the corpus gives meaningful evidence to support that classical works implement distinct narrative story telling strategies and emotional patterns that explore classical themes. These themes include interactions between mortals and gods, the hero's journey, and characters befalling tragedy. The analyses of these thematic patterns further suggest that classical works have distinct literary device applications that separate classical genres from one another. Further research with a larger corpus could reveal even more nuances to the patterns exhibited in classical works.

Deliverables Reference Guide:

All work is in a zip file called: final_project

All Deliverables are in a folder called: Project_Deliverables, but all work is in the above file location for ALL output files and tables

1. Collection of Source files called and located in: Source_Files located in Project_Deliverables
2. A manifest File Describing the source files: The top of the Jupyter Notebook contains the manfiest information, heading called: Manfiest Note
3. Collection of Data files called and located in: Data_Files located in Project_Deliverables
4. PCA Tables outputs called and located in: PCA_Outputs located in Project_Deliverables
5. LDA, word2vec, Sentiment Analysis: Located in the Final_Project.ipynb
6. Jupyter notebook that explores:
    a. Hierarchical cluster diagrams
    b. Heatmaps Showing correlations
    c. Scatterplots
    d. KDA plots
    e. Dispersion plots
    f. T-SNE plots
        i. All are in the Final Project Jupyter Notebook called Final_Project.ipynb with write up of my findings.
7. Two-to-Four-page Final Report, this document.

Appendix

[i] https://www.gutenberg.org/