Final Project Notebook

DS 5001 Exploratory Text Analytics | Spring 2024

Metadata

- Full Name: Sydney Mathiason
- Userid: qex8sd
- GitHub Repo URL: https://github.com/sydneymathiason/ds5001_finalproject
- UVA Box URL: https://virginia.box.com/s/pgrapjjiwxjwcfmv709olsaw04brmmm3

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)

Provide a brief description of your source material, including its provenance and content. Tell us where you found it and what kind of content it contains.

The source material for my responses includes the three books from "The Hunger Games" trilogy, written by Suzanne Collins, and the three books from the "Divergent" series, penned by Veronica Roth. These books are popular young adult dystopian novels that have garnered widespread acclaim and have been adapted into successful film franchises.

"The Hunger Games" trilogy is set in a post-apocalyptic nation called Panem, where the ruling Capitol holds an annual event called the Hunger Games, where young participants from each district fight to the death in a televised spectacle.

The "Divergent" series is set in a futuristic Chicago where society is divided into factions based on personality traits. The story follows Beatrice "Tris" Prior as she navigates the challenges of belonging to multiple factions and uncovering the secrets of her society.

Both series explore themes of oppression, rebellion, and identity, making them compelling reads for audiences of various ages.

I found the text files for these books at https://archive.org/.

Source Features (1)

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

- Source URL: https://archive.org/
- UVA Box URL: https://virginia.box.com/s/d6ucpdxdpsp8yc1kpo1si1qao5pgdp76
- Number of raw documents: 6
- Total size of raw documents (e.g. in MB): 3.7 MB
- File format(s), e.g. XML, plaintext, etc.: plaintext

Source Document Structure (1)

Provide a brief description of the internal structure of each document. That, describe the typical elements found in document and their relation to each other. For example, a corpus of letters might be described as having a date, an addressee, a salutation, a set of content paragraphs, and closing. If they are various structures, state that.

There are six books, each book is made up of chapters, which have paragraphs and sentences within them. I replaced all contractions with n't to not to help adjust for the lone t's showing up in my topic models.

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.box.com/s/5ww9wpueszxyo14av9ers76kygvfesdu
- GitHub URL for notebook used to create: https://github.com/sydneymathiason/ds5001_finalproject/blob/main/CreateLIB.ipynb
- Delimitter: , (comma)
- Number of observations: 6
- List of features, including at least three that may be used for model summarization (e.g. date, author, etc.):
  - 'title', 'author', 'year', 'genre', 'cover color', 'file', 'txt_str','n_char'
- Average length of each document in characters: 600883.16

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.box.com/s/vajiw6iau60ns3y5alcvtvs2u2zidmlx
- GitHub URL for notebook used to create: https://github.com/sydneymathiason/ds5001_finalproject/blob/main/CreateTokenCSV.ipynb
- Delimitter: , (comma)

- Number of observations Between (should be >= 500,000 and <= 2,000,000 observations.): 643777
- OHCO Structure (as delimitted column names): ['book_num', chap_num', 'para_num', 'sent_num', 'token_num']
- Columns (as delimitted column names, including `token_str`, `term_str`, `pos`, and `pos_group`): ['token_str', 'term_str', 'pos', 'pos_group']

VOCAB (2)

The unique word types (terms) in the corpus.

- UVA Box URL: https://virginia.box.com/s/q07uauzsx9a71ecuwlzp35oa1rt4blgf
- GitHub URL for notebook used to create: https://github.com/sydneymathiason/ds5001_finalproject/blob/main/CreateVOCAB.ipynb
- Delimitter: , (comma)
- Number of observations: 18111
- Columns (as delimitted names, including `n`, `p `', `i`, `dfidf`, `porter_stem`, `max_pos` and `max_pos_group`, `stop`):
  - ['n', 'n_chars', 'p', 'i', 'max_pos', 'max_pos_group', 'stop','porter_stem', 'dfidf']
- 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.

['catch', 'lift', 'somewhere', 'twelve', 'true', 'moving', 'ones', 'having', 'deep', 'went', 'gray', 'presses', 'fast', 'wish', 'found', 'throw', 'near', 'high', 'pass', 'lean']

Derived Tables

BOW (3)

A bag-of-words representation of the CORPUS.

- UVA Box URL: https://virginia.box.com/s/xn498n0t1htc5xd6wz4c5sqghu42dgbq
- GitHub URL for notebook used to create: https://github.com/sydneymathiason/ds5001_finalproject/blob/main/CreateVocab.ipynb
- Delimitter: , (comma)
- Bag (expressed in terms of OHCO levels): ['chap_num']
- Number of observations: 187554
- Columns (as delimitted names, including `n`, `tfidf`):

DTM (3)

A represenation of the BOW as a sparse count matrix.

- UVA Box URL: https://virginia.box.com/s/9h2i0exvgbrpv6f6jwds4lkrlla9p9ju

- UVA Box URL of BOW used to generate (if applicable): https://virginia.box.com/s/xn498n0t1htc5xd6wz4c5sqghu42dgbq
- GitHub URL for notebook used to create: https://github.com/sydneymathiason/ds5001_finalproject/blob/main/CreateVocab.ipynb
- Delimitter: , (comma)
- Bag (expressed in terms of OHCO levels): ['chap_num']

TFIDF (3)

A Document-Term matrix with TFIDF values.

- UVA Box URL: https://virginia.box.com/s/00k9823on565249zv97udl82s1g05gki
- UVA Box URL of DTM or BOW used to create: https://virginia.box.com/s/9h2i0exvgbrpv6f6jwds4lkrlla9p9ju
- GitHub URL for notebook used to create: https://github.com/sydneymathiason/ds5001_finalproject/blob/main/CreateVocab.ipynb
- Delimitter: , (comma)
- Description of TFIDIF formula ($LaTeX$ OK): max TF (DTCM.T / DTCM.T.max()) * standard IDF (np.log2(N / DF))

Reduced and Normalized TFIDF_L2 (3)

A Document-Term matrix with L2 normalized TFIDF values.

- UVA Box URL: https://virginia.box.com/s/cd47lss9bq1czlfcujtq78qp99wl0d4g
- UVA Box URL of source TFIDF table: https://virginia.box.com/s/00k9823on565249zv97udl82s1g05gki
- GitHub URL for notebook used to create: https://github.com/sydneymathiason/ds5001_finalproject/blob/main/CreateVocab.ipynb
- Delimitter: , (comma)
- Number of features (i.e. significant words): 5000
- Principle of significant word selection: top 5000 nouns, verbs, and adjectives, not including proper nouns

Models

PCA Components (4)

- UVA Box URL: https://virginia.box.com/s/gdnqvs1awox0v1dg6gvp1orustxr5d0a
- UVA Box URL of the source TFIDF_L2 table: https://virginia.box.com/s/cd47lss9bq1czlfcujtq78qp99wl0d4g
- GitHub URL for notebook used to create: https://github.com/sydneymathiason/ds5001_finalproject/blob/main/CreateModels.ipynb
- Delimitter: , (comma)

- Number of components: 10
- Library used to generate: created a get_pca function similar to what was covered in Module 07.01, creating from scratch
- Top 5 positive terms for first component: serum gun factionless simulation faction
- Top 5 negative terms for second component: initiates gun jungle knife water

PCA DCM (4)

The document-component matrix generated.

- UVA Box URL: https://virginia.box.com/s/ma2zpe2mph5cqai1v9uo2qk47hoklzhd
- GitHub URL for notebook used to create: https://github.com/sydneymathiason/ds5001_finalproject/blob/main/CreateModels.ipynb
- Delimitter: , (comma)

PCA Loadings (4)

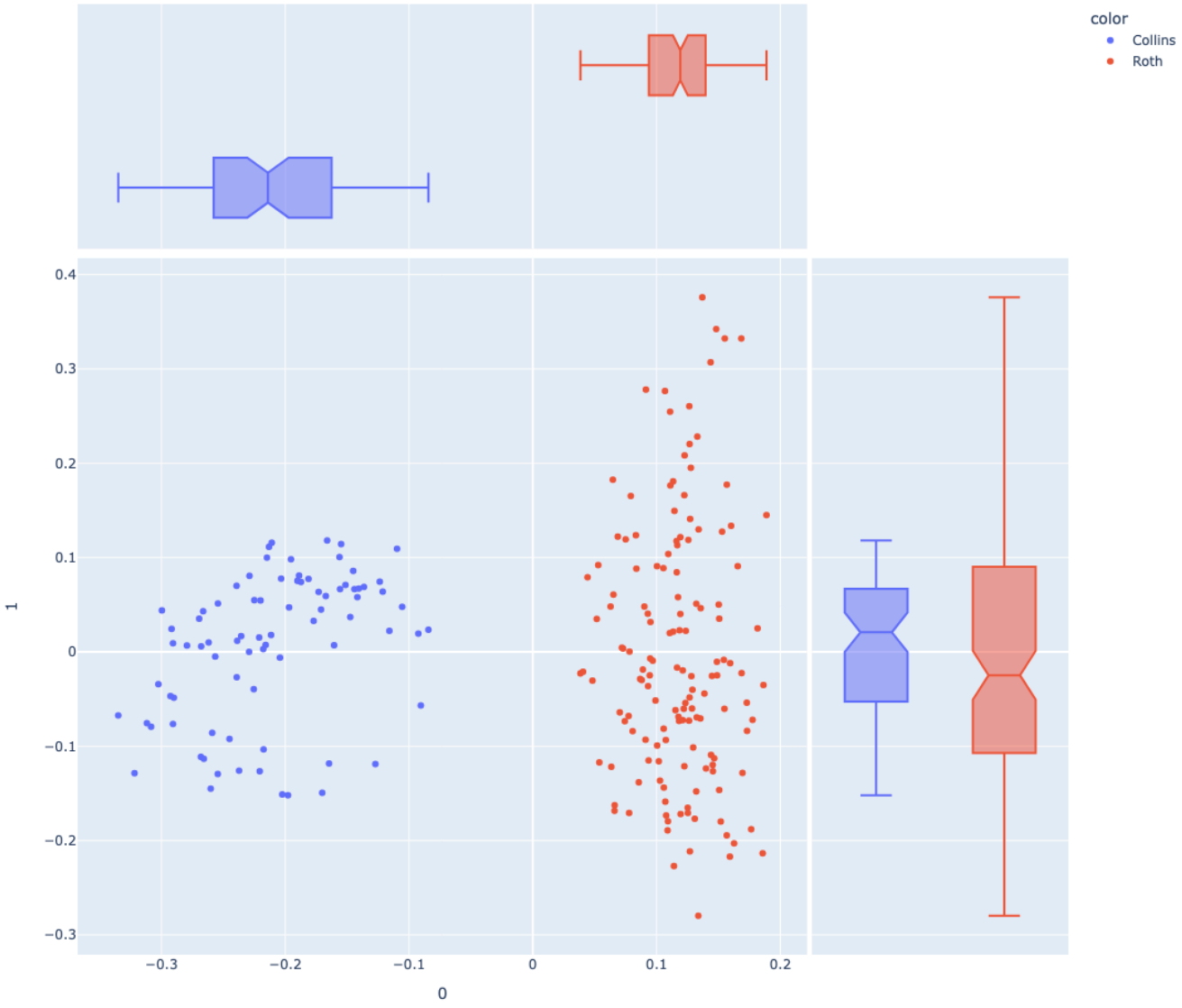
The component-term matrix generated.

- UVA Box URL: https://virginia.box.com/s/d7t8ilh1dlym9on2wjrsucjv4lntns73
- GitHub URL for notebook used to create: https://github.com/sydneymathiason/ds5001_finalproject/blob/main/CreateModels.ipynb
- Delimitter: , (comma)

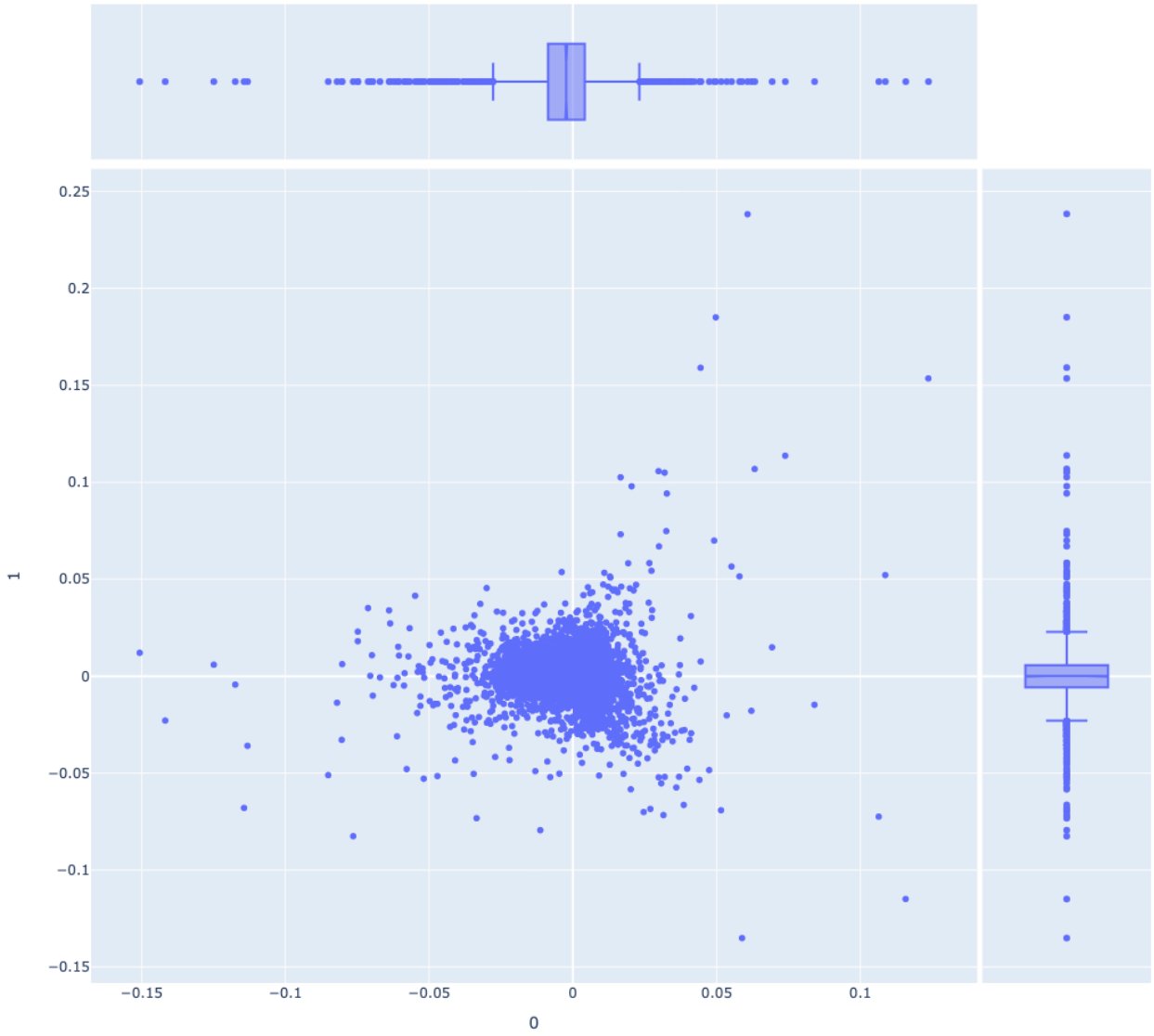
PCA Visualization 1 (4)

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.)



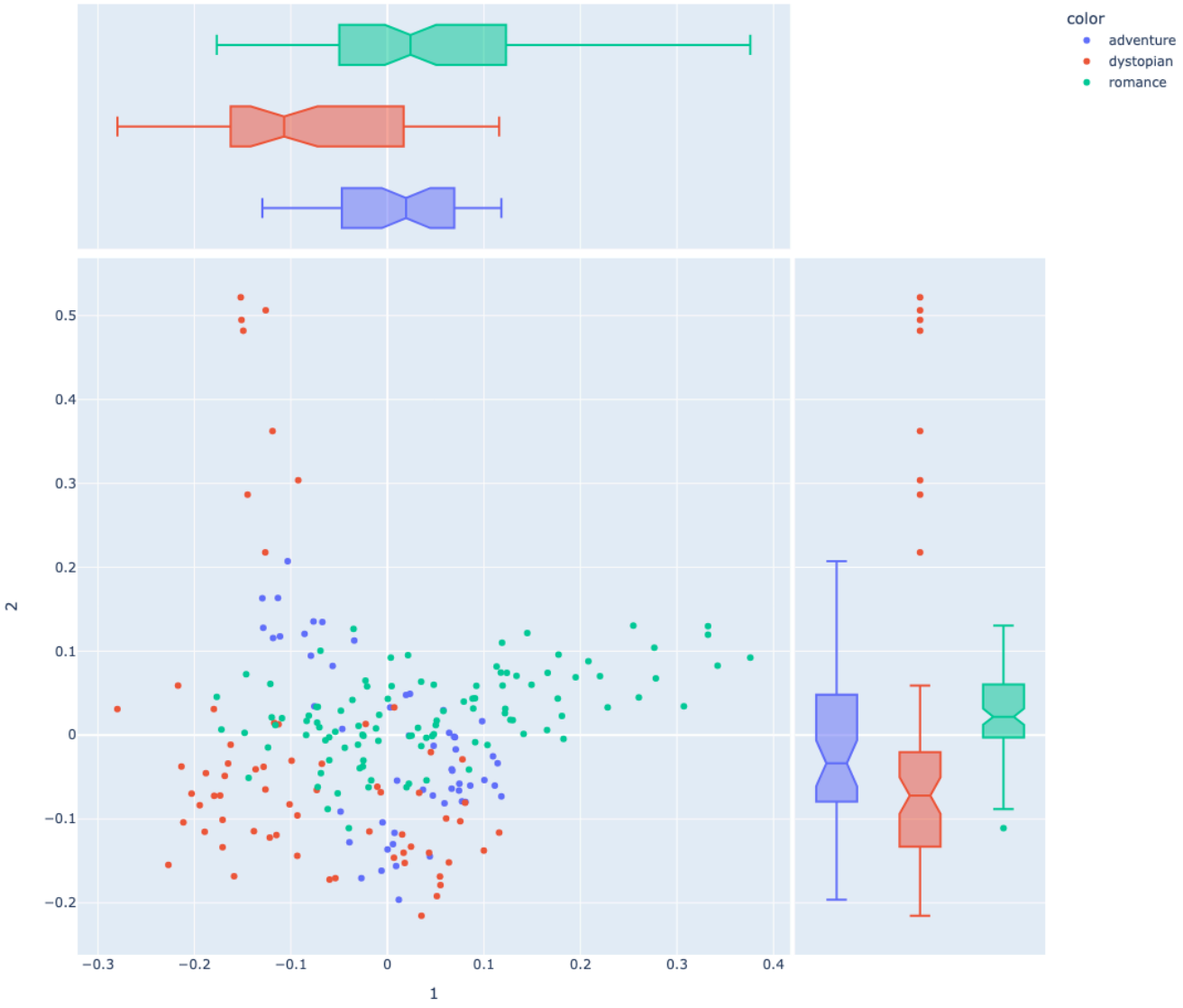
Briefly describe the nature of the polarity you see in the first component:

The first component's polarity is pretty neutral overall, with most points relatively close to zero. But there are some values spread out a little wider in both directions.

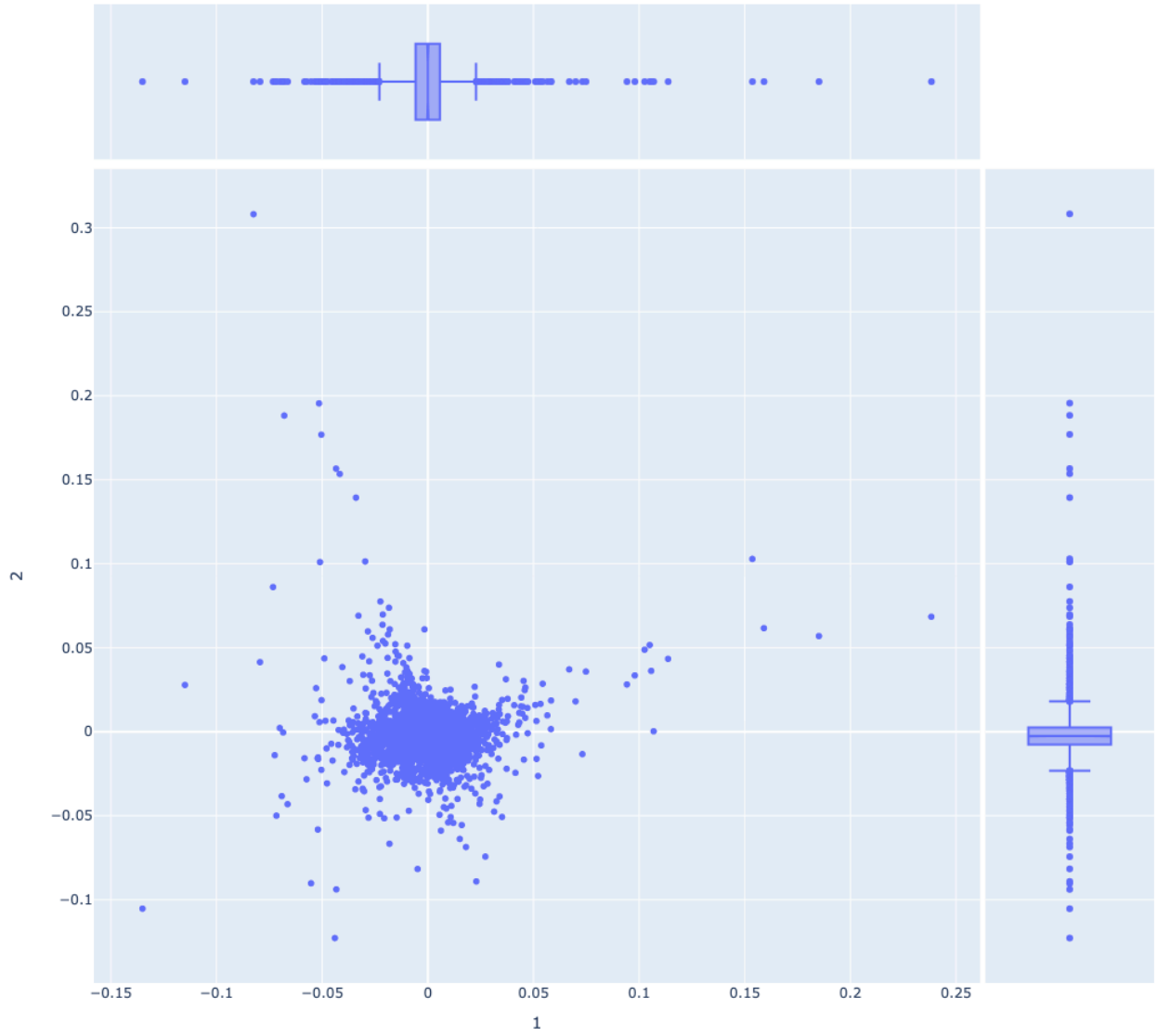
PCA Visualization 2 (4)

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.)



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

In the second component the polarity is still pretty neutral, with a few points spread out in both directions. In this case the positive points go march farther than the negative compoennts.

LDA TOPIC (4)

- UVA Box URL: https://virginia.box.com/s/qaz295j62nupypmsir2uz1km3zq778yt
- UVA Box URL of count matrix used to create: https://virginia.box.com/s/c9ri0xszzybq4x6j83il9io3ozqzl1ly
- GitHub URL for notebook used to create: https://github.com/sydneymathiason/ds5001_finalproject/blob/main/CreateModels.ipynb
- Delimitter: , (comma)
- Libary used to compute: sklearn (LDA)
- A description of any filtering, e.g. POS (Nouns and Verbs only): Limited to just Nouns and Verbs (no proper nouns)
- Number of components: 20

- Any other parameters used: norm_docs=True,norm_level=2,center_by_mean=True, center_by_variance=False
- Top 5 words and best-guess labels for topic five topics by mean document weight:
  - T00: people time mother way eyes : **life's journey**
  - T01: eyes head hand hands way : **sensory**
  - T02: eyes head hand hands face : **body parts**
  - T03: time hand arena head way : **pathways**
  - T04: room time eyes serum people : **perception**

LDA THETA (4)

- UVA Box URL: https://virginia.box.com/s/e5clzkua9ddfa8s92uvkvkzze2b1xw06
- GitHub URL for notebook used to create: https://github.com/sydneymathiason/ds5001_finalproject/blob/main/CreateModels.ipynb
- Delimitter: , (comma)

LDA PHI (4)

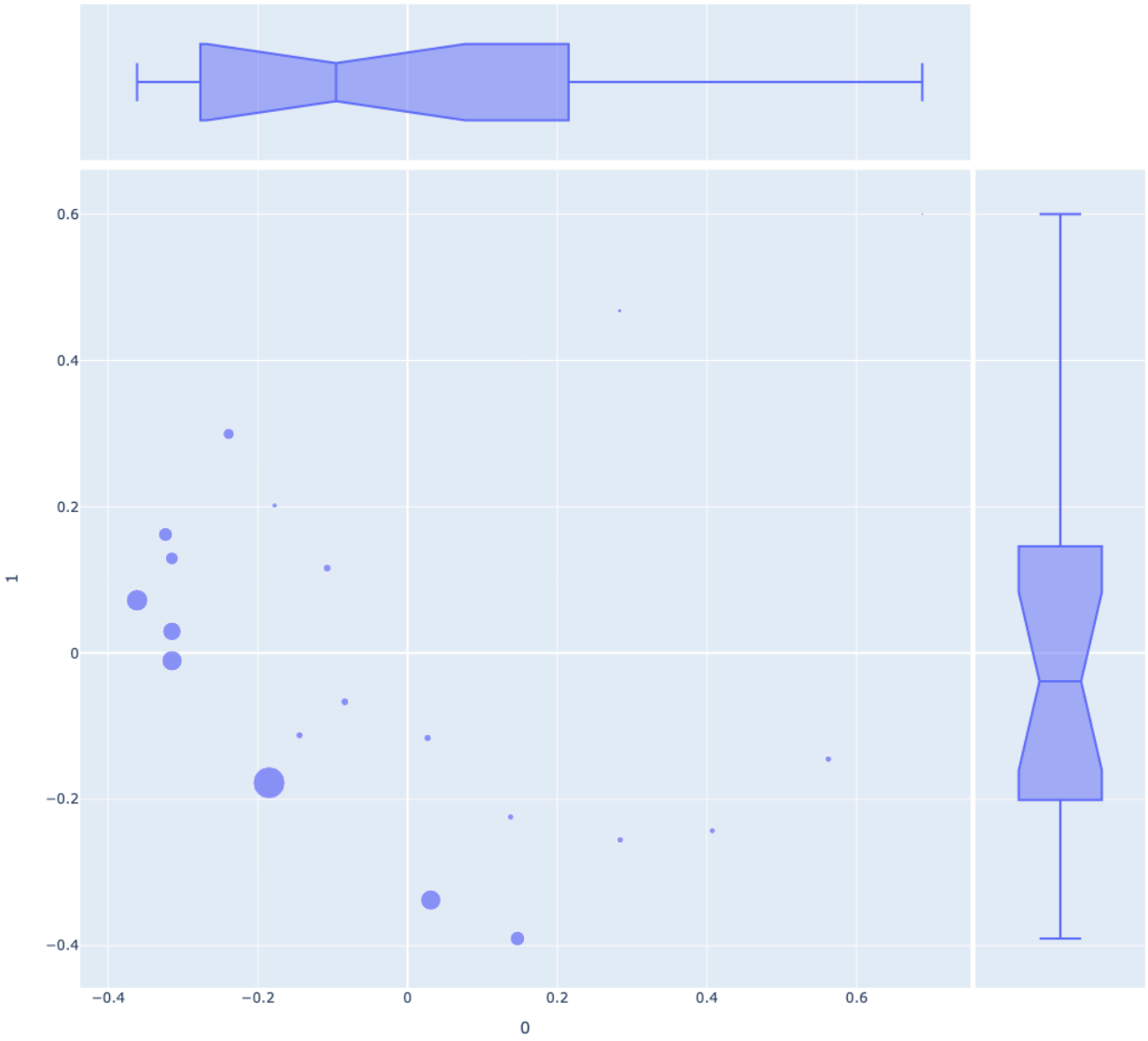
- UVA Box URL: https://virginia.box.com/s/uyyj83oz2auxutk9q9iu4bgi8bm977qp
- GitHub URL for notebook used to create: https://github.com/sydneymathiason/ds5001_finalproject/blob/main/CreateModels.ipynb
- Delimitter: , (comma)

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).

Provide a brief interpretation of what you see.



Most of the points are towards the negative, bottom left corner, while also having two arcs with the larger points being below the smaller points.

Sentiment VOCAB_SENT (4)

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

- UVA Box URL: https://virginia.box.com/s/ph2odnt2ac9fbu1ds7y77460k350gudy
- UVA Box URL for source lexicon: https://virginia.box.com/s/vfvbpjaomiho7fffgipuge1cwtl2x7hk
- GitHub URL for notebook used to create: https://github.com/sydneymathiason/ds5001_finalproject/blob/main/CreateSentiment.ipynb
- Delimitter: , (comma)

Sentiment BOW_SENT (4)

Sentiment values from VOCAB_SENT mapped onto BOW.

- UVA Box URL: https://virginia.box.com/s/xhyynnqyi9pemosx95h898i1ydekt4gc
- GitHub URL for notebook used to create: https://github.com/sydneymathiason/ds5001_finalproject/blob/main/CreateSentiment.ipynb
- Delimitter: , (comma)

Sentiment DOC_SENT (4)

Computed sentiment per bag computed from BOW_SENT.

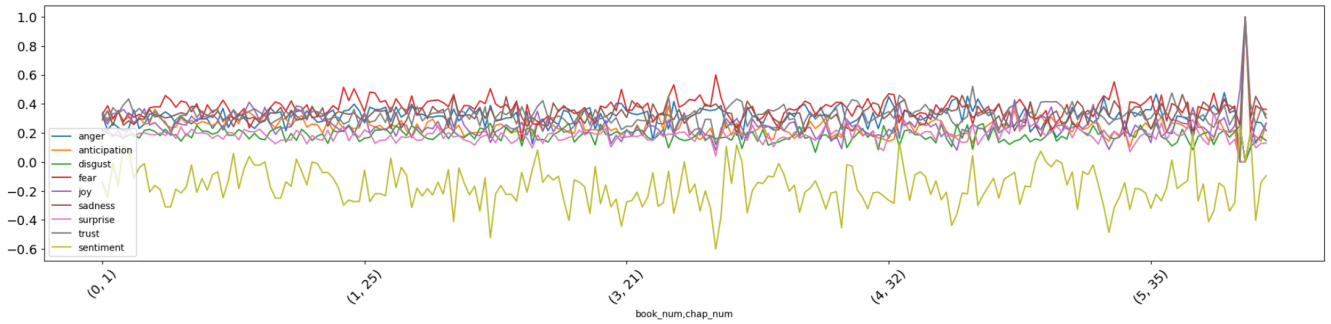
- UVA Box URL: https://virginia.box.com/s/3spj3ywlwp0mx1nejuvkeph91kkmkavm
- GitHub URL for notebook used to create: https://github.com/sydneymathiason/ds5001_finalproject/blob/main/CreateSentiment.ipynb
- Delimitter: , (comma)
- Document bag expressed in terms of OHCO levels: ['book_num', 'chap_num']

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.



VOCAB_W2V (4)

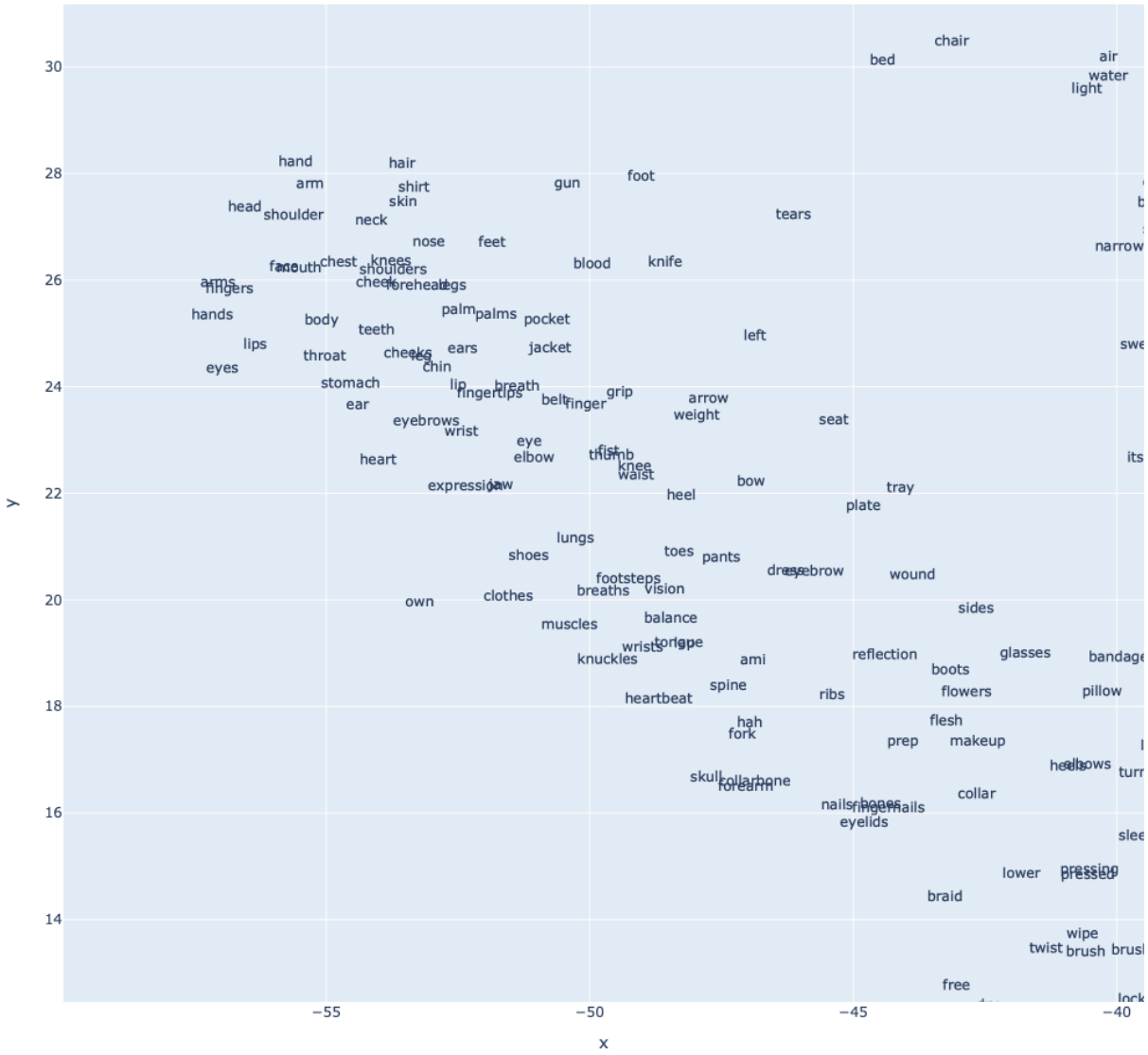
A table of word2vec features associated with terms in the VOCAB table.

- UVA Box URL: https://virginia.box.com/s/999px8fkw3zb0qr14um7wupgxlx8lhf9
- GitHub URL for notebook used to create: https://github.com/sydneymathiason/ds5001_finalproject/blob/main/CreateW2V.ipynb
- Delimitter: , (comma)
- Document bag expressed in terms of OHCO levels: ['book_num', 'chap_num']
- Number of features generated: 256
- 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.



One cluster in the plot that captures my attention is the area that has all of the different bodies parts. Towards the bottom right corner it morphs into what you might put on your body, such as clothes, glasses, and makeup.

Riffs

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

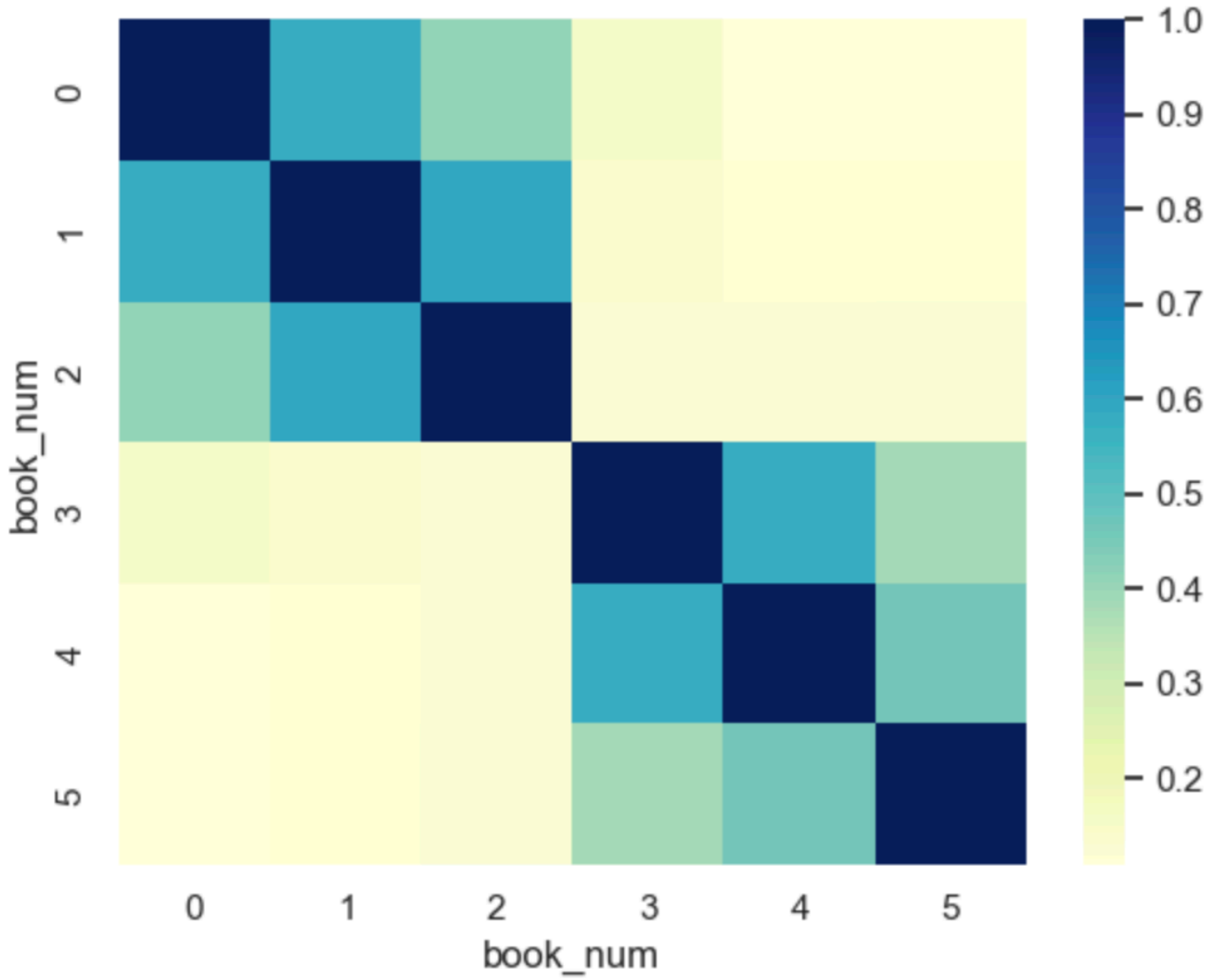
These should provide insight into how features in the LIB table are related.

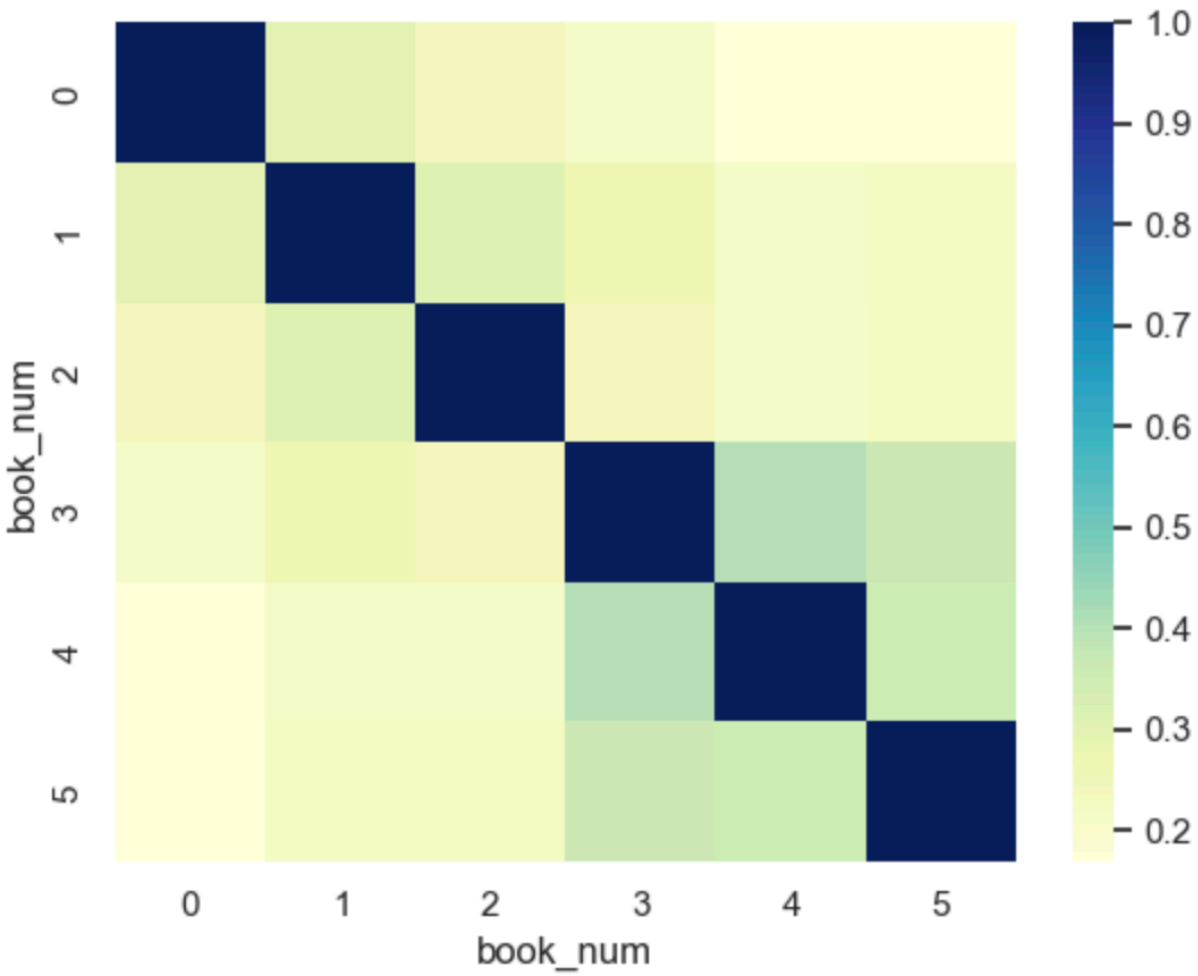
The nature of this relationship is left open to you -- it may be correlation, or mutual information, or something less well defined.

In doing so, consider the following visualization types:

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

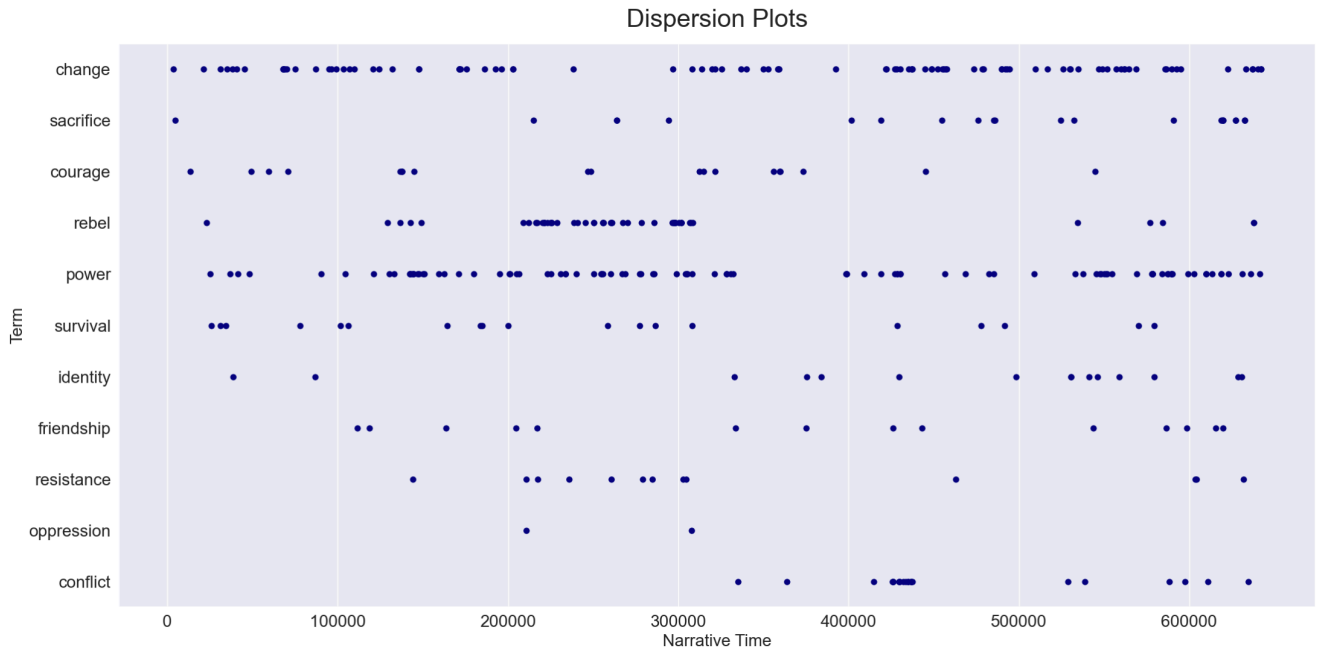
Riff 1 (5)





For this I looked at the correlation between the books and I found it interesting that with the different correlation types there were different levels of correlation. For example with `pearson` (image 1) it is very clear that both series (0-2: Hunger Games) and (3-5: Divergent) are highly correlated with the others books in the series. But for `kendell` (image 2) the Divergent books are very correlated with one another whereas the Hunger Games books are less so.

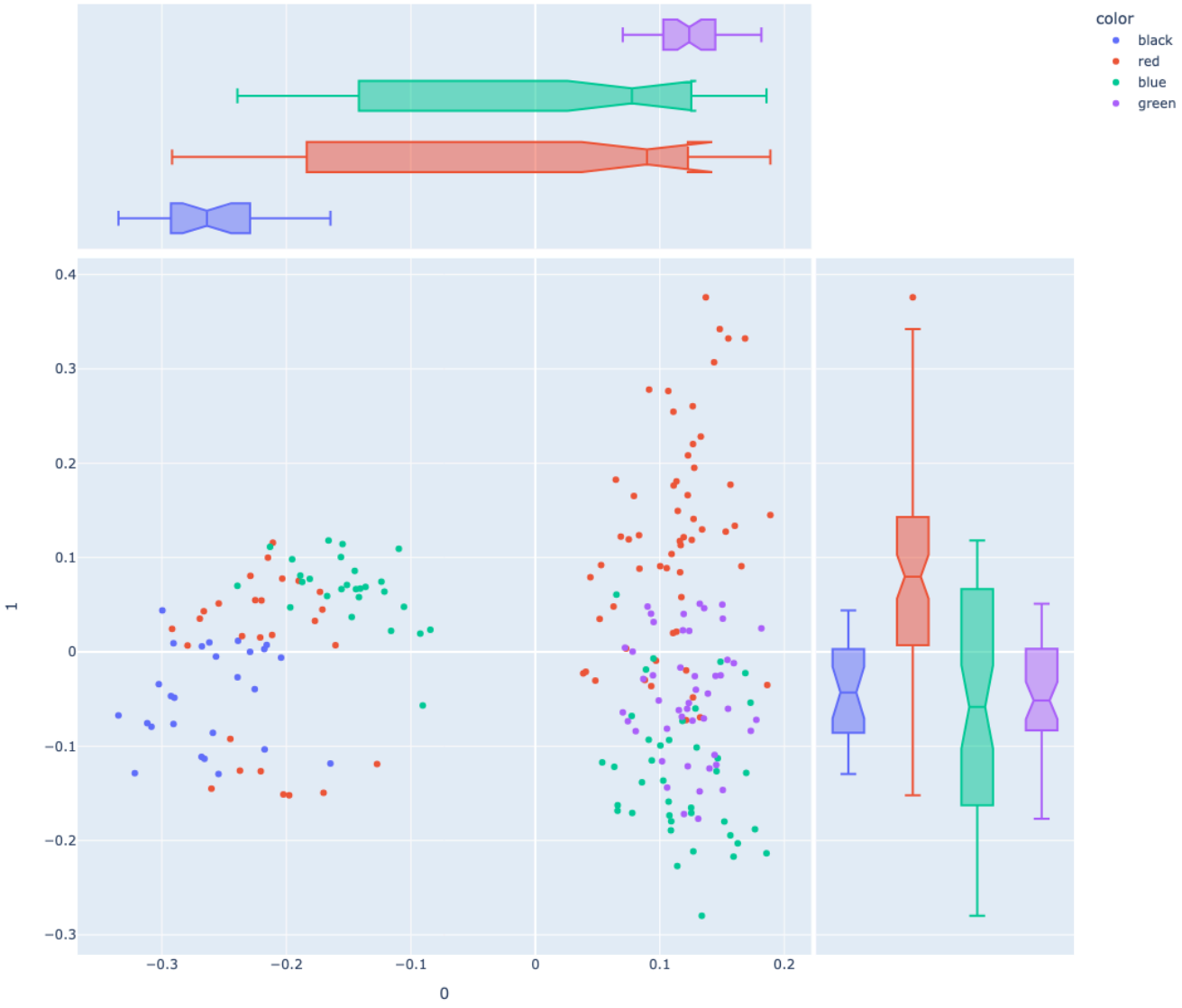
Riff 2 (5)

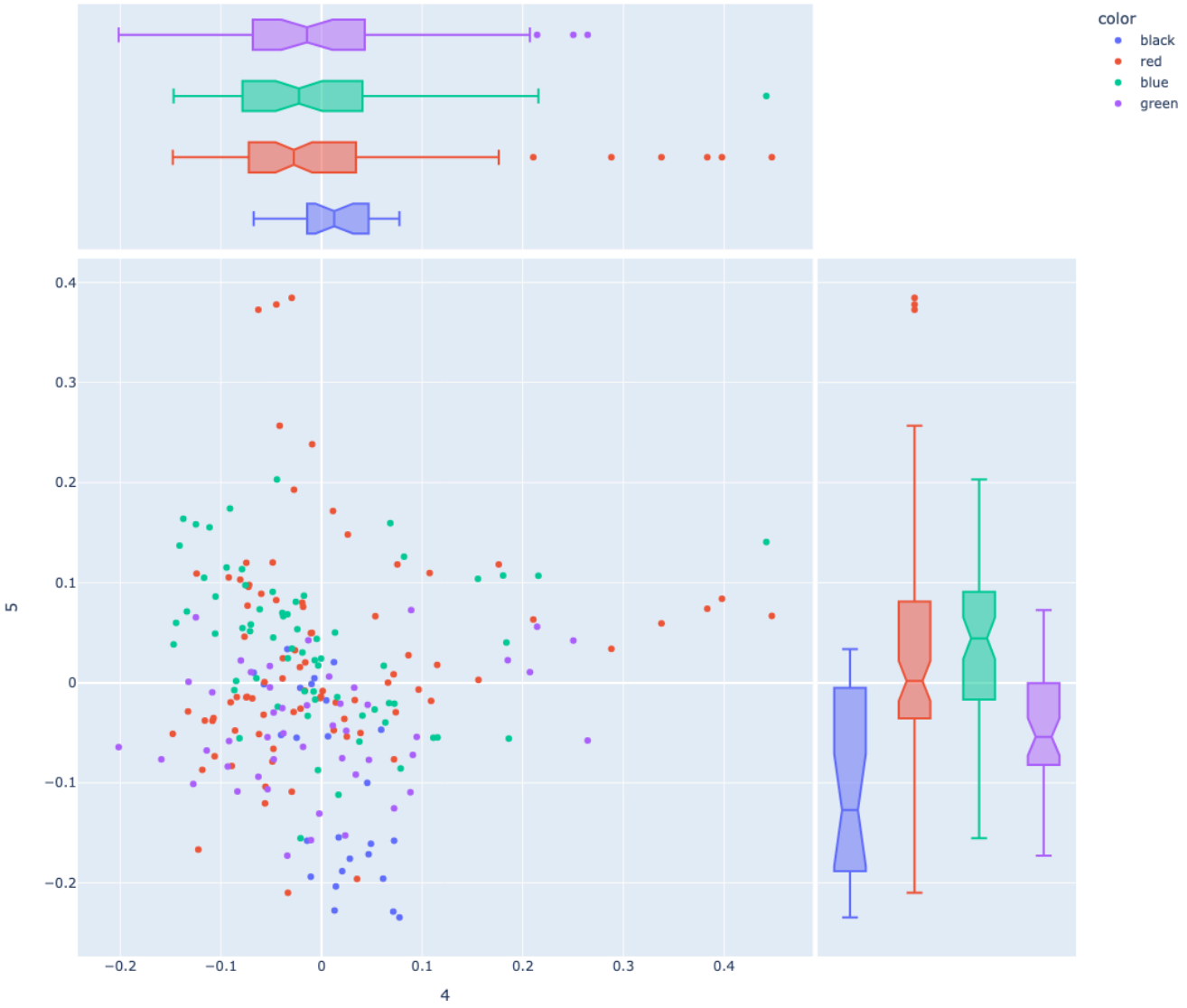


Both the "Divergent" series by Veronica Roth and "The Hunger Games" trilogy by Suzanne Collins share striking similarities in their exploration of themes such as survival, resistance against oppressive systems, and the quest for identity amidst conflict. Through richly crafted narratives and compelling characters, both series captivate readers with tales of courage, sacrifice, and the power of friendship in the face of adversity. Due to this I created a list of words that I thought might appear throughout both books.

Both `change` and `power` are clearly littered across both series very often. `rebel` on the other hand is mostly in the third book of the Hunger Games series (Mockingjay), but does show up a few times in the third book of the Divergent series (Allegiant) as well. `conflict` and `oppression` both only appear to show up in one of the series. the other words are scattered throughout.

Riff 3 (5)


color
black
red
blue
green
0.4
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
1
−0.3
−0.2
−0.1
0
0.1
0.2
0



One thing I was curious about is if the color of the book covers had any hidden correlations. As there is some overlap with both series having a red and blue cover for two of the books.

I decided to look at this by comparing the colors in different PCA component spaces. In the first space the series are clearly separated. But one of the colors (blue) is on the outside edge of both groupings whereas the other overlapping color (red) does not seem to have a pattern.

As I continued to look the the next compoenet spaces all of the points started to converge together, but there did not appear to be any pattern with cover color.

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.

Answer

Throughout this analysis, one of the most intriguing findings was the distinct contrast observed in the principal component analysis (PCA) between authors Roth (Divergent) and Collins (Hunger Games). Roth's representation skewed towards the positive end, while Collins' leaned towards the negative direction.

What adds depth to this observation is the evolution of the characters within their respective narratives. Katniss from the Hunger Games begins as an unwilling participant, forced into a role she never sought. Her journey, however, transforms her into a beacon of hope and a force for change in the face of oppression.

Similarly, Tris from Divergent navigates a society defined by strict factions, embarking on a journey of self-discovery and defiance against societal norms. Her growth into her true self parallels a narrative of empowerment and resistance against injustice.

In conclusion, the contrasting PCA representations of these authors' works underscore the diverse trajectories of their protagonists. Despite their differences, both narratives explore themes of identity, resistance, and empowerment, resonating with audiences through their compelling journeys of personal growth and societal change.

In [ ]: