Thomas McIntyre Gem5cm@virginia.edu

DS5001

08 May 2022

DS5001 Final Project: Game of Thrones

Motivation

Game of Thrones successfully captivated the worlds attention for many years in the 2010s, and once I began watching the series I quickly realized why. I have a knack for being able to predict what is going to happen in plots of many shows and texts, however Game of Thrones was different. I was always kept on my toes and had never really watched anything quite like it. It led me to read some of the books that the HBO series is based on. On the contrary, I remember being forced to read several classics in middle and high school that I could barely get through to pass my English class quizzes. The one that stood out to me as the worst was Jane Eyre. Thus, about halfway through this class I thought of the idea of utilizing these tools to compare some of the Game of Thrones, Jane Eyre, and other classic books to see if there are any underlying themes that may link a story I thoroughly enjoyed, to a tale you couldn't pay me to pick up again. This paper will first describe the corpus and then outline the processes and the findings contained in each of the 7 Jupyter Notebooks provided.

Corpus

The corpus of interest contains 12 different novels. The genres range from fantasy, children's fantasy, crime, gothic, and historical. The fantasy novels are the first three Game of Thrones books and the first Lord of the Rings Book. The two children's fantasy stories are Peter Pan and Peter and Wendy, which are essentially the same novel except the prior was based purely on the play (this will be useful to reconcile our steps to see if the outputs are very similar for these two). There are two Charles Dicken's books, Oliver's Twist and Great Expectations which are tagged as crime and gothic respectively. The two historical books are Homer's The Odyssey and The Iliad, which will be interesting to see how these compare to some of the fantasy adventure books mentioned earlier. The final two books were both tagged as gothic and are Bronte's Jane Eyre and Wuthering Heights. Based on intuition, I would hypothesize that the books that will relate the most to Game of Thrones will be Lord of the Rings, Homer's books, Barrie's books, Dicken's books, and Bronte's books in that order. Please refer to the LIB table in the Appendix for more information on the corpus of interest, and links to the novels.

Notebook 1 Creating LIB and CORPUS/TOKENS Table from Raw Text Files

In the first notebook, I set the OHCO structure to the following levels: book, chapter, paragraph, sentence, and token. One assumption made in the parsing was that a book will be defined as all contents within the covers of the hard copy of the book. For example, The Fellowship of the Ring has two sub-books within the hardcover of the novel, but to keep consistency across the corpus we defined this as only one book and did not reset the chapter count at the start of the second sub book. Another distinction worth noting, was that the Game of Thrones chapters will be defined whenever the story changed point of view, this could easily be parsed by at least three capital letters in a row. For example, "BRAN", "JON", or "NED" in the texts symbolize a change in point of view and in our OHCO was where the chapters were parsed. The main

parts of this notebook were for creating a dictionary to define the start lines, end lines, and chapter RegExs to be passed into the acquire_epubs function within methods.py. This function produced the LIB and DOCS tables. After appending some columns of interest to the LIB table, I then passed the DOCS table into the tokenize function within methods.py to produce our CORPUS table. The two functions mentioned utilize NLTK and other text parsing techniques, please refer to acquire_epubs and tokenize within methods.py for more information. The last piece of this notebook was writing the CORPUS and LIB table to our data folder.

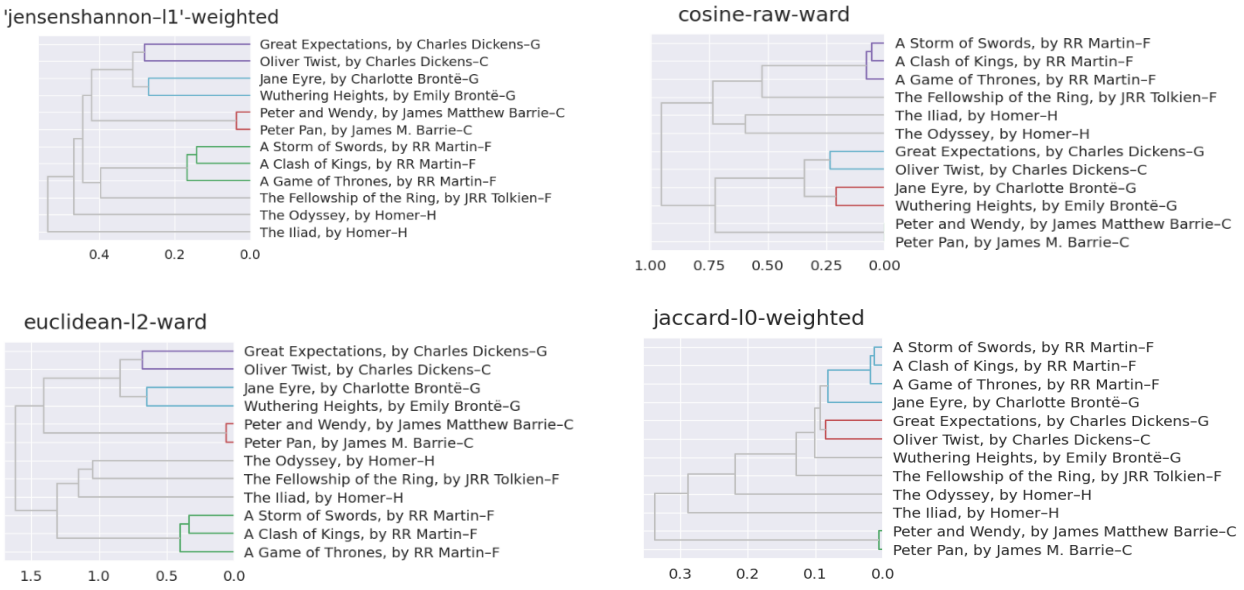
Notebook 2 Using SkLearn to Create VOCAB, DOC, DTM, TFIDF, BOW

This notebook utilized SkLearn's CountVectorizer, TfidfVectorizer, TfidfTransformer and NLTK to create our VOCAB, DOC, DTM, TFIDF, and BOW tables from the two tables created in step 1. The first step utilized gather_docs from methods.py to create the DOC table. The next step retained 5000 terms when using SkLearn's CountVectorizer to create the DTM. The TFIDF table was created using the DTM and SkLearn's TfidfTransformer. Following this step, the VOCAB table was created using the DTM table and appended the following columns tfidf_mean, df, dfidf, p, i, max_pos, n_pos, and stop. The last table created was a BOW at the chapter level and contains a count and tfidf column.

Notebook 3 Distance Measures and Dendrogram Visualizations

In notebook 3 I began to look at different distance measurements to analyze the similarities across all combinations of the 12 books. I created a table called PAIRS, which had 6 different distance measures for each combination. Using the information from the PAIRS table, I plotted 6 different dendrograms to visualize these findings.

Figure 1: Dendrograms with Different Distance Measures and Linkage Methods



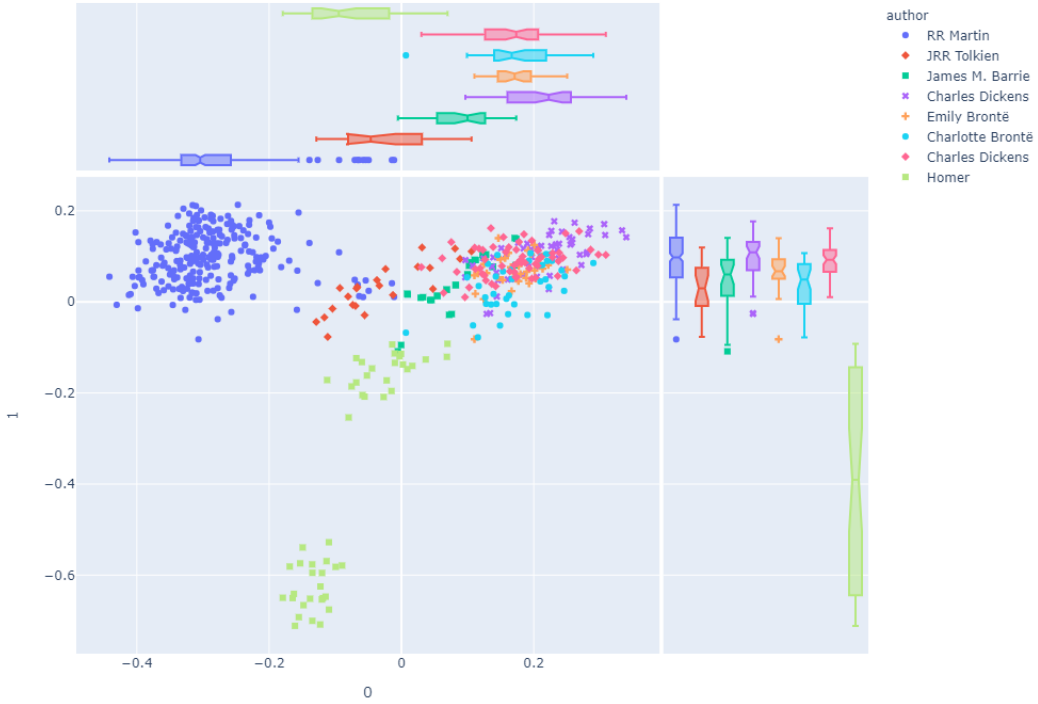
Viewing the dendrograms above, it is apparent that the different distance measures and linkage methods have very similar outcomes. The Game of Thrones books are all grouped with one another in each version, and often are very close to Lord of the Rings and Homer's novels. Interestingly, in the top right dendrogram (cosine-raw-ward), the two Barrie books are not even separated, which makes sense because they are

essentially the same text. The bottom right plot (jaccard-l0-weighted) shows Jane Eyre grouped with Game of Thrones. This is the only dendrogram I investigated that showed this, however this may suggest that there may be some underlying similarities between the two. When reviewing the PAIRS table, the Game of Thrones books are primarily far away from Jane Eyre across the board. Refer to notebook DS5001_STEP_3_Dendos.ipynb for more information.

Notebook 4 Principal Component Analysis

In this notebook we conducted principal component analysis using our TFIDF table at chapter level. The TFIDF table was subset by the top 2000 words from the VOCAB table sorted on DFIDF and of the following parts of speech from the Penn Treebank P.O.S. Tags: "NN","NNS", "VB", "VBD", "VBG", "VBN", "VBP", "VBZ", "JJ", "JJR", "JJS", "RB", "RBR", "RBS". After trimming the TFIDF to the filters above, the get_pca() function from methods.py was used to produce the LOADINGS, DCM, and COMPINF tables for the first 10 components. The following plot was created using the DCM joined onto the DOC table (more component plots can be examined in the DS5001_FINAL_STEP_4_PCA.ipynb notebook)

Figure 2: Scatterplots of First and Second Principal Components by Author



In the scatterplot above the X-axis is the first principal component and the Y-Axis is the second. When viewing the first component it was apparent that Game of Thrones (RR Martin) is distinctly far away from the others, while Homer's novels and Lord of the Rings fall in the middle, Barrie's stories are a bit behind those two, and lastly Dicken's and Bronte's are well on the positive end of this component. This is a very similar picture shown in the dendrograms and falls in line with our original hypothesis. It was also worth

noting that Homer's, The Iliad, drastically stands out in the second principal component. The top 5 words for each pole of the principal components can be seen below in Table 1.

Table 1: Principal Component 1 and 2

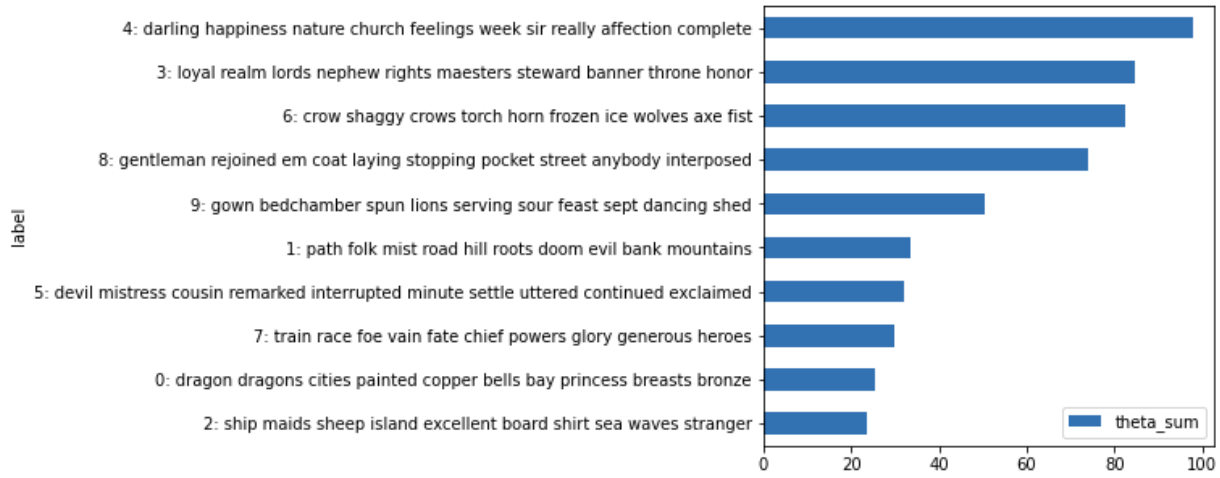
pc_id	pos	neg	eig_val	exp_var
0	sir gentleman dear replied room	king men sword brother knight	0.050579	0.291869
1	lady boy girl door castle	heaven chief god war vain	0.026610	0.153558

The negative pole of component 1 clearly looks tied into Game of Thrones and Lord of the Ring style books, while the very proper pronouns of the positive pole make sense with Dicken's and Bronte's novels. The Iliad must have strong religious ties within the story because the words heaven, vain, and God show up near the top on the negative end of component 2. Other components and tables can be seen in the notebook.

Notebook 5 Topic Models

This notebook heavily used the topicmodel.py script found in the code folder. For the topic modeling, I utilized the same BOW structure as in the PCA notebook (book, chapter as index), and I reduced the table to the top 2000 terms in the same manner. I decided to generate 10 topics to see if any of the Game of Thrones books seem to share topics with Jane Eyre. The topics can be seen on the bar chart below showing their theta sum.

Figure 3: Theta Sum of the 10 Topics



In Figure 3, we can see the top words that belong to each topic as well as their theta sums. On first glance it could be assumed that Topics 3 and 0 probably relate most to Game of Thrones, while Topic 4 would most likely be tied more closely to Bronte's and Dicken's work. To examine the topics futher, I produced the table below showing the means for each topic from our THETA table by author.

Table 2: Topics by Author

author	0	1	2	3	4	5	6	7	8	9
Charles Dickens	0.002800	0.033934	0.003457	0.004505	0.194960	0.064570	0.012596	0.011092	0.669085	0.003002
Charles Dickens	0.001992	0.035530	0.014412	0.008097	0.376252	0.010574	0.021429	0.000806	0.507254	0.023653
Charlotte Brontë	0.003921	0.070976	0.006616	0.010970	0.727808	0.098067	0.007662	0.032265	0.036124	0.005590
Emily Brontë	0.003861	0.029891	0.007618	0.003890	0.210264	0.666408	0.006814	0.008521	0.056340	0.006394
Homer	0.005023	0.015167	0.381482	0.009298	0.036095	0.000868	0.004920	0.510420	0.032835	0.003891
JRR Tolkien	0.004264	0.793100	0.006570	0.021593	0.068601	0.025450	0.022661	0.010761	0.033679	0.013321
James M. Barrie	0.000217	0.059004	0.050472	0.000217	0.666798	0.000217	0.114558	0.023476	0.048622	0.036421
RR Martin	0.100709	0.022533	0.006957	0.338433	0.015326	0.003368	0.308938	0.008190	0.004777	0.190769

In Table 2, it looks as though RR Martin does not have any topics that show up in other authors works. Topic 6, which describes things and animals found in an old village appear to have some small connection with Barrie's work. An interesting finding from this table is that all of our novels are pretty different in style and not many topics are shared across novels, except for Topic 4 which shows up in Bronte's, Dicken's and Barrie's work. This lens of our corpus follows a similar story to as what has been shown so far through the dendrograms, PCA plots, and Topics table. For more information on the topics, please visit the notebook.

Notebook 6 Word Embedding

This notebook utilizes gensim's "word2vec" and sklearn's "TNSE". The TOKENS table had all of the proper nouns removed before passing it onto the functions to produce W2V and COORDS tables. These two tables were created for the whole dataset, RR Martin only, and then everything except RR Martin. Using the word2vec models created for each grouping, I investigated the most similar words to see differences from RR Martin writing and the rest of the corpus. One word of interest I had was "brother" since it showed up as one of the most important words in component 1 investigated earlier.

Table 3: Words Similar to Brother by Subset

Corpus

	term	sim
0	sister	0.914136
1	uncle	0.901152
2	mother	0.875830
3	cousin	0.862297
4	nephew	0.838934
5	husband	0.836720

RR Martin

	term	sim
0	sister	0.936290
1	uncle	0.935960
2	mother	0.904875
3	son	0.899289
4	cousin	0.870137
5	daughter	0.851225

Rest

	term	sim
0	husband	0.922704
1	conduct	0.906784
2	companion	0.893252
3	soul	0.886500
4	cousin	0.886357
5	uncle	0.880395

In Table 3, the most similar words to brother in the corpus and RR Martin are very similar, however when comparing the rest of the corpus to RR Martin there is a distinction in which words are detected as most

similar. More word comparisons and analogy tools can be found in notebook 6. This notebook also contains t-SNE plots for the word embeddings created. The plot for the whole corpus can be seen below in Figure 4.

Figure 4: t-SNE Plot for the Whole Corpus

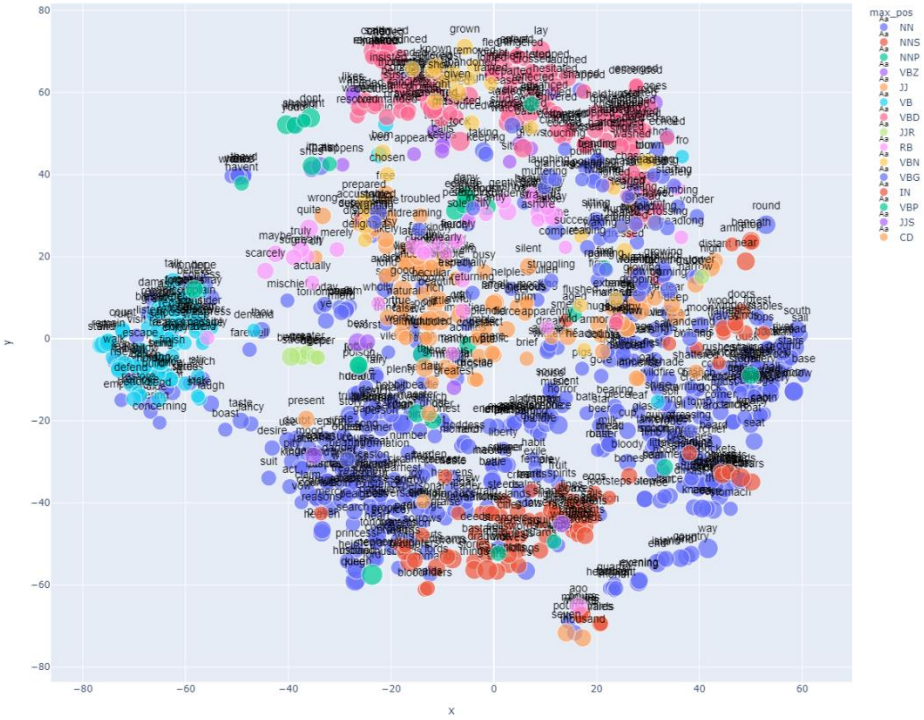


Figure 4 can be better interpreted by zooming in on clusters within the notebook. An obsevation from this view is the blue and purple cluster to the left contains many words about arriving, interacting with or leaving an area/event. The cluster at the bottom right of the plot contains a variety of time related items. For a more in depth look, visit the plots in the notebook.

Notebook 7 Sentiment Analysis

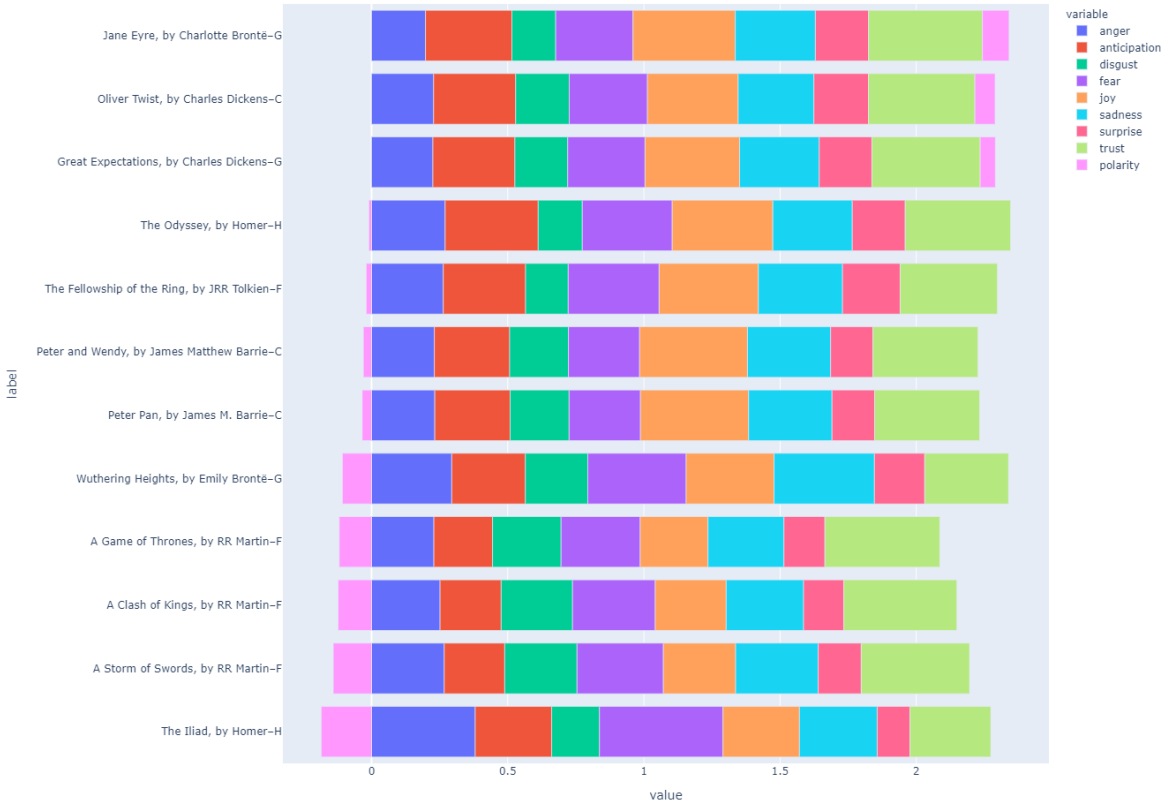
The final notebook investigates sentiment using Salex. The salex_nrc.csv can be found in the data folder and contains a term string index, 7 emotion columns, and a polarity column. The words are all assigned a 0 or 1 depending on if they are associated with that emotion. The polarity column has values -1, 0, and 1. I primarily investigated sentiment at book level to see the key emotions across the corpus.

Table 4: Salex Sentiment Table by Book (Sorted by Polarity)

book_id	anger	anticipation	disgust	fear	joy	sadness	surprise	trust	polarity	tfidf	label
6130	0.380709	0.280774	0.175413	0.453515	0.280774	0.285471	0.120019	0.296890	-0.184888	0.033574	The Iliad, by Homer–H
3	0.267112	0.221901	0.265488	0.317015	0.265352	0.303253	0.157785	0.397735	-0.140640	0.037318	A Storm of Swords, by RR Martin–F
2	0.251574	0.224347	0.262069	0.303696	0.260494	0.284223	0.147621	0.414995	-0.122726	0.034514	A Clash of Kings, by RR Martin–F
1	0.228867	0.215536	0.251542	0.290538	0.248925	0.278951	0.150190	0.421977	-0.118732	0.039821	A Game of Thrones, by RR Martin–F
768	0.294534	0.270243	0.229251	0.361167	0.322706	0.368590	0.184717	0.307524	-0.106950	0.027927	Wuthering Heights, by Emily Brontë–G
16	0.232191	0.276826	0.216411	0.261948	0.397205	0.306132	0.156447	0.385482	-0.034265	0.056761	Peter Pan, by James M. Barrie–C
26654	0.230665	0.276529	0.215827	0.261241	0.396133	0.305306	0.155126	0.385791	-0.030126	0.056165	Peter and Wendy, by James Matthew Barrie–C
4	0.263484	0.302123	0.156534	0.333773	0.364500	0.308849	0.211658	0.357115	-0.019649	0.025314	The Fellowship of the Ring, by JRR Tolkien–F
1727	0.270355	0.341573	0.161860	0.330194	0.369433	0.291740	0.194232	0.387090	-0.010006	0.027994	The Odyssey, by Homer–H
1400	0.225302	0.301128	0.193912	0.283870	0.347330	0.292295	0.193369	0.397608	0.055714	0.030915	Great Expectations, by Charles Dickens–G
730	0.227946	0.301236	0.196745	0.287043	0.333293	0.278111	0.200049	0.390921	0.074147	0.038960	Oliver Twist, by Charles Dickens–C
1260	0.198074	0.317286	0.160706	0.283929	0.375516	0.294131	0.195323	0.418157	0.097662	0.033744	Jane Eyre, by Charlotte Brontë–G

Table 4 shows the Game of Thrones books on the opposite end of the spectrum for polarity than Jane Eyre. The only emotions that seem to relatively correlate between Game of Thrones and Jane Eyre are fear, sadness, and trust. Both groupings contain a variety of sad parts and need for trust throughout the stories, so this is not a surprise.

Figure 5: Stacked Emotion Chart by Book



It is worth noting in Figure 5, that only three stories (Jane Eyre, Oliver Twist, Great Expectations) had positive polarity. On the contrary the three game of thrones books rank 2-4 on being most negative. The other key insight from Figure 5 is the red “anticipation” and the orange “joy” chunks of the Game of Thrones books vs Jane Eyre. Jane Eyre has some of the highest values for these emotions, meanwhile all three Game of Thrones books have near the lowest.

Conclusion / Closing Remarks

The goal of this project was to analyze Game of Thrones and Jane Eyre across a variety of different types of novels in the corpus to determine if the two have some similarities. I hypothesized that we would see the following books in the corpus be closest related to Game of Thrones in this order: Lord of the Rings, Homer’s books, Barrie’s books, Dicken’s books, and Bronte’s books. Throughout notebooks 3-7 we uncovered different insights from a variety of text analytics techniques. The general trend across all notebooks was that Game of Thrones and Jane Eyre are rather distinct from one another. One could have assumed this based on intuition; however, it reconfirms to me why if I enjoyed the stories told within RR Martin’s books, then I would probably not be a big fan on Jane Eyre.

One potential next step that I would want to take with this project would be to look at Jane Eyre against all of RR Martin’s books alone. I would go into more depth in Jane Eyre and each of RR Martin’s novels and try to see which one relates closest Jane Eyre. I would then like to see if the novels rank order in how much I liked the RR Martin story versus how different they are from Jane Eyre.

Appendix

LIB Table

book_id	book_title	book_file	chap_regex	book_length	n_chaps	genre	mood	author	label
1	A Game of Thrones, by RR Martin	corpus/MARTIN_A_GAME_OF_THRONES-pg1.txt	[A-Z]+[A-Z]+[A-Z]+	294315	78	fantasy	adventure	RR Martin	A Game of Thrones, by RR Martin–F
2	A Clash of Kings, by RR Martin	corpus/MARTIN_A_CLASH_OF_KINGS-pg2.txt	[A-Z]+[A-Z]+[A-Z]+	324029	142	fantasy	adventure	RR Martin	A Clash of Kings, by RR Martin–F
3	A Storm of Swords, by RR Martin	corpus/MARTIN_A_STORM_OF_SWORDS-pg3.txt	[A-Z]+[A-Z]+[A-Z]+	417469	95	fantasy	adventure	RR Martin	A Storm of Swords, by RR Martin–F
4	The Fellowship of the Ring, by JRR Tolkien	corpus/TOLKIEN_THE_FELLOWSHIP_OF_THE_RING-pg4.txt	_Chapter	180888	22	fantasy	adventure	JRR Tolkien	The Fellowship of the Ring, by JRR Tolkien–F
16	Peter Pan, by James M. Barrie	corpus/BARRIE_PETER_PAN-pg16.txt	((Chapter)\s+\D+)	47631	17	childrensfantasy	adventure	James M. Barrie	Peter Pan, by James M. Barrie–C
730	Oliver Twist, by Charles Dickens	corpus/DICKENS_OLIVER_TWIST-pg730.txt	((CHAPTER)\s+\D+)	160895	53	crime	satire	Charles Dickens	Oliver Twist, by Charles Dickens–C
768	Wuthering Heights, by Emily Brontë	corpus/BRONTE_WUTHERING_HEIGHTS-pg768.txt	((CHAPTER)\s+\D+)	118361	34	gothic	tragic	Emily Brontë	Wuthering Heights, by Emily Brontë–G
1260	Jane Eyre, by Charlotte Brontë	corpus/BRONTE_JANE_EYRE-pg1260.txt	((CHAPTER)\s+\D+)|PREFACE	191636	39	gothic	romance	Charlotte Brontë	Jane Eyre, by Charlotte Brontë–G
1400	Great Expectations, by Charles Dickens	corpus/DICKENS_GREAT_EXPECTATIONS-pg1400.txt	((Chapter)\s+\D+)	188910	59	gothic	regretful	Charles Dickens	Great Expectations, by Charles Dickens–G
1727	The Odyssey, by Homer	corpus/HOMER_THE_ODYSSEY-pg1727.txt	((BOOK)\s+\D+)	118692	24	historical	serious	Homer	The Odyssey, by Homer–H
6130	The Iliad, by Homer	corpus/HOMER-THE-ILIAD-pg6130.txt	((BOOK)\s+\D+)	152654	24	historical	serious	Homer	The Iliad, by Homer–H
26654	Peter and Wendy, by James Matthew Barrie	corpus/BARRIE_PETER_AND_WENDY-pg26654.txt	((CHAPTER)\s+\D+)	48031	17	childrensfantasy	adventure	James M. Barrie	Peter and Wendy, by James Matthew Barrie–C

HTML Links

Non-Gutenberg

https://www.kaggle.com/datasets/ashishsinhaiitr/lord-of-the-rings-text

https://archive.org/stream/3.AStormOfSwords/3.%20A%20Storm%20of%20Swords_djvu.txt

https://archive.org/stream/II.AClashOfKings_201803/II.%20A%20Clash%20of%20Kings_djvu.txt

https://archive.org/stream/1AGameOfThrones/1%20A%20Game%20of%20Thrones_djvu.txt

All other novels are from Project Gutenberg:

Free eBooks | Project Gutenberg