Final Report: Analysis of Hermione in J. K. Rowling's *Harry Potter*

DS5001 Spring 2023 Final Project - Exploratory Text Analytics

Eve Schoenrock, UFU2RG

Contact: ufu2rg@virginia.edu

1 Introduction

Harry Potter has become a worldwide sensation since its inception in 1997, inspiring children and young adults for the past three decades to harness their creativity and act on their imagination. What many of Rowling's avid fans do not know is that the beloved author was struggling with domestic violence during the formative era of the *Sorcerer's Stone*. By the time she and her daughter abandoned the abuse, Rowling had written the first three chapters of her first Harry Potter novel. (https://www.biography.com/authors-writers/jk-rowling-harry-potter-author-rags-to-riches-billionaire (https://www.biography.com/authors-writers/jk-rowling-harry-potter-author-rags-to-riches-billionaire))

As seen by the series's incredible explosion, Rowling's personal and financial situation completely changed after publishing. However, I suspect that experiences from her early life seaped into the Harry Potter series. A mother and daughter abandoning an abusive but comfortable life in favor of their deserved safety and independence serves as a powerful feminist story that is worth exploration. Because of Rowling's experience, I expect to find a strong feminine presence within Harry Potter. I focus my analysis on Hermione, one of the most influential women in *Harry Potter*, asking how she impacts the plotline and what her influence suggests about the importance of female strength.

The corpus I use to answer this consists of the first seven novels of the Harry Potter series, pulled from a GitHub repository. Analyzing Rowling's words enables me to delve into the importance of Hermione.

2 Source Data

The digital versions of the seven Harry Potter novels were downloaded from GitHub at the following link: https://github.com/prakhar21/whiteboard/tree/master/nbviewer/notebooks/data/harrypotter (https://github.com/prakhar21/whiteboard/tree/master/nbviewer/notebooks/data/harrypotter).

The seven separate source files are in .txt format. Each book has a different regex required for chapter parsing. The source files are downloaded and located in UVA Box at the following link: https://virginia.box.com/s/706sjv4ritfbecfx05xp2o1v89qbvxtb (https://virginia.box.com/s/706sjv4ritfbecfx05xp2o1v89qbvxtb)

The corpus contains the seven books in the Harry Potter series, by J. K. Rowling. The seven book titles are included in order of publication: *Sorcerer's Stone*, *Chamber of Secrets*, *Prisoner of Azkaban*, *Goblet of Fire*, *Order of the Phoenix*, *Half-Blood Prince*, *Deathly Hallows*. There are **1,323,386 observations** (token-based) in the total corpus, where all books are combined. *Sorcerer's Stone* has 94,874 token-based observations, *Chamber of Secrets* has 104,607 token-based observations, *Prisoner of Azkaban* has 132,238 token-based observations, *Goblet of Fire* has 230,960 token-based observations, *Order of the Phoenix* has 313,490 token-based observations, *Half-Blood Prince* 207,051 token-based observations, and *Deathly Hallows* 240,166 token-based observations. Among all books, the average document length is 189,055 rounded to the nearest whole number.

The Ordered Content of Hierarchy Objcects (OHCO) for this project follows: `['book_id', 'chap_num', 'para_num', 'sent_num', 'token_num']`. Each level of the OHCO denotes book levels in the Harry Potter series. Throughout this report, I focus on book-level and/or chapter-level analysis.

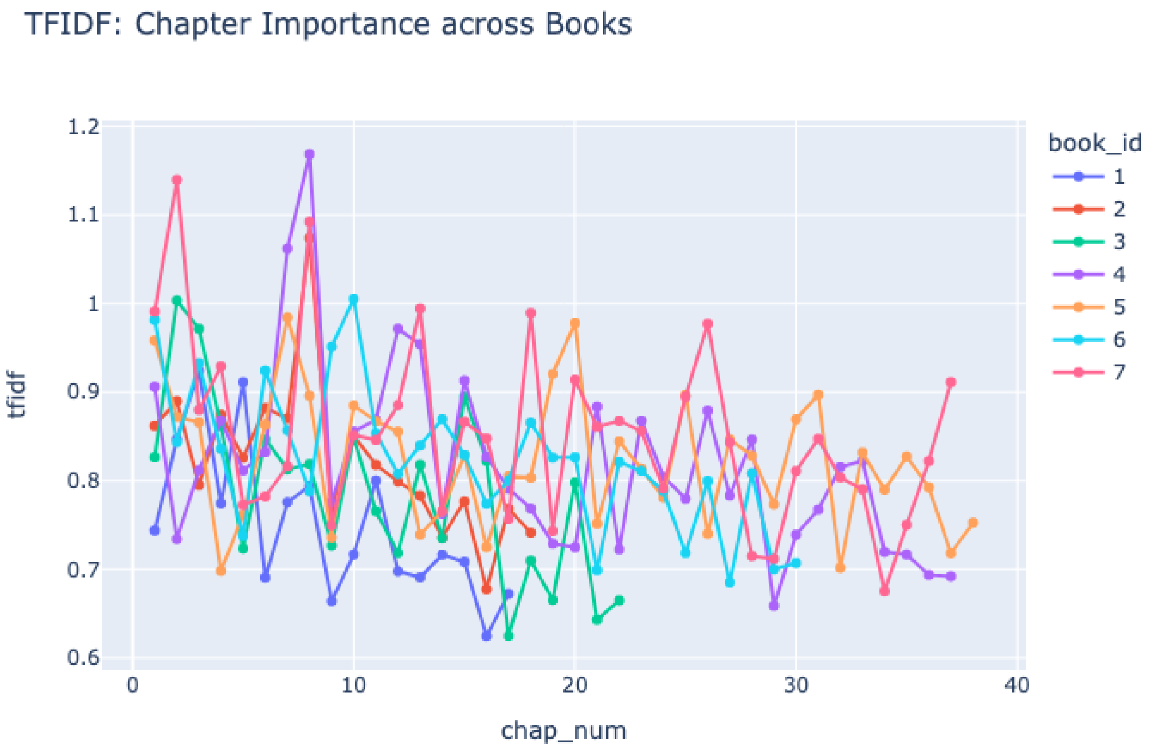
3 Data Model

Descriptions of all analytical tables generated from tokenization, annotation, and corpus analysis are contained in the GitHub repository associated with this report, located at the link: https://github.com/eveschoen/harry-potter/blob/main/data_model_description.md (https://github.com/eveschoen/harry-potter/blob/main/data_model_description.md).

The tables saved to the GitHub repository are `CORPUS`, `LDA-PHI`, `LDA-THETA`, `LIB`, `PCA-DCM`, `PCA-LOADINGS`, `SA-DOCEMOTIONS`, `SA-VOCAB`, `VOCAB`, and `W2V-VOCAB`. Other relevant tables are included in exploratory notebooks, `exploratory_analysis.ipynb`, `hermione.ipynb`, and `visualization.ipynb`.

4 Exploration

To establish a foundation for analysis on Hermione's character, I first identify crucial moments in the Harry Potter series by analyzing mean TFIDF across chapters of each book. In this case, chapters are considered moments. The connected scatterplot, *TFIDF: Chapter Importance across Books*, is displayed below.

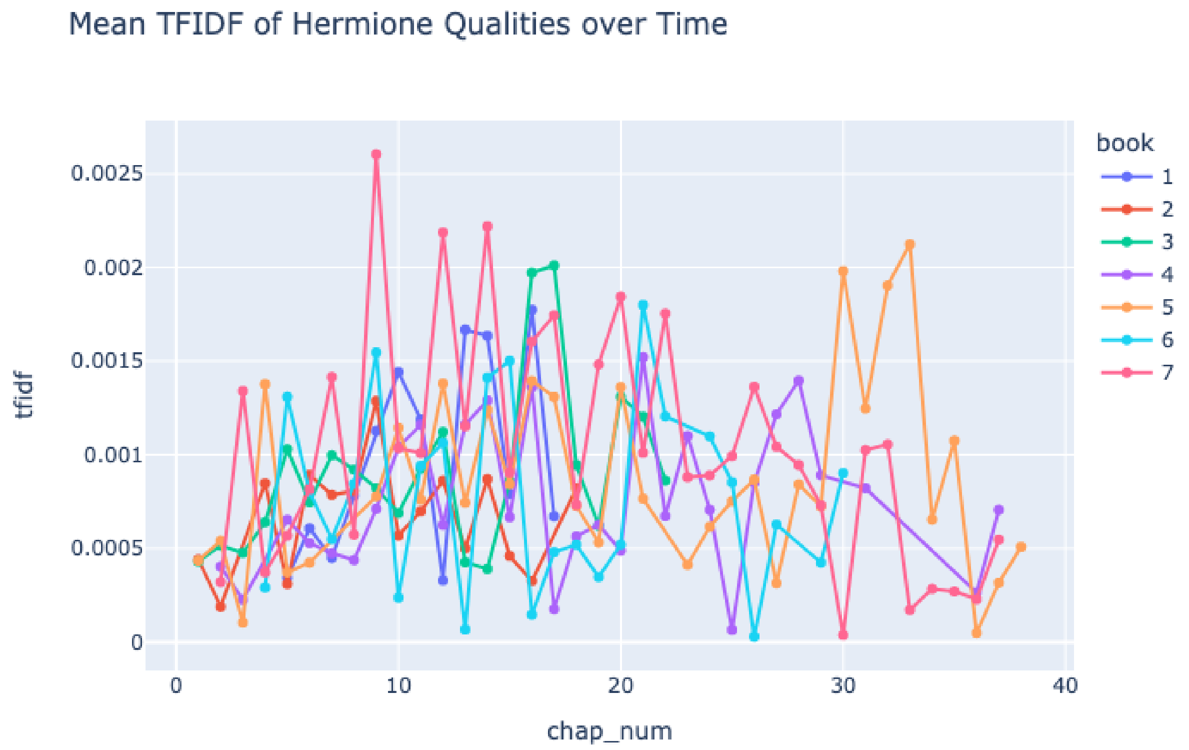


The most important chapter/book combinations according to TFIDF in the Harry Potter series are often early in a novel. Key peaks are:

- Book 7, Chapter 2
- Book 4, Chapter 7
- Book 7, Chapter 8
- Book 2, Chapter 8
- Book 6, Chapter 10
- Book 7, Chapter 13
- Book 7, Chapter 18
- Book 5, Chapter 20
- Book 7, Chapter 26

For TFIDF measurement in this visualization, I grouped the dataset by book and chapter and took the sum TFIDF per chapter.

Another TFIDF plot, specific to Hermione and her qualities, is shown below, titled *Mean TFIDF of Hermione Qualities over Time*.



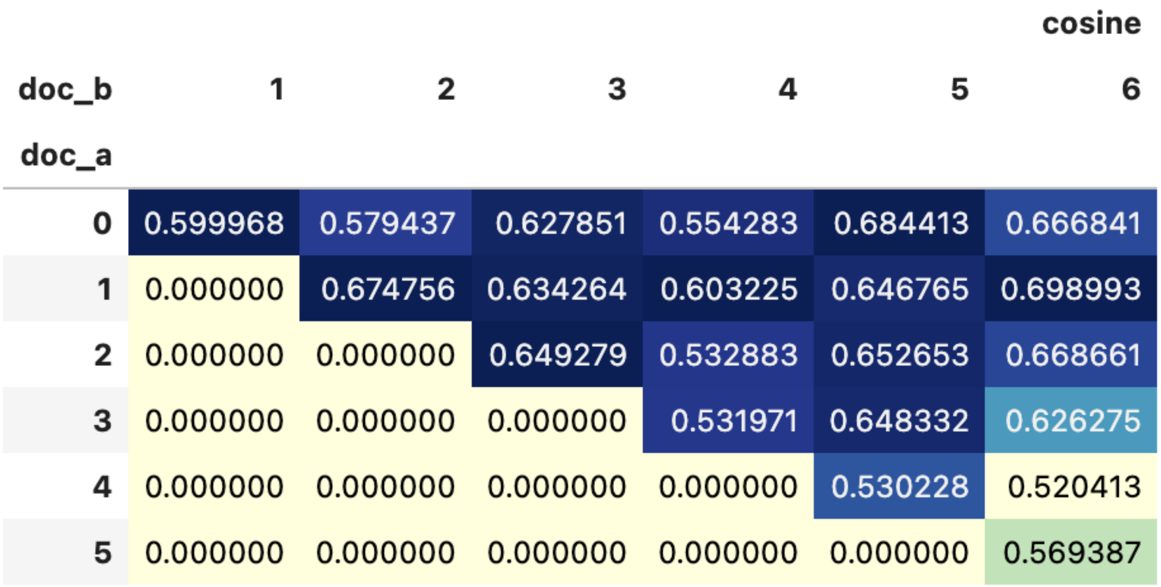
The peaks indicate moments where Hermione is most important/relevant according to TFIDF. Hermione and her qualities are defined by the following terms: *hermione*, *granger*, *smart*, *intelligent*, *clever*, *logical*, and *bossy*. The terms describing Hermione are based on the wikipedia page about Hermione Granger. (https://en.wikipedia.org/wiki/Hermione_Granger#:~:text=Hermione's%20most%20prominent%20features%20inc

(https://en.wikipedia.org/wiki/Hermione_Granger#:~:text=Hermione's%20most%20prominent%20features%20inc

There are clear TFIDF peaks in books 3, 4, 5, 6, and 7. Hermione seems less important in books 1 and 2 because there are more troughs in TFIDF. This matches the plot of the series because in the beginning Harry and Ron are unsure about Hermione and whether they like her. At the end of book 5, Hermione peaks in importance. Throughout all of book 7, Hermione exhibits great importance - this fits the plot of the novel because Hermione repeatedly saves Harry from great trouble.

For this visualization, I inspect mean TFIDF of Hermione language (`['hermione', 'granger', 'smart', 'intelligent', 'clever', 'logical', 'bossy']`) after grouping by book and chapter. Mean TFIDF is preferred for this visualization because sum TFIDF might artificially inflate the importance of chapters where one or more of these terms is considered highly important.

Following analysis of Hermione's importance and chapter importance across *Harry Potter*, I investigate the cosine similarity between each book. This serves as a good baseline of understanding by identifying books that are closely correlated with one another. The heatmap of cosine similarity across books in the corpus is shown below.



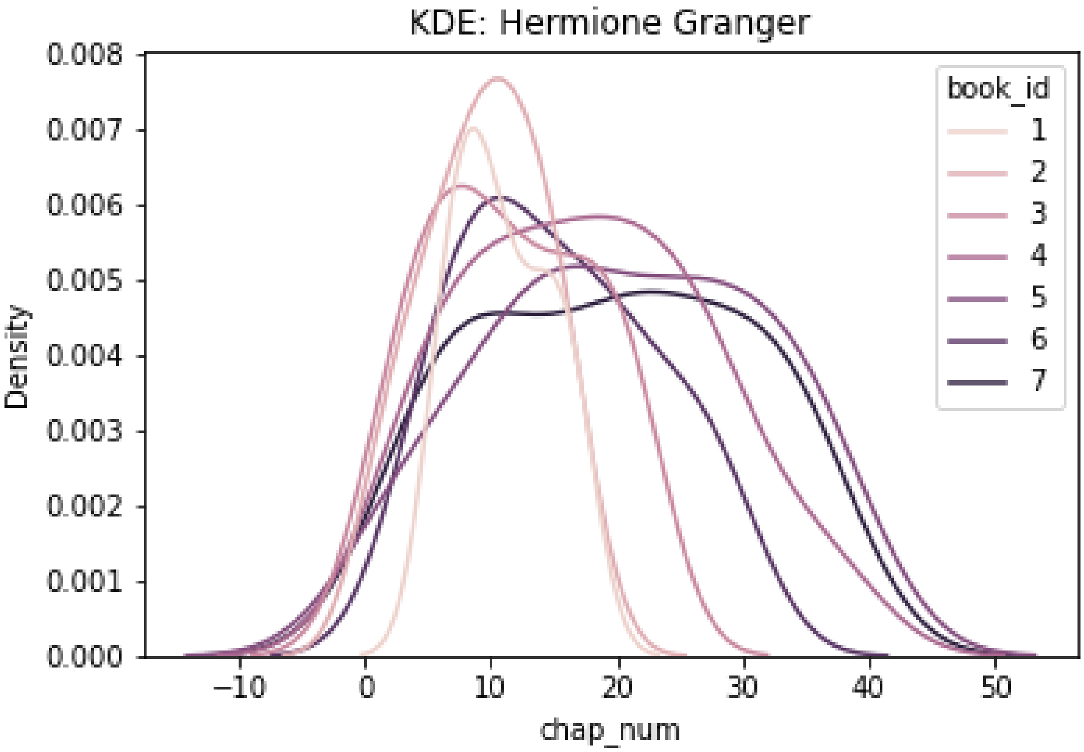
	cosine					
doc_b	1	2	3	4	5	6
doc_a						
0	0.599968	0.579437	0.627851	0.554283	0.684413	0.666841
1	0.000000	0.674756	0.634264	0.603225	0.646765	0.698993
2	0.000000	0.000000	0.649279	0.532883	0.652653	0.668661
3	0.000000	0.000000	0.000000	0.531971	0.648332	0.626275
4	0.000000	0.000000	0.000000	0.000000	0.530228	0.520413
5	0.000000	0.000000	0.000000	0.000000	0.000000	0.569387

Although these results are interesting, I find that all books are closely related, likely because they all follow the same characters throughout a series. It makes sense that books would be correlated - that is the intention! The Harry Potter series builds upon itself. The three greatest similarities between books follow:

- Book 2 and book 7
- Book 1 and book 6
- Book 2 and book 3

Note that books where Hermione is seen a significant character (books 3, 6, and 7) are correlated according to cosine with books where Hermione is not as relevant (books 1 and 2). The parameters for the heatmap are cosine similarity and `colors = "YlGnBu"` .

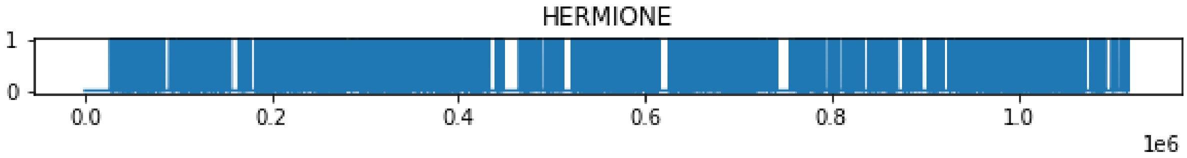
Next, a KDE plot shows the estimated distribution of densities across chapters by book of the individual words, "hermione" and "granger."



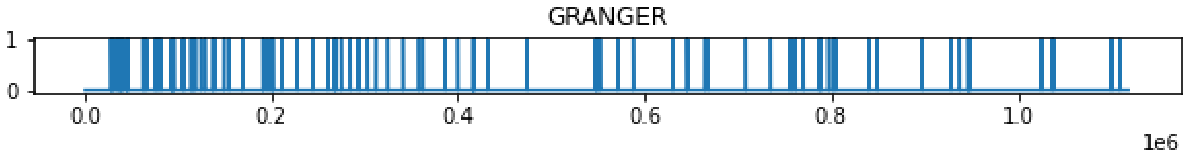
Based on the KDE plot above, "Hermione" and "Granger" appear most frequently in the middle of each book, which is where we anticipate the climax of each novel to occur. It is also clear that in the first three books, there is a higher density peak than in the last four books - this phenomenon correpsonds with books where Hermione is deemed unimportant (books 1-3) and more important (books 4-7) based on the *Mean TFIDF of Hermione Qualities over Time*. Density peaks occur in novels where Hermione is considered less important because her mentioning is much more concentrated to a specific section, whereas in the last four novels density width is greater, indicating Hermione is relevant for greater periods of time.

Again, parameters for this graph are a grouping by book and chapter. Densities are confined to the terms `['hermione', 'granger']`.

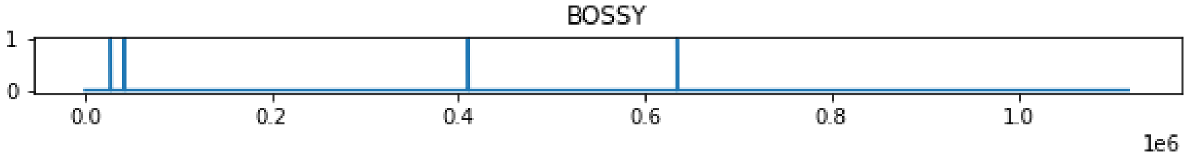
Word dispersion plots are shown below. They demonstrate the frequency of term count through the duration of the Harry Potter series.



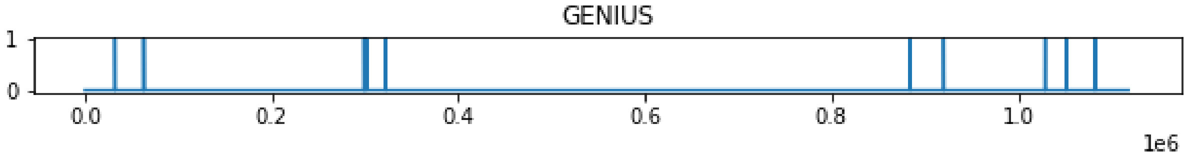
The word dispersion plot above corresponds to the term, "hermione," the first name of the female character of interest in my analysis. Hermione's first name appears consistently throughout the entire series, with only brief pauses in mention that can almost go unnoticed.



The word dispersion plot above correpsonds to the term, "granger," the last name of the female character of interest. Hermione's last name appears more infrequently than her first name, especially later in the Harry Potter series.



Hermione is often described as "bossy," especially in the beginning of the Harry Potter series because the other main characters, Ron and Harry, do not appreciate her personality at first. The word dispersion plot for the term, "bossy" is shown above. The word dispersion plot above reflects Harry's perception of Hermione throughout the plot of the series.

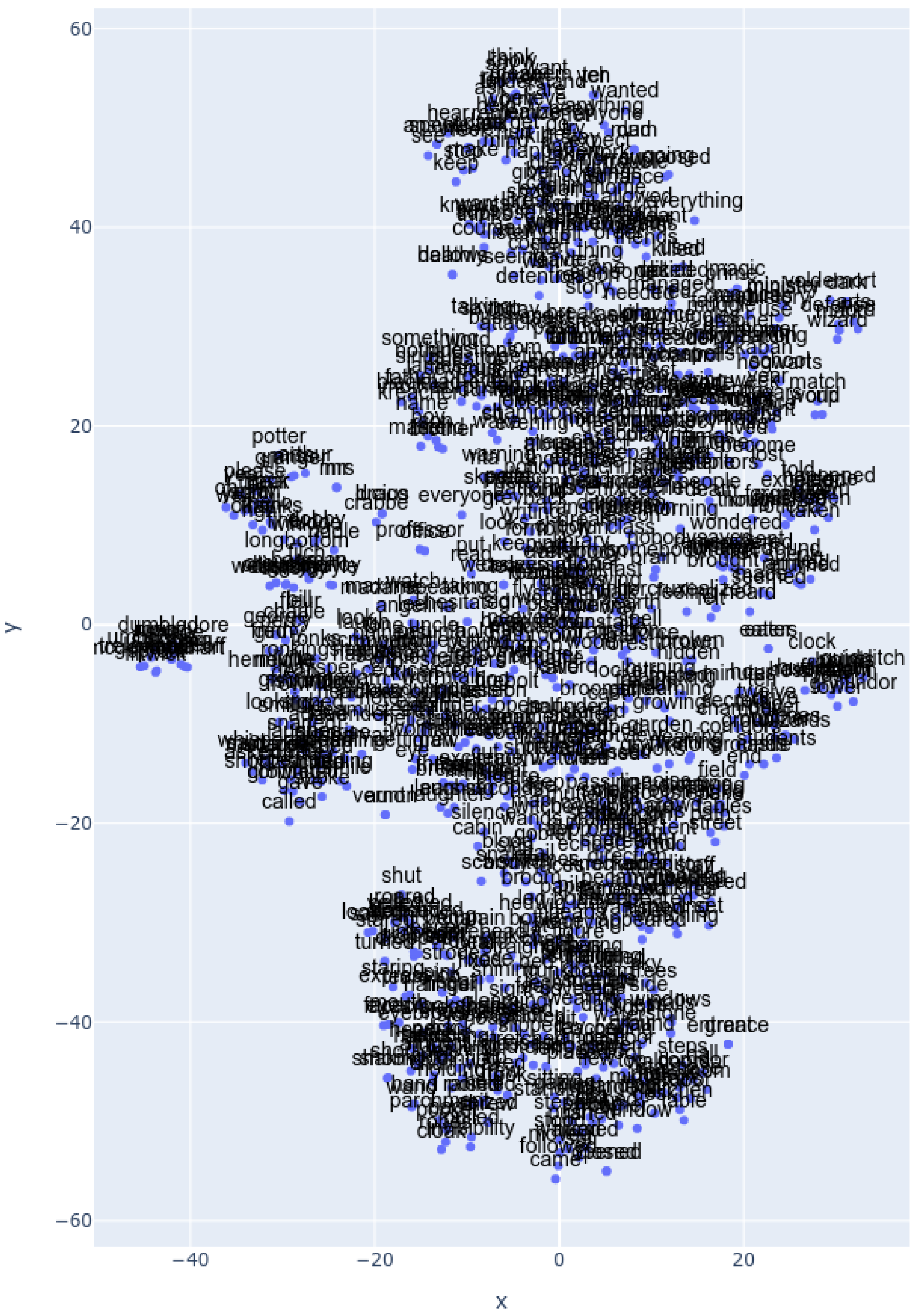


Hermione is also described as "genius," a contrasting word to "bossy." Her immense intelligence awards her this compliment. The word dispersion plot for the term, "genius" is shown above. It is interesting to see that there are clusters where the word "genius" is used.

There are no specific parameters used for generation of these word plots other than the terms themselves and use of the corpus table.

The tSNE plot below shows word clusters where utterances that are semantically similar are located together.

Rowling tSNE

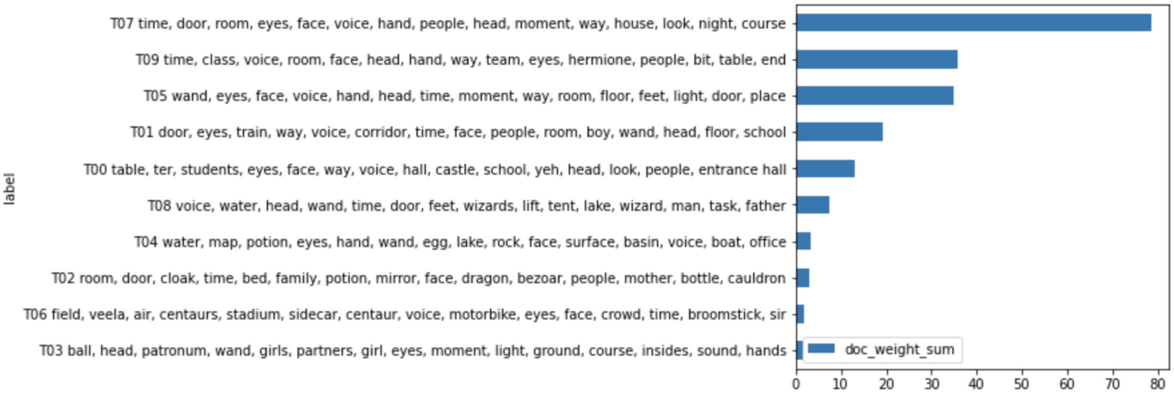


Hermione's last name, "Granger," is next to Harry's last name, "Potter." Additionally, "Hermione" is in a cluser with "Harry," "Ron," "Ginny," and "Snape" who are all main characters in the series. The tSNE evaluation relies on the following parameters:

```
learning_rate = 200
perplexity = 20
n_comps = 2
init = 'random'
n_iter = 1000
rand_state = 42
```

Stopwords were dropped for visualization purposes.

PCA was irrelevant with regard to the posed question of Hermione's importance throughout the Harry Potter series. However, topic modeling proves to be useful.

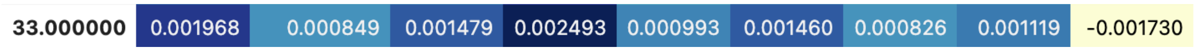


Hermione is shown in the the second most important topic in terms of document weight by book/chapter combination, topic T09. T09 is thematically focused on the classroom setting at Hogwarts, which is a setting that values intellect and cleverness. The LDA topic analysis suggests that throughout all books and chapters, out of all topics, T09 is the second most common topic, and it happens to contain Hermione as the *only character* in the top fifteen words associated with the topic. The topic of Hogwarts schooling suggests that a key value throughout the Harry Potter series is intelligence, and it is a compliment that Hermione is portrayed as the most intelligent character.

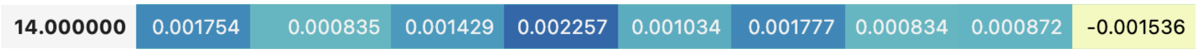
The parameters for the LDA topic model are displayed:

```
ngram_range = [1,2]
n_terms = 4000
n_topics = 10
max_iter = 20
n_top_terms = 15
```

Following topic modeling, I conduct sentiment analysis to investigate the presence of trust in chapters where Hermione and her qualities are most relevant according to TFIDF. Upon investigation, I find no evidence to suggest that Hermione impacts the perceived trust levels in a chapter. Instead, I find that in two of the highest peaks in *Mean TFIDF of Hermione Qualities over Time*, book 5 chapter 33 and book 7 chapter 14, have relatively high levels of fear.



33.000000	0.001968	0.000849	0.001479	0.002493	0.000993	0.001460	0.000826	0.001119	-0.001730



14.000000	0.001754	0.000835	0.001429	0.002257	0.001034	0.001777	0.000834	0.000872	-0.001536

The first image is the sentiment breakdown of book 5 chapter 33, and the second image is the sentiment breakdown of book 7 chapter 14. The fourth column denotes fear associated with a chapter in a given book. I do not believe this means fear exists in the chapter because of Hermione, but I do anticipate that Hermione is helpful and reliable in fearsome situations, so her importance is often amplified in tense circumstances.

5 Interpretation

From a baseline understanding of *Harry Potter*, I know that Hermione is one of the most influential characters in the series. Exploratory text analysis also indicates that Hermione may be one of the most important characters, after Harry Potter himself. Hermione's paramount role in guiding and aiding Harry throughout his time at Hogwarts and beyond implies that female strength and intellect can be a strong backbone for success, as seen when Harry finally defeats Voldemort.

First, I find that important books and chapters according to TFIDF do not correspond with chapters where Hermione and her associated qualities are most important. In other words, overall importance and Hermione-specific importance (according to TFIDF) are inversely related. An inverse relationship in importance is interesting to note because it suggests that Hermione does not drive intrigue throughout the series. In fact, Hermione's TFIDF measure typically displays a negative slope between perceived important chapters and the preceding chapter; Hermione peaks in importance leading up to interesting moments, but in the actual interesting moment she takes a supporting role. This pattern might suggest a subtle, humble form of importance throughout the series.

Hermione's character further embraces the subtle yet strong supporting role in terms of novel correlation. Based on cosine similarity, I find the top three most similar book pairs to be books 2 and 7, books 1 and 6, and books 2 and 3. Recall that according to TFIDF, Hermione is least important in books 1 and 2, and she grows in importance from books 3 to 7. Cosine similarity shows that books where Hermione is considered unimportant (1 and 2) are most similar to books where she peaks in importance (3, 6, and 7). These findings offer many interpretations - maybe Hermione is not relevant in terms of cosine similarity so the correlations mean nothing, but it is also possible that Hermione elevates the plotline of a book to match the energy of books without her. Either way, the fact that cosine similarity correlations often occur between early and late books indicates continuity throughout the series which is a compliment to Rowling's intentionality in writing the series.

The kernel density estimation (KDE) plot represents the density of the words "Hermione" and "Granger" appearing throughout each book of the Harry Potter series. Hermione's full name is most densely concentrated in the center of each book, which is likely where the climax occurs. Her amplified occurrence in the climactic setting implies that Hermione is essential in times of intensity. Likewise, a common sentiment in chapters where Hermione is importance according to TFIDF is fear. This likely suggests that Hermione shines in strenuous moments, again being crucial in times of hardship. Hermoine being essential in intense situations bolsters the idea that feminine strength is pervasive throughout *Harry Potter*.

Word dispersion plots represent the use of a given word through the timeline of a novel, in this case a series. I surmise that consistent word appearance is an indication of word relevance and importance. "Hermione" is one such word that appears incessantly after she is first introduced, which serves as another testament to her importance. Hermione's last name, "Granger" is used more infrequently. "Granger" is more common in the beginning of the series, but dwindles as time proceeds. The fading of mention of "Granger" indicates that Hermione continually becomes more informal and intimate with the main character, Harry, over time because there is less of a need to mention her last name. After about a third of the series as passed, she has likely become so integral to the plot that only her first name is required for recognition - this observation matches the low TFIDF metrics for Hermione and her associated qualities in the first two books. In addition to "Hermione" and "Granger," I develop word dispersion plots for the terms "bossy" and "genius," which are two differing but true traits of the female character. Use of the term "bossy" is infrequent, and not much can be concluded based on so few observations. "Genius" is also used infrequently, but the term appears in clusters through the series, rather than continually. Plot clusters of the term "genius" suggest that Hermione's intelligence is most appreciated in spurts throughout the Harry Potter series (potentially during times of distress), and it may indicate that her intellect is underappreciated in times of comfort.

Hermione further fills a supporting role to Harry as seen in the Rowling tSNE plot. Hermione's first name is located in a small cluster with other main character names, including "Harry" and "Ron." Her association in this grouping solidifies how crucial she is to the series, just like Harry and Ron. "Granger" is most closely clustered with "Potter," and no other character last names are found in this cluster. Based on this observation, Hermione is semantically related to Harry in a special way. Last names being semantically related might be caused by co-occurence in the Hogwarts school setting, where pupils are referred to by their last names. The relation between "Granger" and "Potter" is likely due to the nature of their classroom relationship, where Hermione is constantly helping Harry succeed. The semantic relationship between Hermione and the main character, Harry, illustrates that strong, intelligent women can often be helpers and motivators of success.

Lastly, Latent Dirichlet Allocation (LDA) topic modeling solidifies the evidence that Hermione is essential to *Harry Potter*. I have found through previous analysis that intelligence is highly valued in the series and T09 emphasizes this point. As mentioned, T09 is the second most frequent topic across books and chapters, and it is thematically associated with Hogwarts schooling. Hermione is the only character name that appears in the top fifteen words of T09 - in fact, Hermione is the only character that appears in the top fifteen words out of all ten displayed topics. Her presence in T09 suggests that Hermione is crucial to Hogwarts and the learning that occurs there. It is possible that her name is important to the topic because she is the most exemplary student at Hogwarts during that time, or she may be integral to all learning that occurs in the series. No matter the interpretation on Hermione's existence in T09, it is irrefutable that Hermione is a key character that stands for intelligence and cleverness.

As anticipated, I find strong female presence in *Harry Potter* through Hermione Granger. She proves to be a valuable, but humble character through TFIDF measures, she appears at the climax of each book, she is semantically related to Harry Potter himself, and she is the only name to appear in the top fifteen words of all ten machine-generated topics. In the plotline, word dispersion plots, and LDA topic model, Hermione is demonstrated as a main character of great intellect. Based on the exploratory text analysis, it seems that Rowling is making a point that women are capable, intelligent, and independent beings. Although this deducation sounds like common sense, it likely represents affirmations that Rowling repeatedly made for herself as she experienced the traumas associated with her life. Women like Hermione are brave, smart, helpful, and good friends. Rowling emphasizes these values through Hermione Granger, offering insight into the type woman toward which Rowling herself was striving.