Final Project Notebook

DS 5001 Exploratory Text Analytics | Spring 2024

Metadata

- Full Name: Chaitali Harge
- Userid: afj8am
- GitHub Repo URL: https://github.com/ChaitaliH/Text-Analysis-of-Harry-Potter-Books
- UVA Box URL: https://virginia.app.box.com/folder/262017511177

Overview

The goal of the final project is for you to create a **digital analytical edition** of a corpus using the tools, practices, and perspectives you've learning in this course. You will select a corpus that has already been digitized and transcribed, parse that into an F-compliant set of tables, and then generate and visualize the results of a series of fitted models. You will also draw some tentative conclusions regarding the linguistic, cultural, psychological, or historical features represented by your corpus. The point of the exercise is to have you work with a corpus through the entire pipeline from ingestion to interpretation.

Specifically, you will acquire a collection of long-form texts and perform the following operations:

- **Convert** the collection from their source formats (F0) into a set of tables that conform to the Standard Text Analytic Data Model (F2).
- **Annotate** these tables with statistical and linguistic features using NLP libraries such as NLTK (F3).
- **Produce** a vector representation of the corpus to generate TFIDF values to add to the TOKEN (aka CORPUS) and VOCAB tables (F4).
- **Model** the annotated and vectorized model with tables and features derived from the application of unsupervised methods, including PCA, LDA, and word2vec (F5).
- **Explore** your results using statistical and visual methods.
- **Present** conclusions about patterns observed in the corpus by means of these operations.

When you are finished, you will make the results of your work available in GitHub (for code) and UVA Box (for data). You will submit to Gradescope (via Canvas) a PDF version of a Jupyter notebook that contains the information listed below.

Some Details

- Please fill out your answers in each task below by editing the markdown cell.
- Replace text that asks you to insert something with the thing, i.e. replace `(INSERT IMAGE HERE)` with an image element, e.g. `![](image.png)`.
- For URLs, just paste the raw URL directly into the text area. Don't worry about providing link labels using `[label](link)`.
- Please do not alter the structure of the document or cell, i.e. the bulleted lists.
- You may add explanatory paragraphs below the bulleted lists.
- Please name your tables as they are named in each task below.
- Tasks are indicated by headers with point values in parentheses.

Raw Data

Source Description (1)

The source material utilized for this analysis comprises the text content of all seven books from the renowned Harry Potter series, organized into CSV files. The dataset includes structured information such as book titles and chapter names, with each row corresponding to a text within the series. Acquired from a online Github repository specializing in literary datasets, the content offers a comprehensive glimpse into the enchanting world crafted by author J.K. Rowling. Through the magical narrative arc spanning across the books, readers are transported into the captivating realm of Hogwarts School of Witchcraft and Wizardry, where young wizard Harry Potter and his companions embark on thrilling adventures, confront dark forces, and unravel the mysteries of the wizarding world. With its richly detailed characters, intricate plotlines, and imaginative settings, the Harry Potter series has captivated audiences of all ages, making it a compelling source for exploration and analysis.

Whole series consists of following 7 books:

1. Harry Potter and the Philosopher's Stone

2. Harry Potter and the Chamber of Secrets
3. Harry Potter and the Prisoner of Azkaban
4. Harry Potter and the Goblet of Fire
5. Harry Potter and the Order of the Phoenix
6. Harry Potter and the Half-Blood Prince
7. Harry Potter and the Deathly Hallows

Source Features (1)

Add values for the following items. (Do this for all following bulleted lists.)

- Source URL: https://github.com/gastonstat/harry-potter-data/tree/main
- UVA Box URL: https://virginia.app.box.com/folder/262017511177
- Number of raw documents: 1
- Total size of raw documents (e.g. in MB): 9.6 MB
- File format(s), e.g. XML, plaintext, etc.: CSV

Source Document Structure (1)

The dataset comprises individual chapters from the Harry Potter series, each encapsulated within a structured format. Each entry in the CSV file contains three key elements: "text", "book", and "chapter". The "text" field encapsulates the narrative content, dialogues, and descriptive passages of each chapter, serving as the primary focus for analysis. The "book" field identifies the corresponding title of the Harry Potter book to which the chapter belongs, offering contextual information about the overarching storyline. Meanwhile, the "chapter" field provides a distinct identifier or name for each chapter within its respective book, facilitating organization and navigation. Together, these elements establish a cohesive internal structure, enabling systematic exploration and understanding of the textual content across the Harry Potter series.

Parsed and Annotated Data

Parse the raw data into the three core tables of your addition: the `LIB`, `CORPUS`, and `VOCAB` tables.

These tables will be stored as CSV files with header rows.

You may consider using `|` as a delimitter.

Provide the following information for each.

LIB (2)

The source documents the corpus comprises. These may be books, plays, newspaper articles, abstracts, blog posts, etc.

Note that these are *not* documents in the sense used to describe a bag-of-words representation of a text, e.g. chapter.

- UVA Box URL: https://virginia.app.box.com/folder/262017511177
- GitHub URL for notebook used to create: https://github.com/ChaitaliH/Text-Analysis-of-Harry-Potter-Books
- Delimitter: ','
- Number of observations: 7
- List of features, including at least three that may be used for model summarization (e.g. date, author, etc.): year, title, book_id
- Average length of each document in characters: 895711 characters

CORPUS (2)

The sequence of word tokens in the corpus, indexed by their location in the corpus and document structures.

- UVA Box URL: https://virginia.app.box.com/folder/262017511177
- GitHub URL for notebook used to create: https://github.com/ChaitaliH/Text-Analysis-of-Harry-Potter-Books
- Delimitter: ','
- Number of observations Between (should be >= 500,000 and <= 2,000,000 observations.): 1098761
- OHCO Structure (as delimitted column names): 'book_id', 'chap_num', 'sent_num', 'token_num'
- Columns (as delimitted column names, including `token_str`, `term_str`, `pos`, and `pos_group`): 'book_id', 'chap_num', 'sent_num', 'token_num', 'token_str', 'term_str', 'pos_tuple', 'pos' and 'pos_group'

VOCAB (2)

The unique word types (terms) in the corpus.

- UVA Box URL: https://virginia.app.box.com/folder/262017511177
- GitHub URL for notebook used to create: https://github.com/ChaitaliH/Text-Analysis-of-Harry-Potter-Books
- Delimitter: ','
- Number of observations: 24600
- Columns (as delimitted names, including `n`, `p`', `i`, `dfidf`, `porter_stem`, `max_pos` and `max_pos_group`, `stop`): 'term_str', 'n', 'max_pos', 'dfidf', 'p', 'i'
- Note: Your VOCAB may contain ngrams. If so, add a feature for `ngram_length`.
- List the top 20 significant words in the corpus by DFIDF.

Top 20 significant words by DFIDF:
'distant', 'eagerly', 'impatiently', 'send', 'deal', 'poor', 'terrified', 'sideways', 'control', 'particularly', 'somehow', 'lupin', 'asking', 'screamed', 'fall', 'information','class', 'placed', 'potion', 'dunno'

Derived Tables

BOW (3)

A bag-of-words representation of the CORPUS.

- UVA Box URL: https://virginia.app.box.com/folder/262017511177
- GitHub URL for notebook used to create: https://virginia.app.box.com/folder/262017511177
- Delimitter: ','
- Bag (expressed in terms of OHCO levels): Book_id, Chap_num
- Number of observations: 283480
- Columns (as delimitted names, including `n`, `tfidf`): 'book_id', 'chap_num', 'term_str', 'n'

DTM (3)

A represenation of the BOW as a sparse count matrix.

- UVA Box URL: https://virginia.app.box.com/folder/262017511177
- UVA Box URL of BOW used to generate (if applicable): https://virginia.app.box.com/file/1520819864632
- GitHub URL for notebook used to create:https://virginia.app.box.com/folder/262017511177
- Delimitter: ','
- Bag (expressed in terms of OHCO levels): book_id, chap_num

TFIDF (3)

A Document-Term matrix with TFIDF values.

- UVA Box URL: https://virginia.app.box.com/folder/262017511177
- UVA Box URL of DTM or BOW used to create: https://virginia.app.box.com/file/1520819864632
- GitHub URL for notebook used to create: https://virginia.app.box.com/folder/262017511177
- Delimitter: ','
- Description of TFIDIF formula ($LaTeX$ OK):

$$\text{TF-IDF} = \text{TF} \times \text{IDF}$$

TF (Term Frequency) is calculated based on 'max' method. IDF (Inverse Document Frequency) is calculated based on 'standard' method. Then the TF-IDF score is computed for each term in the document-term matrix.

Reduced and Normalized TFIDF_L2 (3)

A Document-Term matrix with L2 normalized TFIDF values.

- UVA Box URL: https://virginia.app.box.com/folder/262017511177
- UVA Box URL of source TFIDF table: https://virginia.app.box.com/file/1520802882728
- GitHub URL for notebook used to create: https://virginia.app.box.com/folder/262017511177
- Delimitter: ','
- Number of features (i.e. significant words): 1000
- Principle of significant word selection: L2 normalization (Using Euclidean distance)

Models

PCA Components (4)

- UVA Box URL: https://virginia.app.box.com/folder/262017511177
- UVA Box URL of the source TFIDF_L2 table: https://virginia.app.box.com/file/1520822322815
- GitHub URL for notebook used to create: https://virginia.app.box.com/folder/262017511177
- Delimitter: ','
- Number of components: 10
- Library used to generate: scipy, NumPy
- Top 5 positive terms for first component: c, eaters, bill, albus, eater
- Top 5 negative terms for second component: uncle, aunt, dursleys, kitchen, mrs

PCA DCM (4)

The document-component matrix generated.

- UVA Box URL: https://virginia.app.box.com/folder/262017511177
- GitHub URL for notebook used to create: https://virginia.app.box.com/folder/262017511177
- Delimitter: ','

PCA Loadings (4)

The component-term matrix generated.

- UVA Box URL: https://virginia.app.box.com/folder/262017511177
- GitHub URL for notebook used to create: https://virginia.app.box.com/folder/262017511177
- Delimitter: ','

PCA Visualization 1 (4)

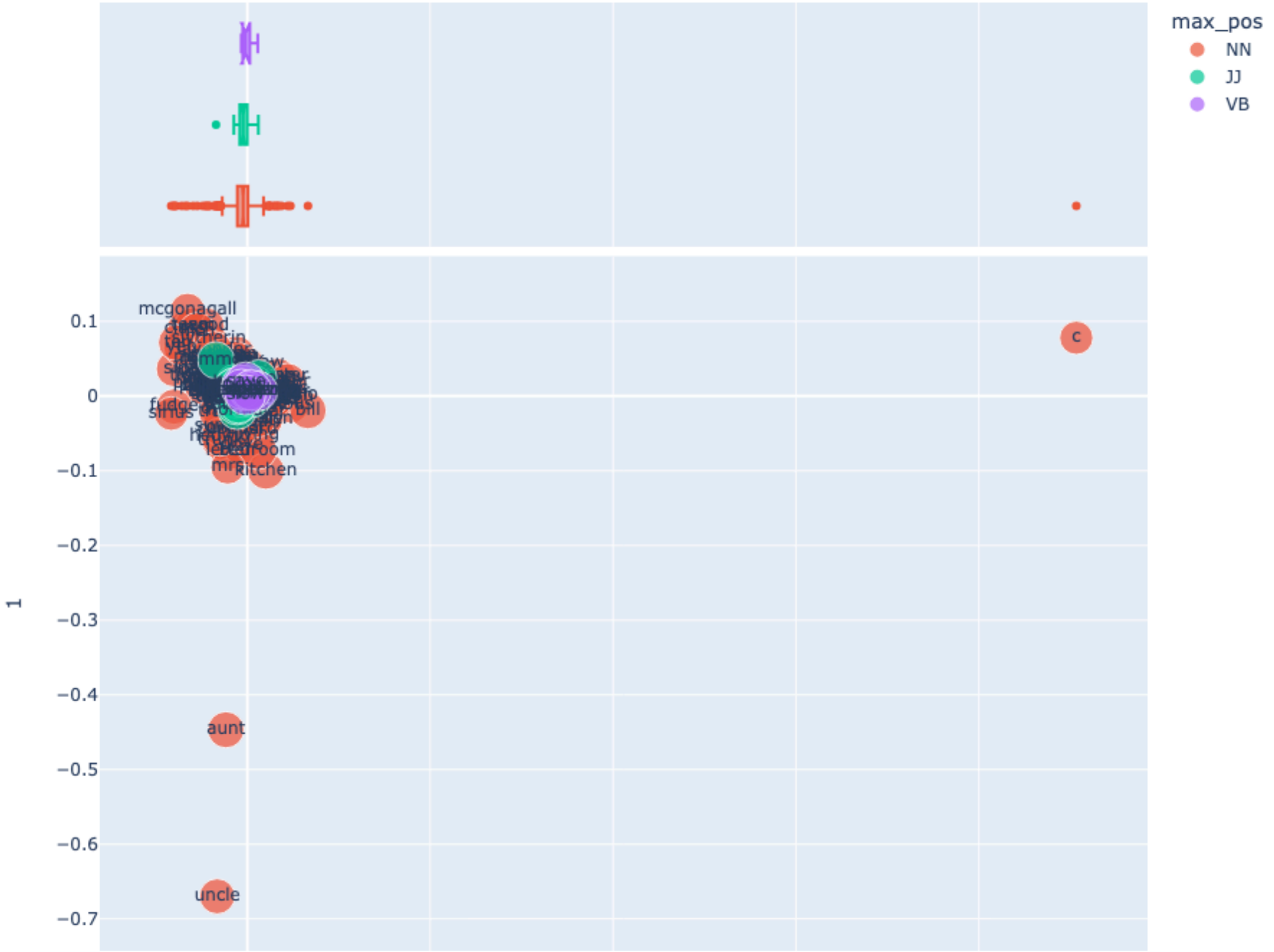
Include a scatterplot of documents in the space created by the first two components.

Color the points based on a metadata feature associated with the documents.

Also include a scatterplot of the loadings for the same two components. (This does not need a feature mapped onto color.)


title
Philosopher's Stone
Chamber of Secrets
Prisoner of Azkaban
Goblet of Fire
Order of the Phoenix
Half-Blood Prince
Deathly Hallows
1
0.2
0
−0.2
−0.4
−0.6
−0.8


−0.2
0
0.2
0.4
0.6
0.8
0


max_pos
NN
JJ
VB
1
0.1
0
−0.1
−0.2
−0.3
−0.4
−0.5
−0.6
−0.7
mcgonagall
c
aunt
uncle
kitchen

0 0.2 0.4 0.6 0.8

0

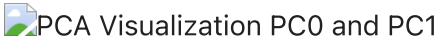
Description:
In the PCA visualization of PC0 and PC1, book "Deathly Hallows" has the wide range and it is way far than other books. Also this book has highest absolute value, and it also shows biggest quantile range compared to other books. The fact that the book "Deathly Hallows" appears to have a wider range in the PCA visualization suggests that it exhibits greater variability or distinctiveness in terms of the language or topics covered compared to other books. This could indicate that "Deathly Hallows" contains a diverse range of themes or chapters that contribute significantly to the variation captured by PC0 and PC1. In the loadings plot, words appearing cluttered and overlapping each other at the center may indicate that these words are common across multiple books and do not contribute significantly to distinguishing one book from another in the PCA space.

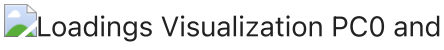
PCA Visualization 2 (4)

Include a scatterplot of documents in the space created by the second two components.

Color the points based on a metadata feature associated with the documents.

Also include a scatterplot of the loadings for the same two components. (This does not need a feature mapped onto color.)

 PCA Visualization PC0 and PC1

 Loadings Visualization PC0 and PC1

Briefly describe the nature of the polarity you see in the second component:

Description: In PCA visualization of PC7 and PC9, the book "Goblet of Fire" has the wide range of quantile. Books "Half Blood Prince" and "Prisoner of Azkaban" also show somewhat wide range. Wide Range of Quantile in PCA Visualization: The observation that the book "Goblet of Fire" exhibits the widest range of quantile in the PCA visualization suggests that it has a high variability or diversity of content along PC7 and PC9. This could indicate that "Goblet of Fire" covers a broad range of themes or topics that contribute significantly to the variation captured by these principal components. Similarly, "Half Blood Prince" and "Prisoner of Azkaban" also show somewhat wide ranges, indicating their diverse content compared to other books in the dataset.

In the loadings plot, the widest range observed for nouns suggests that nouns play a significant role in distinguishing the books along PC7 and PC9. This could imply that the occurrence or usage of nouns contributes the most to the variation observed in these principal components.

LDA TOPIC (4)

- UVA Box URL: https://virginia.app.box.com/folder/262017511177
- UVA Box URL of count matrix used to create: https://virginia.app.box.com/file/1520821933747
- GitHub URL for notebook used to create: https://virginia.app.box.com/folder/262017511177
- Delimitter: ','
- Libary used to compute: sklearn, scikitlearn
- A description of any filtering, e.g. POS (Nouns and Verbs only): NNP are excluded
- Number of components: 40
- Any other parameters used: ngram_range = (1, 2) , n_terms = 4000, n_topics = 40, max_iter = 20, n_top_terms = 9
- Top 5 words and best-guess labels for topic five topics by mean document weight:
  - T15: harry said ron hermione professor
  - T39: harry said ron hermione know
  - T17: harry hermione said wand ron
  - T28: said harry weasley mrs mrs
  - T34: harry uncle vernon uncle vernon

LDA THETA (4)

- UVA Box URL: https://virginia.app.box.com/folder/262017511177
- GitHub URL for notebook used to create: https://virginia.app.box.com/folder/262017511177
- Delimitter: ','

LDA PHI (4)

- UVA Box URL: https://virginia.app.box.com/folder/262017511177
- GitHub URL for notebook used to create: https://virginia.app.box.com/folder/262017511177

- Delimitter: ','

LDA + PCA Visualization (4)

Apply PCA to the PHI table and plot the topics in the space opened by the first two components.

Size the points based on the mean document weight of each topic (using the THETA table).

Color the points basd on a metadata feature from the LIB table.

Provide a brief interpretation of what you see.

Note: After consultation with Professor Alvarado, this subsection is eliminated from report.

Sentiment VOCAB_SENT (4)

Sentiment values associated with a subset of the VOCAB from a curated sentiment lexicon.

- UVA Box URL: https://virginia.app.box.com/folder/262017511177
- UVA Box URL for source lexicon: https://virginia.app.box.com/file/1520808091145
- GitHub URL for notebook used to create: https://virginia.app.box.com/folder/262017511177
- Delimitter: ','

Sentiment BOW_SENT (4)

Sentiment values from VOCAB_SENT mapped onto BOW.

- UVA Box URL: https://virginia.app.box.com/folder/262017511177
- GitHub URL for notebook used to create: https://virginia.app.box.com/folder/262017511177
- Delimitter: ','

Sentiment DOC_SENT (4)

Computed sentiment per bag computed from BOW_SENT.

- UVA Box URL: https://virginia.app.box.com/folder/262017511177
- GitHub URL for notebook used to create: https://virginia.app.box.com/folder/262017511177
- Delimitter: ','
- Document bag expressed in terms of OHCO levels: book_id, chap_num

Sentiment Plot (4)

Plot sentiment over some metric space, such as time.

If you don't have a metric metadata features, plot sentiment over a feature of your choice.

You may use a bar chart or a line graph.

Description:
This sentiment plot depicts the emotional trajectory of "Book 7: Deathly Hallows." One striking observation is the prominent positioning of the 'fear' line consistently at the top, indicating a prevailing sense of apprehension and tension that permeates the narrative from start to finish. The sentiment line exhibits notable fluctuations throughout the book, suggesting dynamic shifts in mood and tone as the story unfolds. This visualization offers valuable insights into the emotional landscape of the narrative, capturing the ebb and flow of sentiments that characterize the journey depicted within "Deathly Hallows."

VOCAB_W2V (4)

A table of word2vec features associated with terms in the VOCAB table.

- UVA Box URL: https://virginia.app.box.com/folder/262017511177
- GitHub URL for notebook used to create: https://virginia.app.box.com/folder/262017511177
- Delimitter: ','
- Document bag expressed in terms of OHCO levels: book_id, chap_num
- Number of features generated: 246

- The library used to generate the embeddings: gensim

Word2vec tSNE Plot (4)

Plot word embedding featues in two-dimensions using t-SNE.

Describe a cluster in the plot that captures your attention.

Description:

Upon examining the tSNE plot above and focusing on specific clusters, it becomes evident that character names from certain books appear to be closely intertwined or overlapping. This suggests a strong connection between characters that are closely related within the narrative. Notably, terms such as 'Dobby', 'Voldemort', 'Sirius', and 'Wormtail' are positioned near each other, indicating their significant presence and interrelation within the book series.

Riffs

Provde at least three visualizations that combine the preceding model data in interesting ways.

These should provide insight into how features in the LIB table are related.

The nature of this relationship is left open to you -- it may be correlation, or mutual information, or something less well defined.

In doing so, consider the following visualization types:

- Hierarchical cluster diagrams
- Heatmaps
- Scatter plots
- KDE plots
- Dispersion plots
- t-SNE plots

- etc.

Riff 1 (5)

Interpretation: Overall, upon examining the sentiment plots across the books from "Philosopher's Stone" to "Deathly Hallows," a noticeable trend emerges: the positivity in the storyline appears to slightly decrease, while negativity gradually increases towards the end. This shift could be attributed to pivotal events such as Voldemort's attempts to kill Harry Potter, resulting in the deaths or sacrifices of several key characters who are dear to him. Consequently, the overarching happiness in the stories does not seem to exhibit an upward trend over time. Additionally, it's observed that the compound sentiment decreases towards the end, indicating a shift towards more negative emotional tones in the narrative's climax.

Riff 2 (5)

Interpretation:

Below is the description for above scatter plots. For First Plot: X-axis: corr_phi - This represents the correlation between topics based on their word distributions (PHI matrix). Y-axis: corr_theta - This represents the correlation between topics based on their document distributions (THETA matrix). Color: phi_cosine - This likely represents some measure of similarity between topics based on their word distributions.

For Second Plot: Color: theta_cosine - This likely represents some measure of similarity between topics based on their word distributions.

Looking at the plots above, it has low PHI correlation and high THETA correlation. It shows the scenario of ideal topic model. Low PHI correlation indicates that the topics are distinct and not redundant. High THETA correlation suggests that topics often co-occur, forming meaningful complexes of meaning.

Riff 3 (5)

Interpretation:

The WordCloud analysis provides valuable insights into the central themes and characters of the Harry Potter books, offering a visual representation of their narrative essence. Prominent characters such as Hermione, Ron, and Dumbledore emerge as central figures in the storyline, indicating their significant roles in Harry Potter's journey. Their frequent appearance in the WordCloud suggests their importance in shaping the plot and influencing the protagonist's experiences. Words like "think" and "time" hint at the intellectual and temporal elements woven into the narrative, reflecting the characters' contemplative moments and the passage of time within the story. Meanwhile, terms like "dark" and "night" evoke the overarching theme of darkness and danger that Harry Potter and his friends must confront throughout their adventures. The WordCloud provides a snapshot of the storyline, offering glimpses into the central conflicts, character dynamics, and thematic elements that define the Harry Potter series. It serves as a visual summary of the books, encapsulating their essence and providing readers with a quick understanding of the narrative landscape.

Interpretation (4)

Describe something interesting about your corpus that you discovered during the process of completing this assignment.

At a minumum, use 250 words, but you may use more. You may also add images if you'd like.

Interpretation:

During the process of completing this assignment, I encountered several intriguing insights about the Harry Potter corpus, which added depth and richness to my analysis. One notable discovery was the evolving sentiment trajectory across the seven books, revealing nuanced shifts in the emotional landscape of the series.

As I delved into the sentiment analysis, I observed a fascinating trend in the sentiment polarity throughout the books. Starting from "The Philosopher's Stone," the sentiment appeared to be predominantly positive, reflecting the innocence and wonder of Harry's introduction to the magical world. However, as the series progressed, particularly in the later books such as "The Deathly Hallows," I noticed a gradual decline in positivity and a corresponding increase in negativity.

This shift in sentiment seemed to mirror the escalating tension and darker themes that permeate the later installments of the series. With Voldemort's return and the looming threat of war, the narrative tone becomes progressively darker, mirroring the characters' growing

anxieties and the escalating stakes of the wizarding world's conflict. Additionally, poignant moments of loss and sacrifice contribute to the overall increase in negative sentiment as the series approaches its climactic conclusion.

Moreover, while negative sentiment may dominate in the later books, it is often intertwined with moments of bravery, resilience, and camaraderie among the characters. This complexity adds depth to the narrative, showcasing the characters' multifaceted emotions and the moral complexities they face in their fight against darkness.

Overall, this exploration of sentiment dynamics offered valuable insights into the emotional arc of the Harry Potter series, highlighting its thematic depth and narrative complexity. It provided a deeper understanding of how sentiment evolves across the books, reflecting the characters' growth, the challenges they confront, and the overarching themes of courage, friendship, and the triumph of good over evil.

In [ ]: