Final Report

DS 5001 Spring 2023 Final Project

Rachel Grace Treene

rg5xm@virginia.edu

Introduction

The corpus used for this project is the seven Harry Potter books written by J.K. Rowling. It has 1,118,267 tokens and 23,096 unique terms. The books in the corpus tell the story of Harry Potter and his adventures, culminating in an epic battle between himself and Voldemort, the evil wizard.

In this project, I consider what Quidditch, the main sport played by people in the wizarding world, can tell us about the structure and similarity of the books to one another. When considering similarity between books, can we trace any patterns relating to the subject of Quidditch that might mirror book similarity? My investigation has three parts: first, I ask what we can learn about the correlation between the books in the Harry Potter series. Second, I ask what patterns we can find showing the prevalence of Quidditch throughout the books. Third, I consider what similarities exist between patterns of book correlation and patterns of the prevalence and inclusion of Quidditch throughout the series.

Source Data

The data comes from text files, obtained from a GitHub repository which used the Harry Potter corpus for an NLP project.

Provenance

The original text files are located at the following repository: https://github.com/ErikaJacobs/Harry-Potter-Text-Mining. A few errors were located after downloading which were fixed to ensure more accurate work could be done.

Location

The slightly edited source files for this project have been added to the current GitHub repository and can be found in the data directory. The link to that directory is as follows: https://github.com/rachelgracetreene/text-analytics-final-project/tree/main/data.

Description

The subject matter of the corpus is the fictional accounts of Harry Potter, a boy who is a wizard in England in the 1990s. The source files are structured in lines, where a line is an observation, and the average document length in terms of lines is 15744.57 lines. The average document length in terms of tokens is 189055.14 tokens.

Format

The source files are TXT files. The data in the files is plaintext.

Data Model

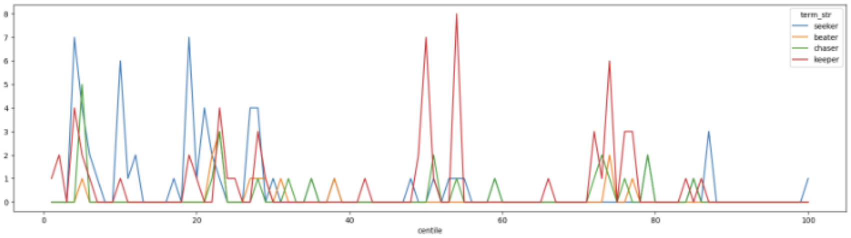
We processed and annotated the corpus and converted it into the standard text analytic data model (F2) format. The main challenge during the processing was defining the correct regular expression to select chapters; this was made difficult since the chapter titles were in all caps, but some text within chapters was also capitalized. We defined a method to chunk the books by chapter without selecting extraneous text as chapter headers. Our OHCO included book number, chapter number, paragraph number, sentence number, and token number. Processing, annotating, and analyzing the corpus produced tables with features that are described in the data_model Jupyter notebook (https://github.com/rachelgracetreene/text-analytics-final-project/blob/main/data_model.ipynb).

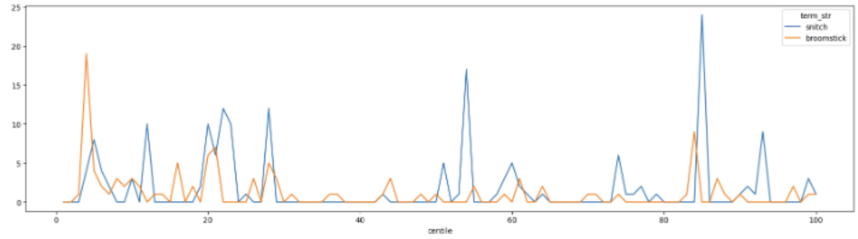
All of the methods used to perform processing, annotating the VOCAB table, LDA, Word2Vec, Sentiment Analysis, and Semantic Search are in a custom Python package called HarryPotterETA (https://github.com/rachelgracetreene/text-analytics-final-project/blob/main/HarryPotterETA.py).

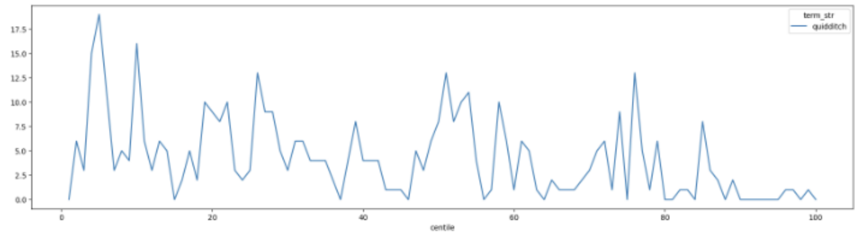
Exploration

First, I converted each book to F2 format. Then I combined each book into one corpus, from which I extracted a vocabulary. I added part of speech, stopwords, and stems to the vocabulary, and calculated term frequency inverse document frequency (TFIDF) and document frequency inverse document frequency (DFIDF).

With the corpus, I generated KDE plots showing the occurrence of the word Quidditch across all seven books, as well as the occurrence of Quidditch-related words like 'seeker', 'beater', 'chaser', 'keeper', 'snitch', and 'broomstick.' The plots are below.

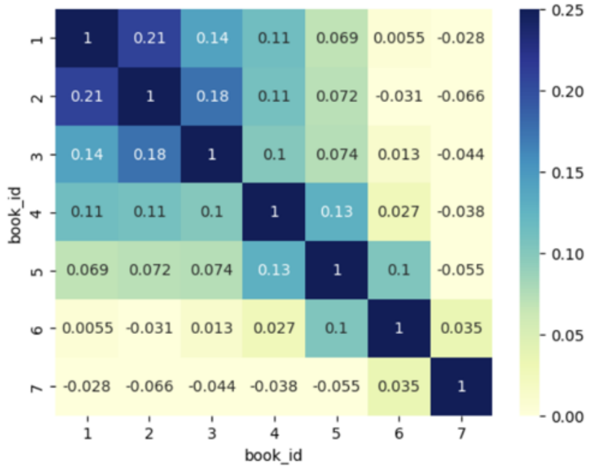






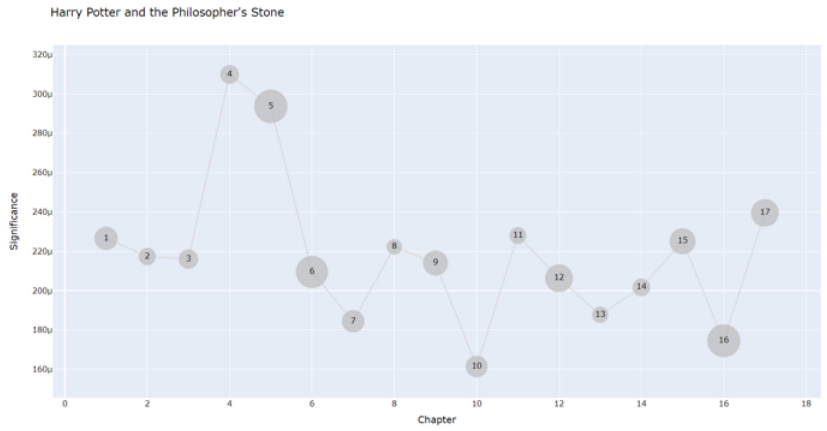
When comparing the KDE of the positions in Quidditch,'Keeper' and 'Seeker' are clearly more prevalent terms than 'Chaser' and 'Beater.' 'Keeper' and 'Seeker' occur more often by comparison. 'Broomstick' has a more random occurrence, while 'Snitch' has a few peaks that become more significant throughout the series. Quidditch is repeated regularly throughout the series, with the most incidence in the beginning and the least at the end.

To compare book similarity, I normalized TFIDF grouped by book and created a table called PAIRS that calculates book similarities with various metrics. I also created a correlation matrix of the books using the Kendall rank and edited the matrix to be a heatmap. It is shown below.



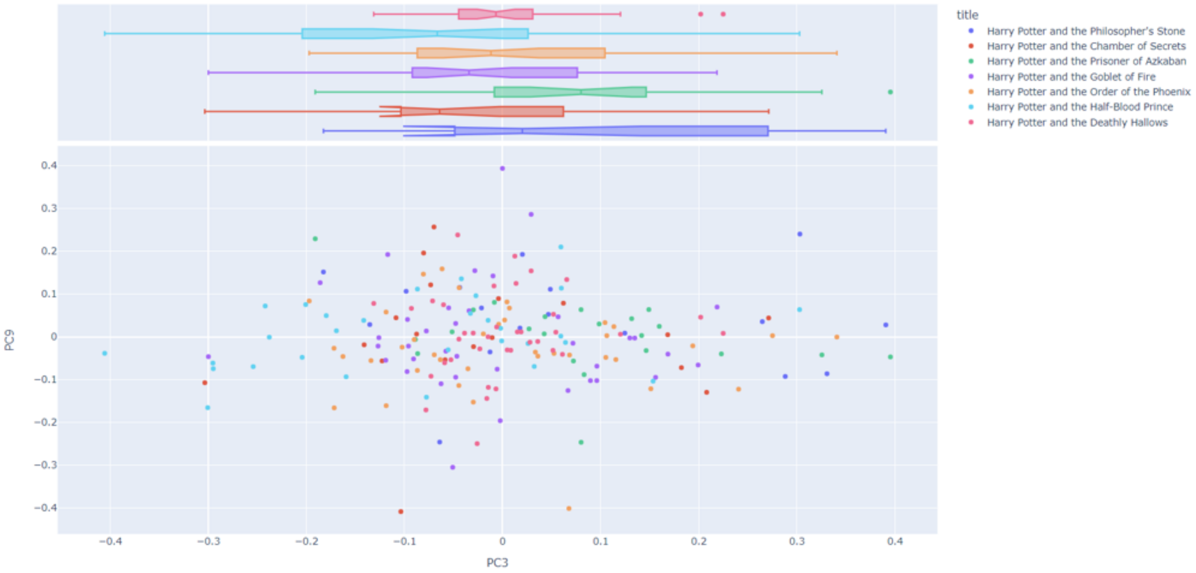
From the graph above, it is clear that books 1, 2, and 3, have strongest correlation with one another. Book 4 appears to be about equally correlated with books 1, 2, and 3. Books 5, 6, and 7 are most correlated with the previous book - that is, book 5 is most correlated with book 4, book 6 is most correlated with book 5, and book 7 is most correlated with book 6. The strength of correlation also decreases as the series continues.

Using mean TFIDF, I generated a significance measurement for each chapter of the corpus and plotting the significance for each book. The observations for my research question is similar for all books, so I only show the significance plot for Book 1. It is below.

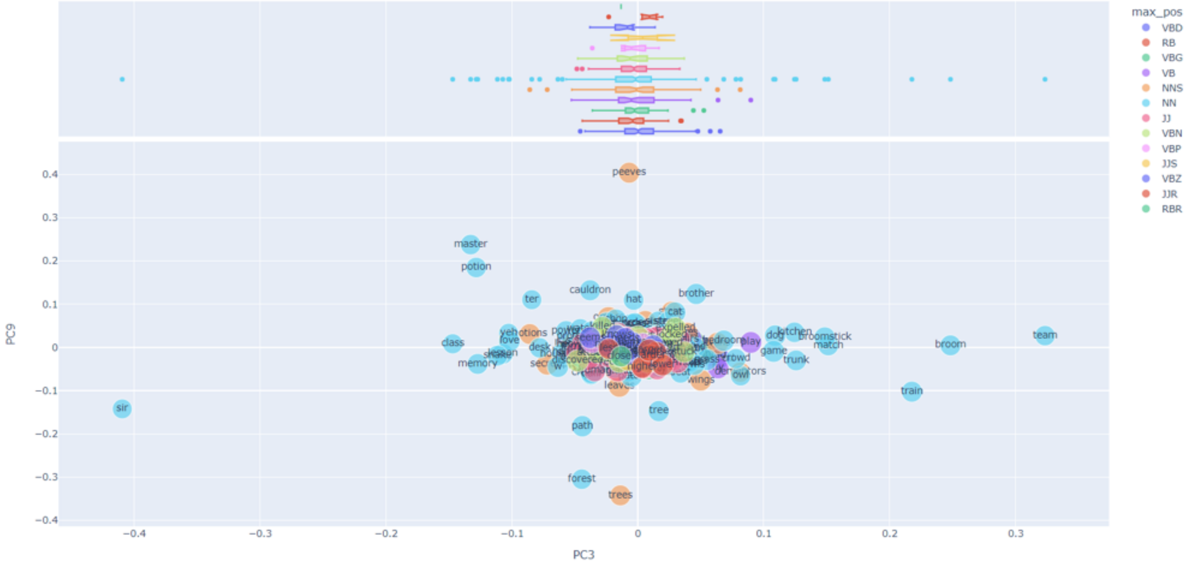


In book 1, the chapters containing a Quidditch match are chapters 11 and 13. In this plot, chapters 11 and 13 do not appear to be 'significant.' In fact, when plotting significance for each book, the same phenomenon is observed - all chapters with Quidditch matches are insignificant.

I conducted Principal Component Analysis (PCA) on the top 1000 terms by DFIDF, excluding proper nouns. I chose the top ten components and created a LOADINGS table to display the output. Below is a plot of components 9 and 3 plotted against each other by book and a plot showing the words in each component. The plots demonstrate an opposition between Quidditch and the classroom.

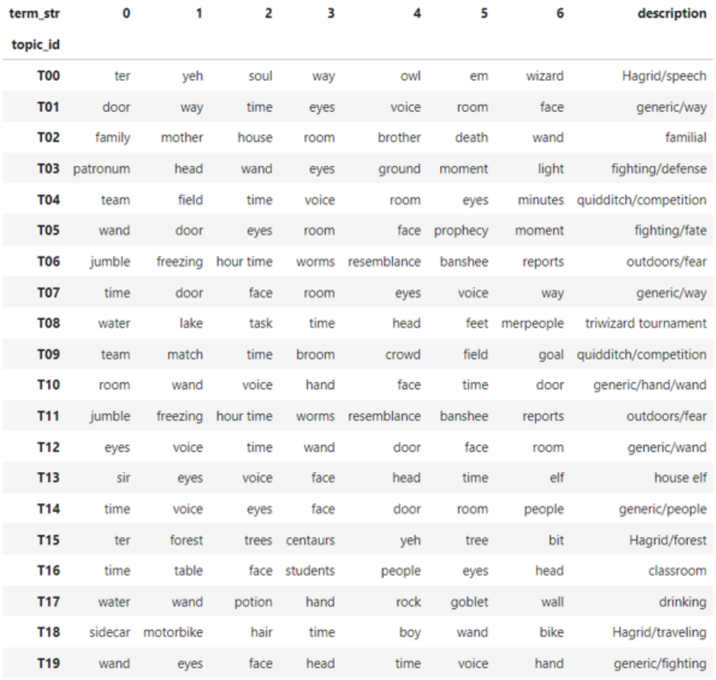


In the graph above, PC3 plotted against PC9 shows an opposition between Quidditch and the classroom. Based on the center measurements of the color-coded box and whisker plots, it is clear that the books 1 and 3 (*Harry Potter and the Philosopher's Stone* and *Harry Potter and the Prisoner of Azkaban*) are located on the 'Quidditch' side of the opposition.



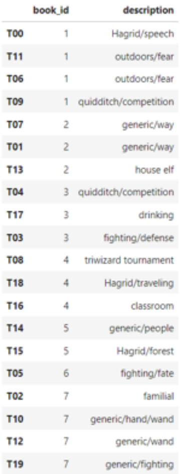
In the graph above, the words corresponding to the components are shown. It is clear that the words pertaining to QUidditch - 'team,' 'broom,' 'train', 'broomstick,' and 'match' are compared with 'sir,' 'class,' 'master,' 'potion,' and 'memory,' words that are associated with the classroom.

I conducted Latent Dirichlet Allocation (LDA) with a vector space. I used an ngram range of [1, 2] and chose 4000 features, with English stopwords. For LDA, I specified 20 topics, 20 components, 5 iteractions, and a learning offset of 5. I chose 7 terms per topic. Then I annotated the generated TOPICS manually with my own description of each topic. The annotated TOPICS table is below.



term_str	0	1	2	3	4	5	6	description
topic_id								
T00	ter	yeh	soul	way	owl	em	wizard	Hagrid/speech
T01	door	way	time	eyes	voice	room	face	generic/way
T02	family	mother	house	room	brother	death	wand	familial
T03	patronum	head	wand	eyes	ground	moment	light	fighting/defense
T04	team	field	time	voice	room	eyes	minutes	quidditch/competition
T05	wand	door	eyes	room	face	prophecy	moment	fighting/fate
T06	jumble	freezing	hour time	worms	resemblance	banshee	reports	outdoors/fear
T07	time	door	face	room	eyes	voice	way	generic/way
T08	water	lake	task	time	head	feet	merpeople	triwizard tournament
T09	team	match	time	broom	crowd	field	goal	quidditch/competition
T10	room	wand	voice	hand	face	time	door	generic/hand/wand
T11	jumble	freezing	hour time	worms	resemblance	banshee	reports	outdoors/fear
T12	eyes	voice	time	wand	door	face	room	generic/wand
T13	sir	eyes	voice	face	head	time	elf	house elf
T14	time	voice	eyes	face	door	room	people	generic/people
T15	ter	forest	trees	centaurs	yeh	tree	bit	Hagrid/forest
T16	time	table	face	students	people	eyes	head	classroom
T17	water	wand	potion	hand	rock	goblet	wall	drinking
T18	sidecar	motorbike	hair	time	boy	wand	bike	Hagrid/traveling
T19	wand	eyes	face	head	time	voice	hand	generic/fighting

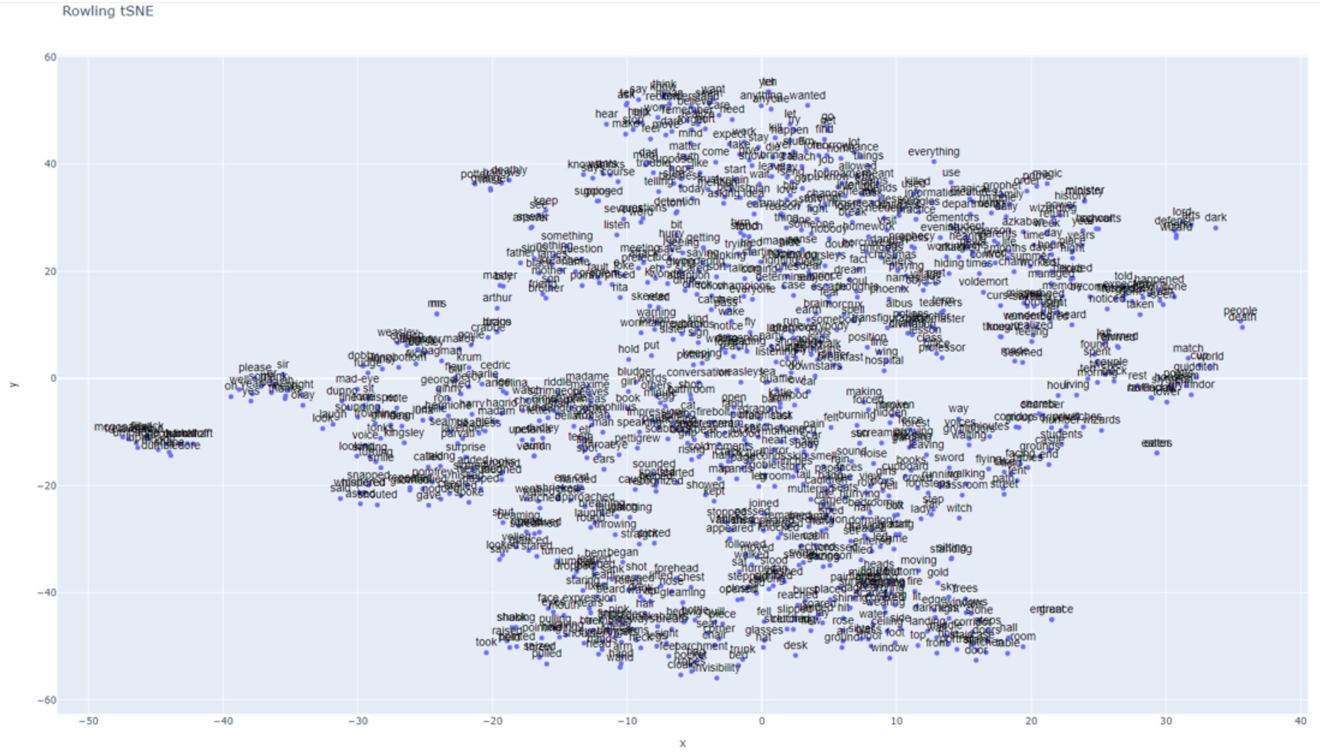
Obviously, topics 4 and 9 pertain to Quidditch based on the words they contain. I generated a table to show, for each topic, which book most contained it. Below are the results.



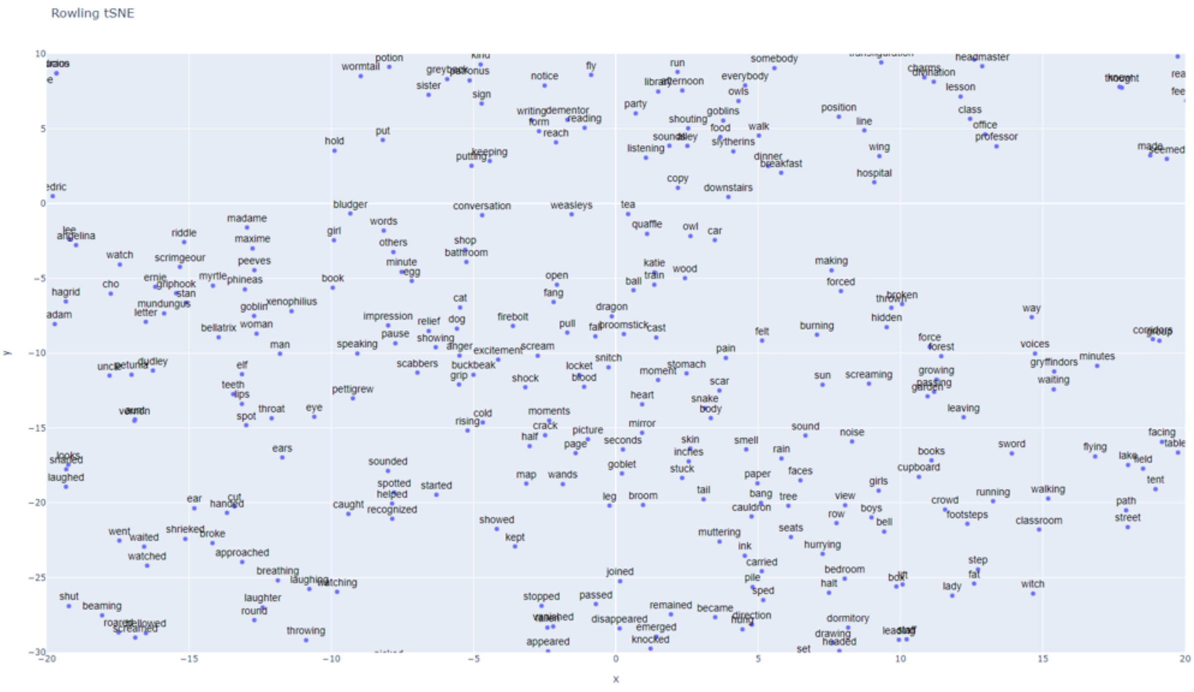
	book_id	description
T00	1	Hagrid/speech
T11	1	outdoors/fear
T06	1	outdoors/fear
T09	1	quidditch/competition
T07	2	generic/way
T01	2	generic/way
T13	2	house elf
T04	3	quidditch/competition
T17	3	drinking
T03	3	fighting/defense
T08	4	triwizard tournament
T18	4	Hagrid/traveling
T16	4	classroom
T14	5	generic/people
T15	5	Hagrid/forest
T05	6	fighting/fate
T02	7	familial
T10	7	generic/hand/wand
T12	7	generic/wand
T19	7	generic/fighting

Topic 9 is most prevalent in book 1, and Topic 4 is most prevalent in book 3.

Next I conducted Word 2 Vec on verbs and nouns, creating a model with a window of 2, a vector size of 256, and a minumum count of 80. For my tSNE table, I used a learning rate of 200, perplexity of 20, 2 components, a random initialization, and 1000 iterations. The resulting tSNE plot is below.

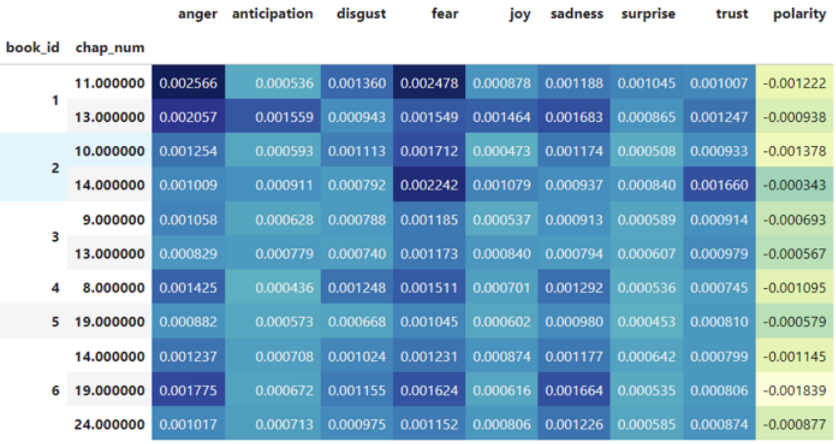


I edited the tSNE plot to show the cluster pertaining to Quidditch more closely, located around (0, -10). The plot is below.



The words 'quaffle,' 'ball,' 'train,' broomstick,' 'snitch,' and 'fall' occur close together, with 'bludger' and 'broom' not far off.

We performed sentiment analysis with the sentiment analysis lexicon and plotted emotions for each book. I selected the chapters from each book in which a Quidditch match occurs and generated the emotions in those chapters as a heatmap for each book. Finally, I combined those rows of the heatmap. The plot is below.



book_id	chap_num	anger	anticipation	disgust	fear	joy	sadness	surprise	trust	polarity
1	11.000000	0.002566	0.000536	0.001360	0.002478	0.000878	0.001188	0.001045	0.001007	-0.001222
	13.000000	0.002057	0.001559	0.000943	0.001549	0.001464	0.001683	0.000865	0.001247	-0.000938
2	10.000000	0.001254	0.000593	0.001113	0.001712	0.000473	0.001174	0.000508	0.000933	-0.001378
	14.000000	0.001009	0.000911	0.000792	0.002242	0.001079	0.000937	0.000840	0.001660	-0.000343
3	9.000000	0.001058	0.000628	0.000788	0.001185	0.000537	0.000913	0.000589	0.000914	-0.000693
	13.000000	0.000829	0.000779	0.000740	0.001173	0.000840	0.000794	0.000607	0.000979	-0.000567
4	8.000000	0.001425	0.000436	0.001248	0.001511	0.000701	0.001292	0.000536	0.000745	-0.001095
5	19.000000	0.000882	0.000573	0.000668	0.001045	0.000602	0.000980	0.000453	0.000810	-0.000579
6	14.000000	0.001237	0.000708	0.001024	0.001231	0.000874	0.001177	0.000642	0.000799	-0.001145
	19.000000	0.001775	0.000672	0.001155	0.001624	0.000616	0.001664	0.000535	0.000806	-0.001839
	24.000000	0.001017	0.000713	0.000975	0.001152	0.000806	0.001226	0.000585	0.000874	-0.000877

The predominant emotions felt throughout all chapters appear to be fear and anger.

Finally, we added code for a semantic search to the project. No visualization was produced that is helpful for this research question.

Interpretation

These visualizations enable an exploration of book correlation and similarity. In particular, the heatmap showing the Kendall sum correlations indicates that books 1, 2, and 3 are most similar to one another, while books 4, 5, 6, and 7 are more different from the others.

The visualizations also enable an exploration of Quidditch occurrence and themes pertaining to the sport throughout the books. The KDE plot of the term 'quidditch' throughout the book suggests a fairly regular appearance of the word; however, it decreases in frequency as the series goes on, with the most prevalence noted in the earliest books. The lack of 'significance' for all chapters with Quidditch matches across the series indicates that Quidditch may occupy a similar role in all the books in which matches occur (notably, book 7 does not have any Quidditch matches). In addition, the heatmap showing emotions during each chapter with a Quidditch match indicates similarity in emotional theme for all matches for all books.

A few visualizations indicate that Quidditch may be a more dominant theme in the first few books of the series. First, the higher peaks of occurrence of 'quidditch' in the KDE plot align with the first two books in particular. Second, results from PCA show that books 1 and 3 skew more favorably toward the Quidditch 'side' of the opposition, compared with the classroom 'side.' Third, results from LDA show that the topics pertaining to Quidditch are most present in books 1 and 3, respectively.

Generally, it seems that the occurrence of Quidditch throughout the books does have some correlation with similarity of books to one another. My results indicate that books 1, 2, and 3 are most similar to one another, and Quidditch is most prevalent in these books based on several measures. Further exploration might expand on this interpretation, and could explore whether, on a book-by-book basis, any kind of narrative similarity in Quidditch matches could be found. I would also be interested in exploring whether the chapters with Quidditch matches play similar roles in the plots of their books. The significance measure used above is an experimental one, so I would hope to explore that differently.