Pavan Kumar Bondalapati
(pb7ak@virginia.edu)
DS 5001
May 11, 2021

Exploratory Text Analysis on Ovid's *Metamorphoses*

Overview

Pūblius Ovidius Nāsō (43 BCE to 18 CE), known as Ovid, was a prolific Roman poet during the fall of the Roman Republic and the subsequent rise of the Roman Empire. Ovid is greatly regarded by many historians for his *magnum opus*, the *Metamorphoses*. Spanning across 15 books, which together contain over 250 myths, this poem accounts the historical events from the creation of the world to deification of Julius Caesar. Given the chronological nature of *Metamorphoses*, there appear to be no clear divisions in this poem; nonetheless, American scholar Brooks Otis (1908-1977) proposed four arguable divisions. Books I and II cover "The Divine Comedy." Books III through VI are centered around "The Avenging Gods." Books VII through XI focus on "The Pathos of Love." The topic of the remaining books underscore "Rome and the Deified Ruler." The goal of this project is to extend the tools of exploratory text analysis to Ovid's *Metamorphoses*. In particular, I will employ clustering on each of the 15 books in this poem to potentially identify the latent chronology or provide evidence for Otis' four divisions. Additionally, I will apply topic modeling on these four divisions as a means to verify whether the narrative descriptions provided by Otis are accurate.

Conversion

All 15 books in the *Metamorphoses* are publicly available on Project Gutenberg (eBook No. for Books I-VII and VIII-XV are 21765 and 26073, respectively) in plain text UTF-8 encoding. This version of the poem is translated into English by Henry Thomas Riley (1816-1878).

Figure 1: First Five Rows of the Library Table

book_num	book_title	book_id	book_file	author	title
1	The Metamorphoses of Ovid, by Publius Ovidius ...	21765	./data_in/21765-0.txt	Publius Ovidius Naso	The Metamorphoses of Ovid
2	The Metamorphoses of Ovid, by Publius Ovidius ...	21765	./data_in/21765-0.txt	Publius Ovidius Naso	The Metamorphoses of Ovid
3	The Metamorphoses of Ovid, by Publius Ovidius ...	21765	./data_in/21765-0.txt	Publius Ovidius Naso	The Metamorphoses of Ovid
4	The Metamorphoses of Ovid, by Publius Ovidius ...	21765	./data_in/21765-0.txt	Publius Ovidius Naso	The Metamorphoses of Ovid
5	The Metamorphoses of Ovid, by Publius Ovidius ...	21765	./data_in/21765-0.txt	Publius Ovidius Naso	The Metamorphoses of Ovid

As shown in “1-Conversion.ipynb,” these plain text files are then converted into a singular library, vocabulary, and token tables. Slightly deviating from the norm, the library table is indexed by “book_num” rather than “book_id,” since it serves as the more informative primary key. In addition to the actual translated poem, this version contains detailed footnotes and a brief explanation at the end of each fable. Aside from the referencing notations, all of this additional information is accounted in the vocabulary and token tables. Shown below are the labeled vocabulary and token tables.

Figure 2: Overview of the Vocabulary Table

term_id	term_str	n	num	stop	stem_porter	stem_snowball	stem_lancaster	pos_max
0		31063	0	0				,
1	1	2	1	0	1	1	1	CD
2	1000	1	1	0	1000	1000	1000	CD
3	109	1	1	0	109	109	109	CD
4	11	3	1	0	11	11	11	CD
...	...	...	...	...	...	...	...	...
14907	ὥπς	1	0	0	ὥπς	ὥπς	ὥπς	NNP
14908	ὦ	1	0	0	ὦ	ὦ	ὦ	NNP
14909	ὦτα	1	0	0	ὦτα	ὦτα	ὦτα	NNP
14910	ὦτον	1	0	0	ὦτον	ὦτον	ὦτον	NN
14911	ῥαίνω	1	0	0	ῥαίνω	ῥαίνω	ῥαίνω	NNP

14912 rows × 8 columns

Figure 3: Overview of the Token Table

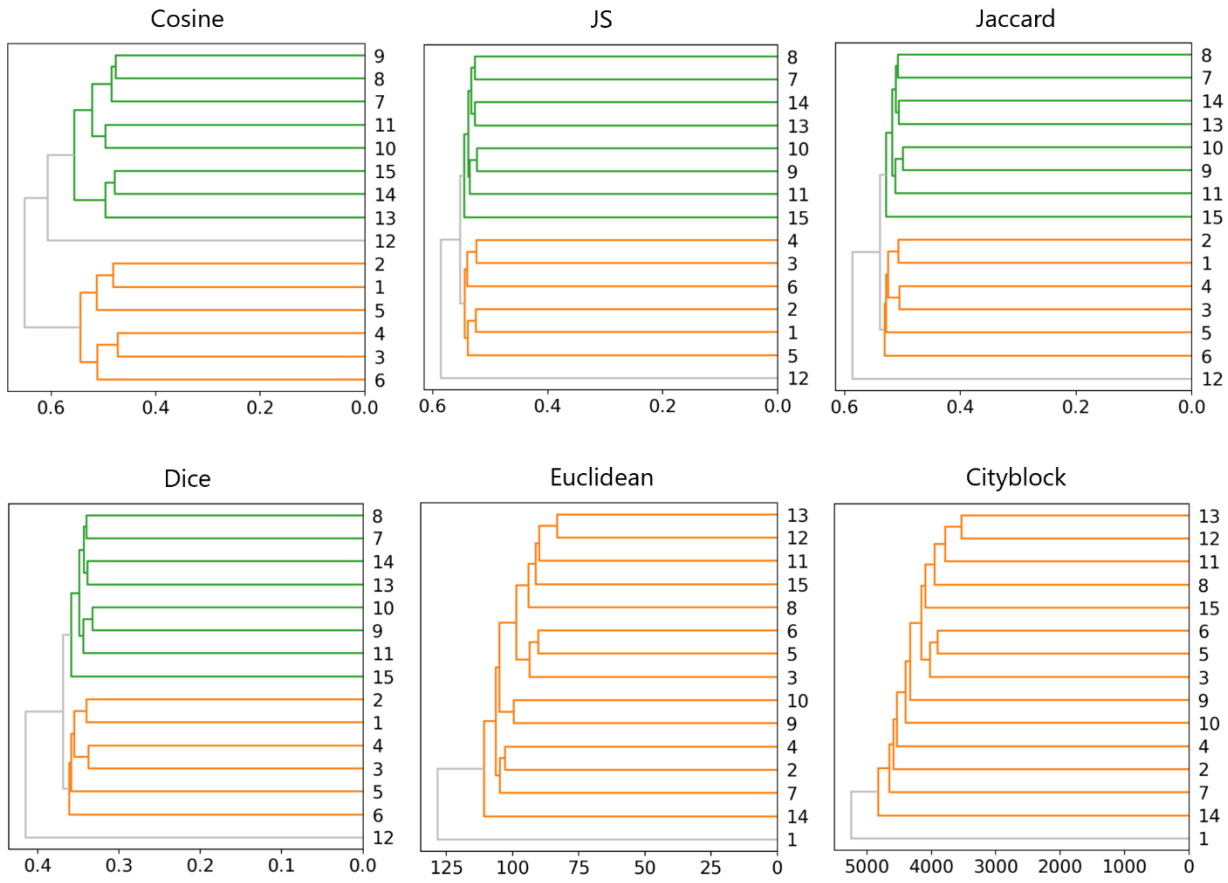
book_num	fable_num	para_num	sent_num	token_num	pos_tuple	pos	token_str	term_str
1	**1**	**0**	**0**	**0**	('God', 'NNP')	NNP	God	god
				1	('reduces', 'VBZ')	VBZ	reduces	reduces
				2	('Chaos', 'NNP')	NNP	Chaos	chaos
				3	('into', 'IN')	IN	into	into
				4	('order', 'NN')	NN	order	order
...	**...**	**...**	**...**	**...**	...	...	...	...
15	**5**	**21**	**9**	**19**	('at', 'IN')	IN	at	at
				20	('what', 'WP')	WP	what	what
				21	('time', 'NN')	NN	time	time
		22	**0**	**0**	('THE', 'DT')	DT	THE	the
				1	('END', 'NN')	NN	END	end

276849 rows × 4 columns

Clustering and Similarity

Using fables as bags and "max" as the term-frequency method, the generated TFIDF table was subsequently collapsed by the books. After reducing the features to 4,000 terms, the relative distances between the 15 books were computed using the following linkage methods: Cosine, Jensen Shannon, Jaccard, Dice, Euclidean, and Cityblock. From Figure 4, the most informative distance measure appears to be the Cosine measure since this metric best captures the chronological nature of the *Metamorphoses* as well as Otis' four divisions (i.e. Books I-II, Books III-VI, Books VII-XI, and Books XII-XV).

Figure 4: Dendrograms by Linkage Method

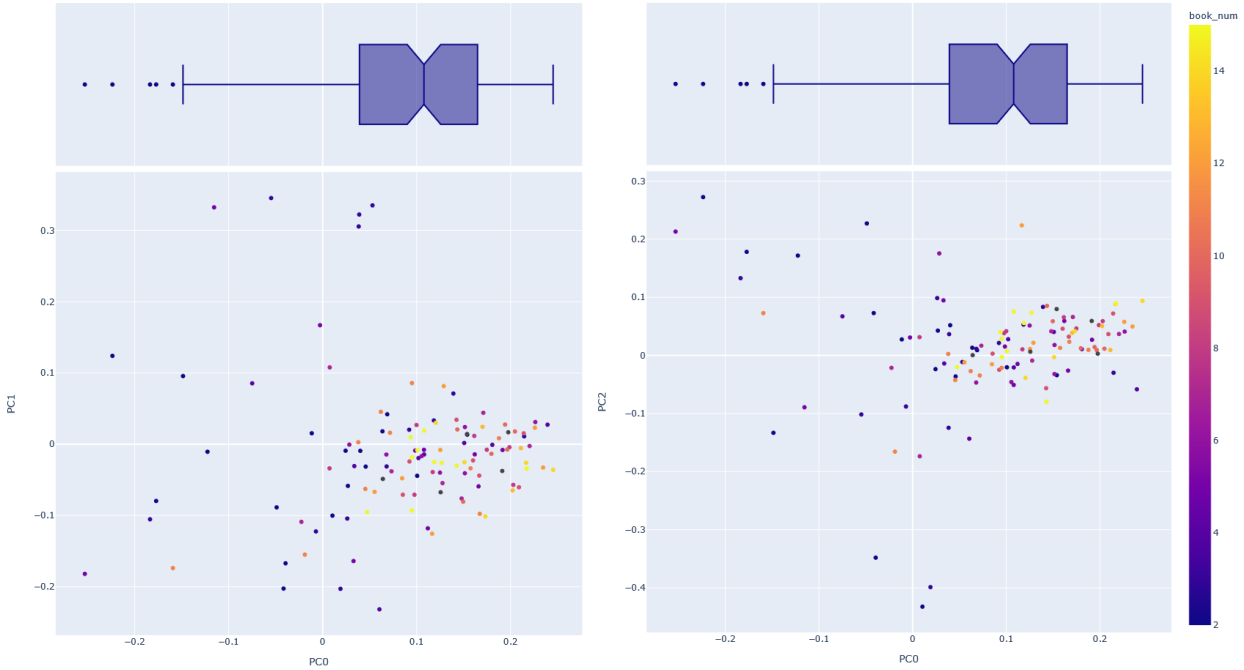


Principal Component Analysis (PCA)

In order to better characterize the fables in the *Metamorphoses*, PCA could elucidate the hidden composition behind these fables. Similar to the clustering section, the TFIDF table was generated using fables as bags and "max" as the term-frequency method. After decomposing the top 4,000 significant terms from the TFIDF table, the

following two figures compare the first and second as well as first and third dimensions of PCA, respectively. Although it is unclear what exactly is captured from these three dimensions, the first dimension appears to marginally capture the temporal order of the books in *Metamorphoses*.

Figure 5: PCA 0 vs. 1 and 0 vs. 2



Topic Modeling

Latent Dirichlet Allocation (LDA) was utilized to model the topics for each of the four divisions described by Otis. The narrative, "The Divine Comedy," was depicted as "story, time, son, daughter, writers, history, tradition, father, man, king." Likewise, the narratives "The Avenging Gods" and "The Pathos of Love" are both portrayed by "story, fact, town, writers, names, daughter, women, poets, death, mother." Finally, "Rome and the Deified Ruler," is described by "island, people, temple, city, honor, death, festival, promontory, altars, Aeneas." Given these generated topics, LDA appears to fairly capture the narratives of "The Divine Comedy" and "Rome and the Deified Ruler." The generated topics for the remaining two narratives seem to be rather broad.

Figure 6: Topic Modeling for “The Divine Comedy”

narrative topic_id	The Divine Comedy	label
31	0.061704	31 story, time, son, daughter, writers, histor...
26	0.061236	26 river, stream, waters, banks, course, noise...
38	0.052203	38 explanation, number, earth, place, water, d...
28	0.042508	28 sun, earth, heat, air, things, weight, circ...
19	0.037464	19 mountain, son, thrace, form, gold, things, ...

Figure 7: Topic Modeling for “The Avenging Gods”

narrative topic_id	The Avenging Gods	label
5	0.045470	5 story, fact, town, writers, names, daughter,...
18	0.041002	18 word, city, country, son, story, oracle, wr...
23	0.040180	23 mother, daughter, sister, father, thou, wif...
38	0.038035	38 explanation, number, earth, place, water, d...
31	0.036340	31 story, time, son, daughter, writers, histor...

Figure 8: Topic Modeling for “The Pathos of Love”

narrative topic_id	The Pathos of Love	label
5	0.046227	5 story, fact, town, writers, names, daughter,...
32	0.043488	32 son, death, father, wife, mother, hand, arm...
2	0.041072	2 sea, waves, ship, waters, time, fountain, ci...
38	0.040839	38 explanation, number, earth, place, water, d...
27	0.036916	27 island, people, temple, city, honour, death...

Figure 9: Topic Modeling for “Rome and the Deified Ruler”

narrative topic_id	Rome and the Deified Ruler	label
27	0.049446	27 island, people, temple, city, honour, death...
31	0.041895	31 story, time, son, daughter, writers, histor...
18	0.038158	18 word, city, country, son, story, oracle, wr...
38	0.037264	38 explanation, number, earth, place, water, d...
30	0.036633	30 son, daughter, blood, flower, wife, death, ...

Word Embedding

Another layer of understanding for the full story inscribed in the *Metamorphoses* can be achieved by word embedding. Here, using paragraphs as the bag, we can generate a "word2vec" model. The clustering of the related words can be visualized by a t-distributed stochastic neighbor embedding (t-SNE) plot. The t-SNE of the word embeddings for all books is shown in Figure 10. At the narrative or fable level, there appears to be not enough words for coherent clustering. As a result, the analogies shown in Figures 11-14 are drawn from a corpus which utilizes all 15 books.

Figure 10: t-SNE of Word Embeddings

Figure 11: Analogy 1 – Man : Son :: Woman : ?

	term	sim
0	daughter	0.954267
1	king	0.875437
2	who	0.846047
3	wife	0.830755
4	mother	0.807898

Figure 12: Analogy 2 – Hero : Strength :: Monster : ?

	term	sim
0	escaped	0.996159
1	remained	0.995932
2	removed	0.995703
3	reached	0.994786
4	weapon	0.994709

Figure 13: Analogy 3 – Man : Strength :: Woman : ?

	term	sim
0	care	0.989563
1	mind	0.989348
2	wretched	0.987711
3	hardly	0.987544
4	remains	0.987483

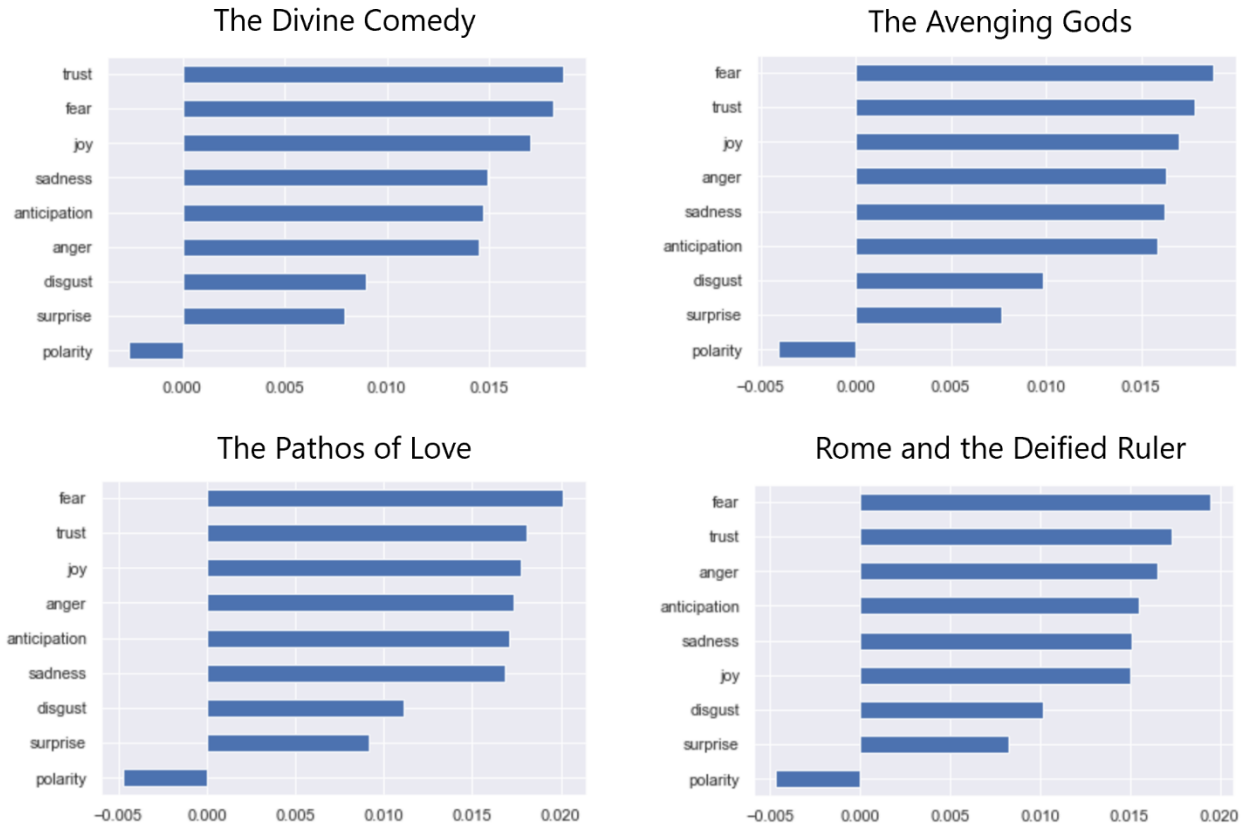
Figure 14: Analogy 4 – Hero : Monster :: Man : ?

	term	sim
0	foundation	0.986603
1	female	0.985043
2	whence	0.984021
3	told	0.982810
4	kind	0.981856

Sentiment Analysis

Finally, the analysis of sentiment variation among the four divisions proposed by Otis could provide insight on the validity of these divisions. Ideally, we would expect each of the four narratives to have a unique variation of sentiment, which provides justification for the division. The sentiment by the four narratives are shown in Figure 15. Although the four narratives share the general distribution of emotions, "The Divine Comedy" and "Rome and the Deified Ruler" as well as "The Avenging Gods" and "The Pathos of Love" share similar sentiments. In summary, the sentiment analysis does not appear to support the divisions proposed by Otis.

Figure 15: Sentiment by Narrative



References

https://en.wikipedia.org/wiki/Ovid (Overview)
https://en.wikipedia.org/wiki/Metamorphoses (Overview)
https://www.gutenberg.org/ebooks/21765 (Conversion)
https://www.gutenberg.org/ebooks/26073 (Conversion)