Final Project Notebook

DS 5001 Text as Data | Spring 2025

In [363…
```
# I couldn't get my images to render using ![]() formatting, so I'm doing this instead
from IPython.display import Image, display
```

Metadata

- Full Name: Samantha Remmey
- Userid: sqr8ap
- GitHub Repo URL: https://github.com/sqr8ap/DS5001-2025-01-R/tree/fp/final_project
- UVA Box URL: https://virginia.box.com/s/a50wn3g3o25hsa9anxt88f9soulr0cmz

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.

I will be conducting analyses on a subset of Virginia Woolf's novels, sourced from Project Gutenberg. The novels are contained in plaintext files, and three of the four novels are broken down by chapter. One novel is broken down into sections separated by line breaks, represented as asterisks in the text file. The novels include Mrs. Dalloway, The Voyage Out, Night and Day, and Jacob's Room. Virginia Woolf is known for themes of identity, social structures & gender, time & memory, and the mind & consciousness.

Source Features (1)

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

- Source URL: https://www.gutenberg.org/ebooks/author/89
- UVA Box URL: https://virginia.box.com/s/a50wn3g3o25hsa9anxt88f9soulr0cmz
- Number of raw documents: 4
- Total size of raw documents (e.g. in MB): 2.547 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.

The corpus consists of novels broken down into chapters or sections. Each plaintext file contains metadata regarding the book itself and its source, Project Gutenberg. The beginning and end of each novel is clearly marked and can be easily parsed. All files are consistent in their structures.

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/x5eyxhuhmuv857qhh08ar1fbrgdg7fis
- GitHub URL for notebook used to create: https://github.com/sqr8ap/DS5001-2025-01-R/blob/fp/final_project/Parse.ipynb
- Delimitter: ,
- Number of observations: 4
- List of features, including at least three that may be used for model summarization (e.g. date, author, etc.): source_file_path, raw_title, book_len, n_chaps, date, n_chars, woolf_age, prot_sex (sex of protagonist)
- Average length of each document in characters: 477816.5

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/c0s20zzarrs0qmwngwz9qjjew25xboj3

- GitHub URL for notebook used to create: https://github.com/sqr8ap/DS5001-2025-01-R/blob/fp/final_project/Parse.ipynb
- Delimitter: ,
- Number of observations Between (should be >= 500,000 and <= 2,000,000 observations.): 427340
- OHCO Structure (as delimitted column names): book_id | 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 | n_chars

VOCAB (2)

The unique word types (terms) in the corpus.

- UVA Box URL: https://virginia.box.com/s/idukv7yfch2lklqpa3p1wofp6gm76s3h
- GitHub URL for notebook used to create: https://github.com/sqr8ap/DS5001-2025-01-R/blob/fp/final_project/Parse.ipynb
- Delimitter: ,
- Number of observations: 18495
- Columns (as delimitted names, including `n`, `p`', `i`, `dfidf`, `porter_stem`, `max_pos` and `max_pos_group`, `stop`): n | n_chars | p | i | stop | stem_porter | max_pos | max_pos_group | n_pos_group | cat_pos_group | n_pos | cat_pos | df | idf | 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.

'pages', 'wants', 'considerable', 'god', 'happiness', 'save', 'pink', 'single', 'john', 'bedroom', 'agree', 'gentlemen', 'interrupted', 'burst', 'explained', 'comes', 'compared', 'hat', 'anyhow', 'force'

Derived Tables

BOW (3)

A bag-of-words representation of the CORPUS.

- UVA Box URL: https://virginia.box.com/s/msaj7297vn5ervmqsukn68mnj6omgleb
- GitHub URL for notebook used to create: https://github.com/sqr8ap/DS5001-2025-01-R/blob/fp/final_project/Parse.ipynb
- Delimitter: ,
- Bag (expressed in terms of OHCO levels): ['book_id', 'chap_num']
- Number of observations: 113953
- Columns (as delimitted names, including `n`, `tfidf`): n | tf | tfidf

DTM (3)

A represenation of the BOW as a sparse count matrix.

- UVA Box URL: https://virginia.box.com/s/lypsnaqz4hu13kf0qzvv8nr9y0n41m8l
- UVA Box URL of BOW used to generate (if applicable): https://virginia.box.com/s/msaj7297vn5ervmqsukn68mnj6omgleb
- GitHub URL for notebook used to create: https://github.com/sqr8ap/DS5001-2025-01-R/blob/fp/final_project/Parse.ipynb
- 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.box.com/s/fcp2fi6sde0xtslzwrzf7m12bipukcxl
- UVA Box URL of DTM or BOW used to create: https://virginia.box.com/s/lypsnaqz4hu13kf0qzvv8nr9y0n41m8l
- GitHub URL for notebook used to create: https://github.com/sqr8ap/DS5001-2025-01-R/blob/fp/final_project/Parse.ipynb
- Delimitter: ,
- Description of TFIDIF formula ($LaTeX$ OK): max tfidf -> $\left(\frac{\text{DTCM}^T}{\max(\text{DTCM}^T)}\right)^T$

Reduced and Normalized TFIDF_L2 (3)

A Document-Term matrix with L2 normalized TFIDF values.

- UVA Box URL: https://virginia.box.com/s/5nryqu5g20rbjclguyhgn5t0zc1sdj0c
- UVA Box URL of source TFIDF table: https://virginia.box.com/s/fcp2fi6sde0xtslzwrzf7m12bipukcxl
- GitHub URL for notebook used to create: https://github.com/sqr8ap/DS5001-2025-01-R/blob/fp/final_project/Parse.ipynb
- Delimitter: ,
- Number of features (i.e. significant words): 2000
- Principle of significant word selection: top 2000 terms with highest mean tfidf

Models

PCA Components (4)

- UVA Box URL: https://virginia.box.com/s/wbqtfqcrrdcitvt7kitlam2id6dklucn
- UVA Box URL of the source TFIDF_L2 table: https://virginia.box.com/s/5nryqu5g20rbjclguyhgn5t0zc1sdj0c
- GitHub URL for notebook used to create: https://github.com/sqr8ap/DS5001-2025-01-R/blob/fp/final_project/Models.ipynb
- Delimitter: ,
- Number of components: 10
- Library used to generate: none; I used my own function from hw7 to apply PCA from scratch
- Top 5 positive terms for first component: that but not was had
- Top 5 negative terms for second component: rachel women really like villa

PCA DCM (4)

The document-component matrix generated.

- UVA Box URL: https://virginia.box.com/s/22beso4gh69h4m8pgpsw6jmg4cagjenf
- GitHub URL for notebook used to create: https://github.com/sqr8ap/DS5001-2025-01-R/blob/fp/final_project/Models.ipynb
- Delimitter: ,

PCA Loadings (4)

The component-term matrix generated.

- UVA Box URL: https://virginia.box.com/s/zqr18gesf75ikj3veamn7u61bqfd29sn
- GitHub URL for notebook used to create: https://github.com/sqr8ap/DS5001-2025-01-R/blob/fp/final_project/Models.ipynb
- Delimitter: ,

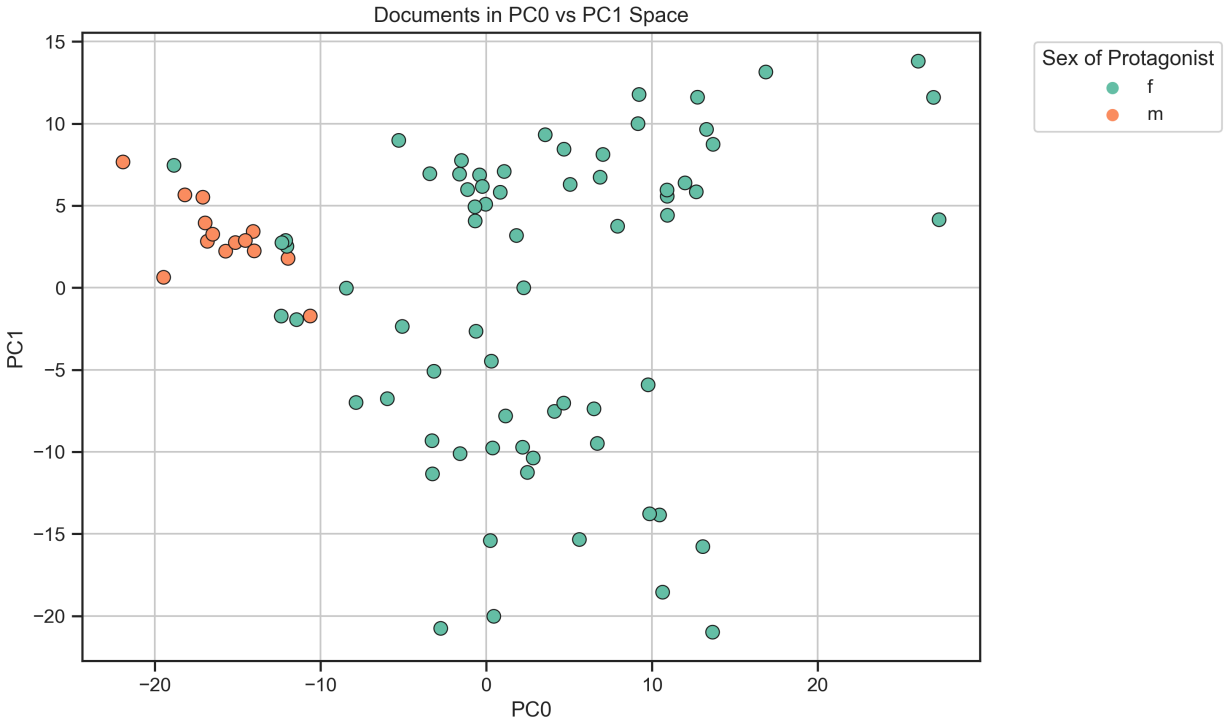
PCA Visualization 1 (4)

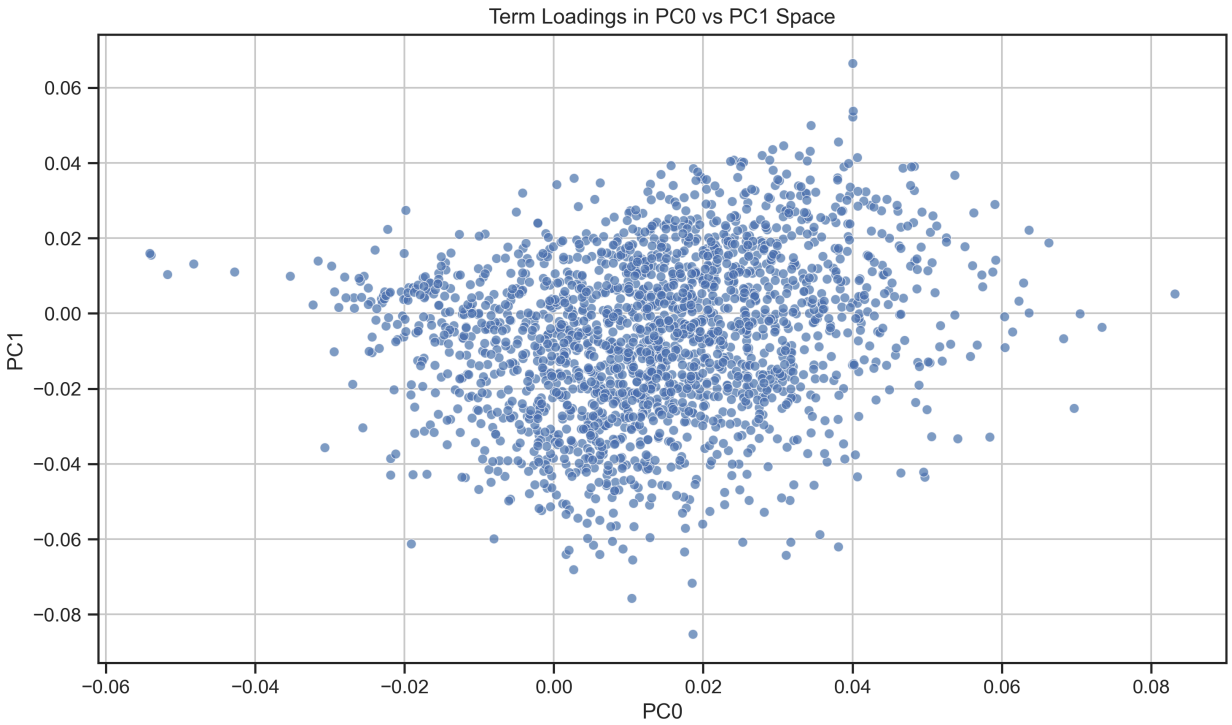
Include a scatterplot of documents in the space created by the first two components.

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.)

In [368…
```
for img_file in ["pca_vis_1a.png", "pca_vis_1b.png"]:
    display(Image(filename=img_file))
```





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

Based on the plot of documents in PC0 vs PC1 space, there's a notable separation along the first component (PC0) between documents with a male protagonist and documents with a female protagonist. Documents with a male protagonist also appear to be tightly clustered along PC1 as well. Documents with a female protagonist are a bit more variable in terms of both components. The plot of the loadings for these two components depicts a cloud of points spanning all directions, though most points map to the positive pole of PC0.

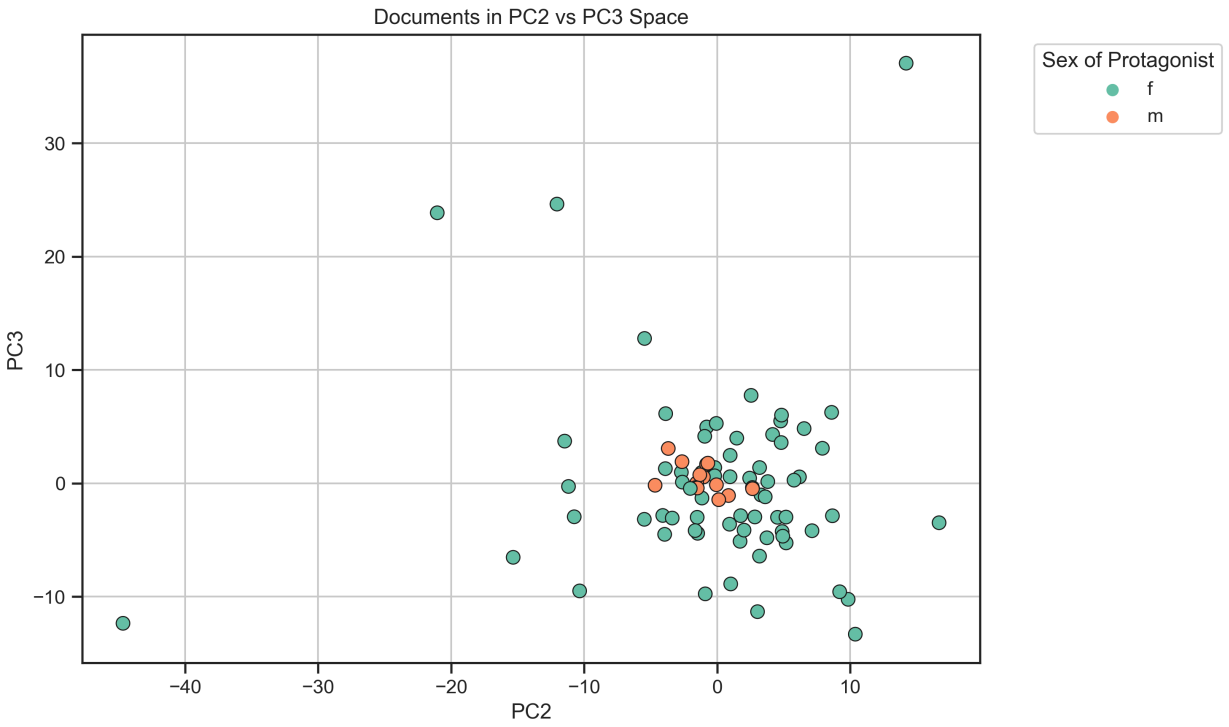
PCA Visualization 2 (4)

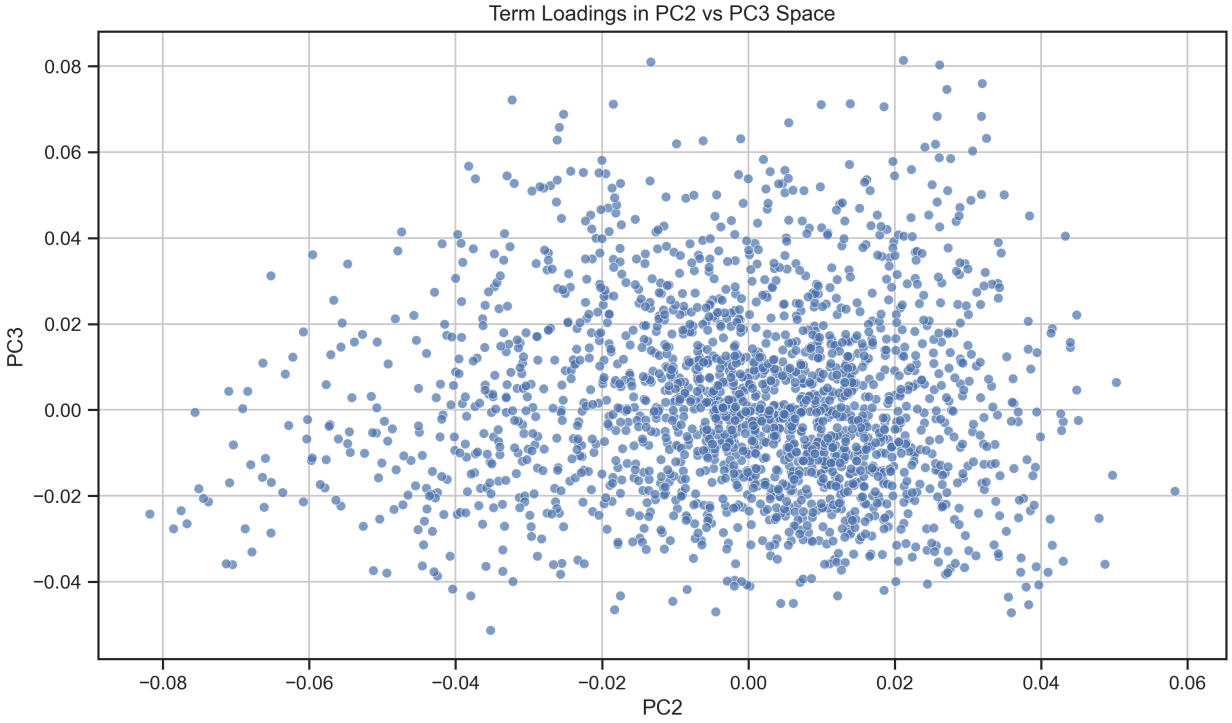
Include a scatterplot of documents in the space created by the second two components.

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.)

In [371…
```
for img_file in ["pca_vis_2a.png", "pca_vis_2b.png"]:
    display(Image(filename=img_file))
```





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

The plot of documents in PC2/PC3 space depicts a tighter cluster that lacks clear separation based on protagonist sex. Points are largely centered around 0 along each component. The plot of the term loadings for these components depicts a wide cloud centered at (0,0).

LDA TOPIC (4)

- UVA Box URL: https://virginia.box.com/s/hgg21jrdhaoe3zt9h95l240ujcdy1hro
- UVA Box URL of count matrix used to create: https://virginia.box.com/s/c8g5xsqz9img34lgzu1hz98i21m6lskq

- GitHub URL for notebook used to create: https://github.com/sqr8ap/DS5001-2025-01-R/blob/fp/final_project/Models.ipynb
- Delimitter: ,
- Libary used to compute: sklearn
- A description of any filtering, e.g. POS (Nouns and Verbs only): common nouns only (filtered out proper)
- Number of components: 20
- Any other parameters used: bag = chapter = ['book_id', 'chap_num'], max_features = 4000, stop_words = 'english', n_components = 20, max_iter = 5, learning_offset = 50, random_state = 0, n_words = 7
- Top 5 words and best-guess labels for top five topics by mean document weight:
  - T00: harsh aback pride prime primeval; PRIMAL
  - T01: mother life time book people; FAMILY
  - T02: people time room things eyes; HOME
  - T03: harsh aback pride prime primeval; PRIMAL
  - T04: harsh aback pride prime primeval; PRIMAL

Note: my topics seem to have some repeats, so here's two other topics with top words and best guess labels:

T19: wind waves eyes ship sea; SEA/VOYAGE

T06: trees river boat bank deck; NATURE

LDA THETA (4)

- UVA Box URL: https://virginia.box.com/s/eupjoex3v491i7o2tvef9i2ychucvnwx
- GitHub URL for notebook used to create: https://github.com/sqr8ap/DS5001-2025-01-R/blob/fp/final_project/Models.ipynb
- Delimitter: ,

LDA PHI (4)

- UVA Box URL: https://virginia.box.com/s/ien5kmrnktewrkkf9jtnh1g6tt4ywxhd
- GitHub URL for notebook used to create: https://github.com/sqr8ap/DS5001-2025-01-R/blob/fp/final_project/Models.ipynb
- Delimitter: ,

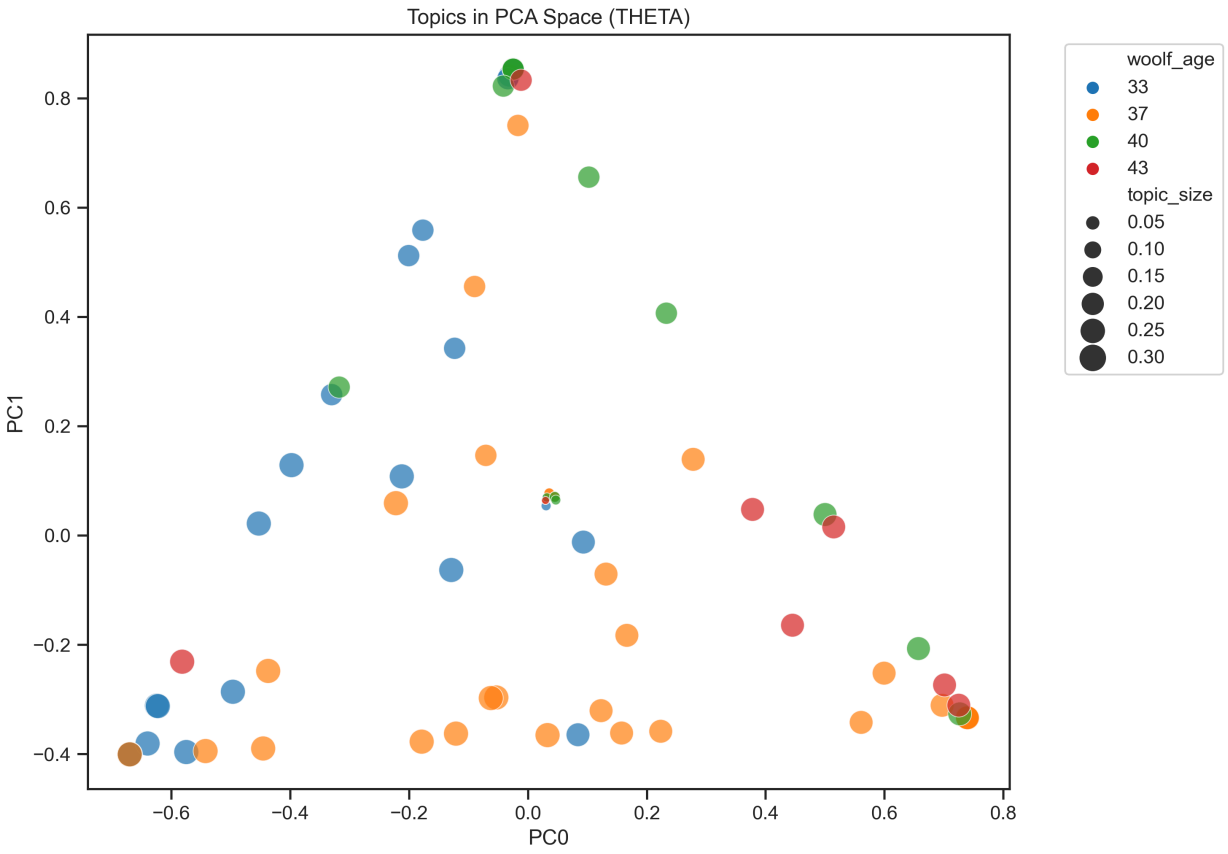
LDA + PCA Visualization (4)

Apply PCA to the THETA 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.

In [374…
```
display(Image(filename="lda_pca.png"))
```



Provide a brief interpretation of what you see.

Although I'm not seeing any major trends here, I can see that all points where Woolf's age is 33 (the first book published in my corpus) fall towards the negative pole of PC0, and almost all points where Woolf's age is 43 (the last book published in my corpus) fall towards the positive pole of PC0. Perhaps PC0 is capturing some thematic element that changed over the course of Woolf's works.

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/2s00r3jldmoiyvknrbrax0y2jkf4v3vs
- UVA Box URL for source lexicon: https://virginia.box.com/s/o4vone8dd46l4bshltzipnorxy681rwk
- GitHub URL for notebook used to create: https://github.com/sqr8ap/DS5001-2025-01-R/blob/fp/final_project/Models2.ipynb
- Delimitter: ,

Sentiment BOW_SENT (4)

Sentiment values from VOCAB_SENT mapped onto BOW.

- UVA Box URL: https://virginia.box.com/s/abr11n2d2g3rus6d5usre30fnphc1vwn
- GitHub URL for notebook used to create: https://github.com/sqr8ap/DS5001-2025-01-R/blob/fp/final_project/Models2.ipynb
- Delimitter: ,

Sentiment DOC_SENT (4)

Computed sentiment per bag computed from BOW_SENT.

- UVA Box URL: https://virginia.box.com/s/h3vs21ro8ck464kex3pio2x9tm8jhh68
- GitHub URL for notebook used to create: https://github.com/sqr8ap/DS5001-2025-01-R/blob/fp/final_project/Models2.ipynb
- Delimitter: ,
- Document bag expressed in terms of OHCO levels: ['book_id', 'chap_num']

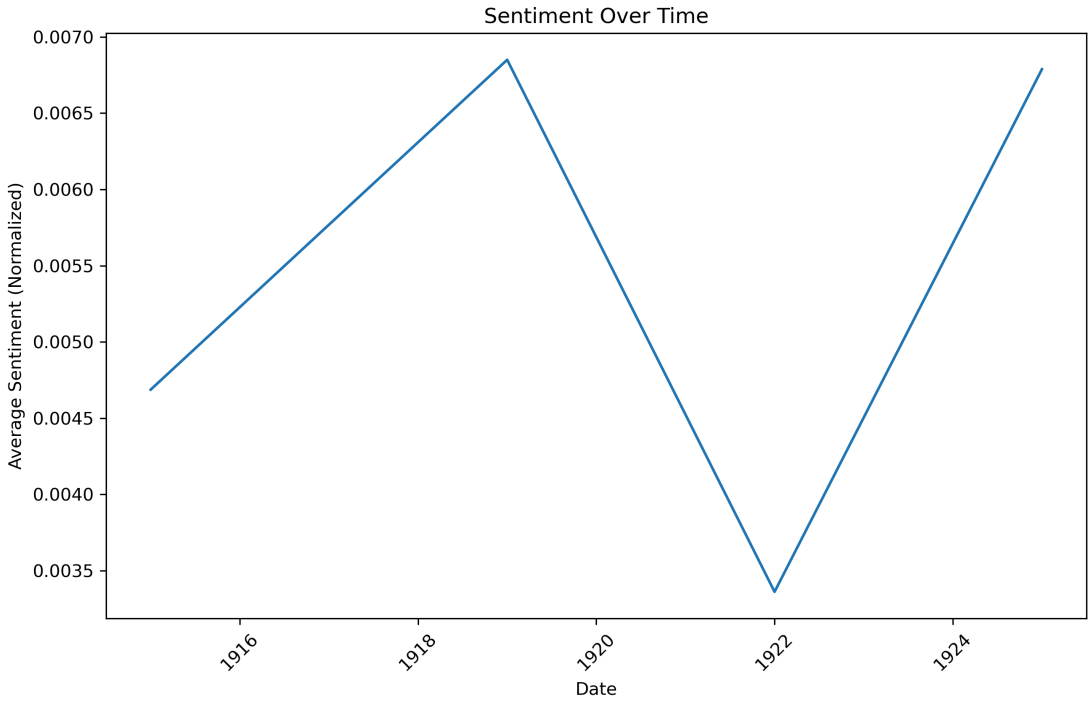
Sentiment Plot (4)

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.

In [383…
```
display(Image(filename="sent_time.png"))
```



VOCAB_W2V (4)

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

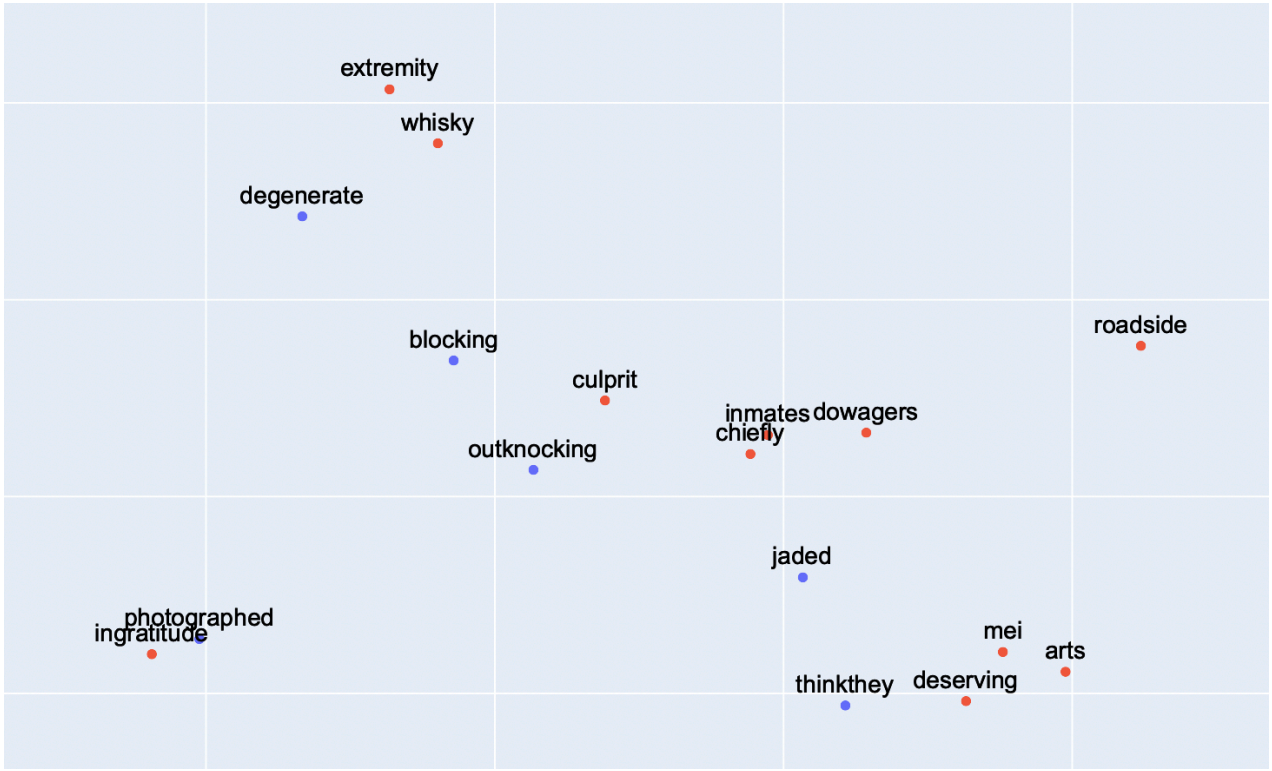
- UVA Box URL: https://virginia.box.com/s/alv7j90c87o4ty8be5d5mg30zaw758s7
- GitHub URL for notebook used to create: https://github.com/sqr8ap/DS5001-2025-01-R/blob/fp/final_project/Models3.ipynb
- Delimitter: ,
- Document bag expressed in terms of OHCO levels: Paragraph - ['book_id', 'chap_num', 'para_num', 'sent_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.

In [386…
```
display(Image(filename="tsne.png"))
```



This cluster contains words like 'degenerate,' 'culprit,' 'inmates' and 'jaded,' suggesting this region in space may be capturing some element/theme related to crime or moral corruption.

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.

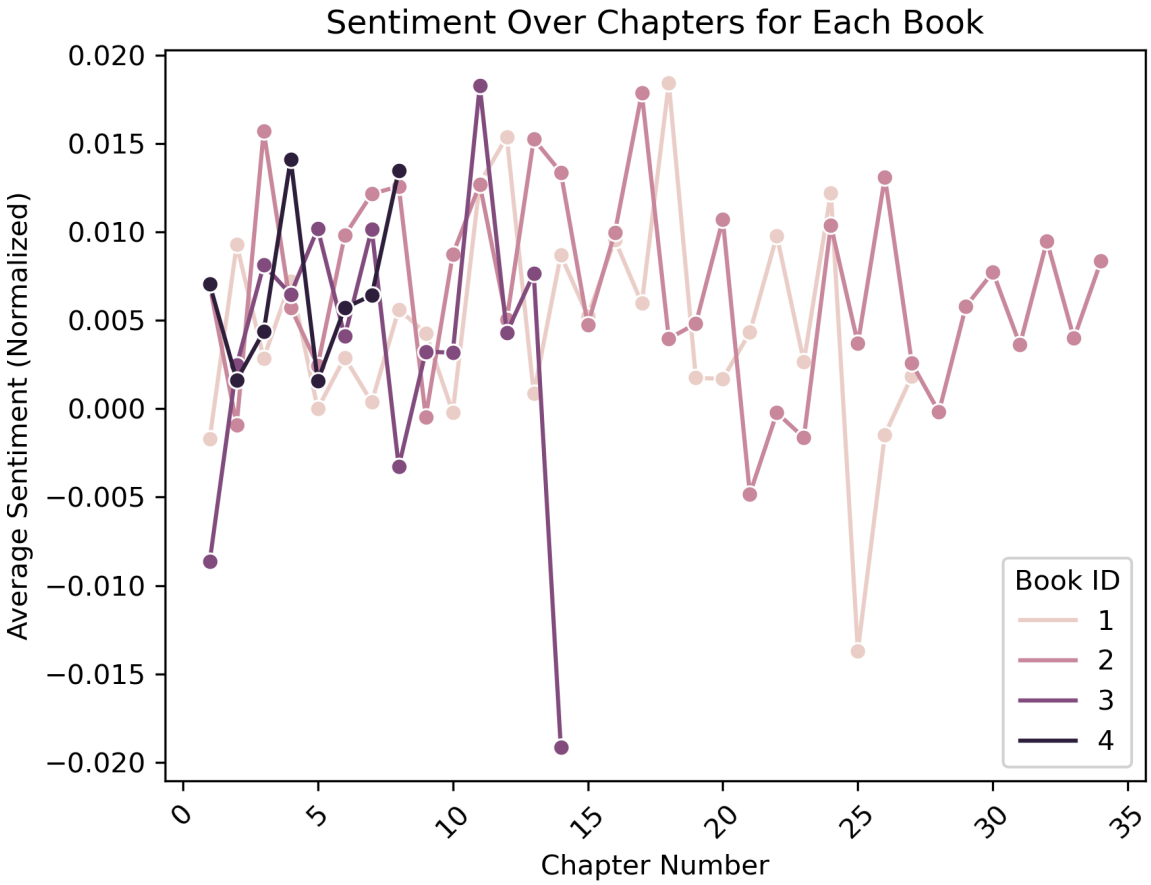
The nature of this relationship is left open to you -- it may be correlation, or mutual information, or something less well defined.

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)

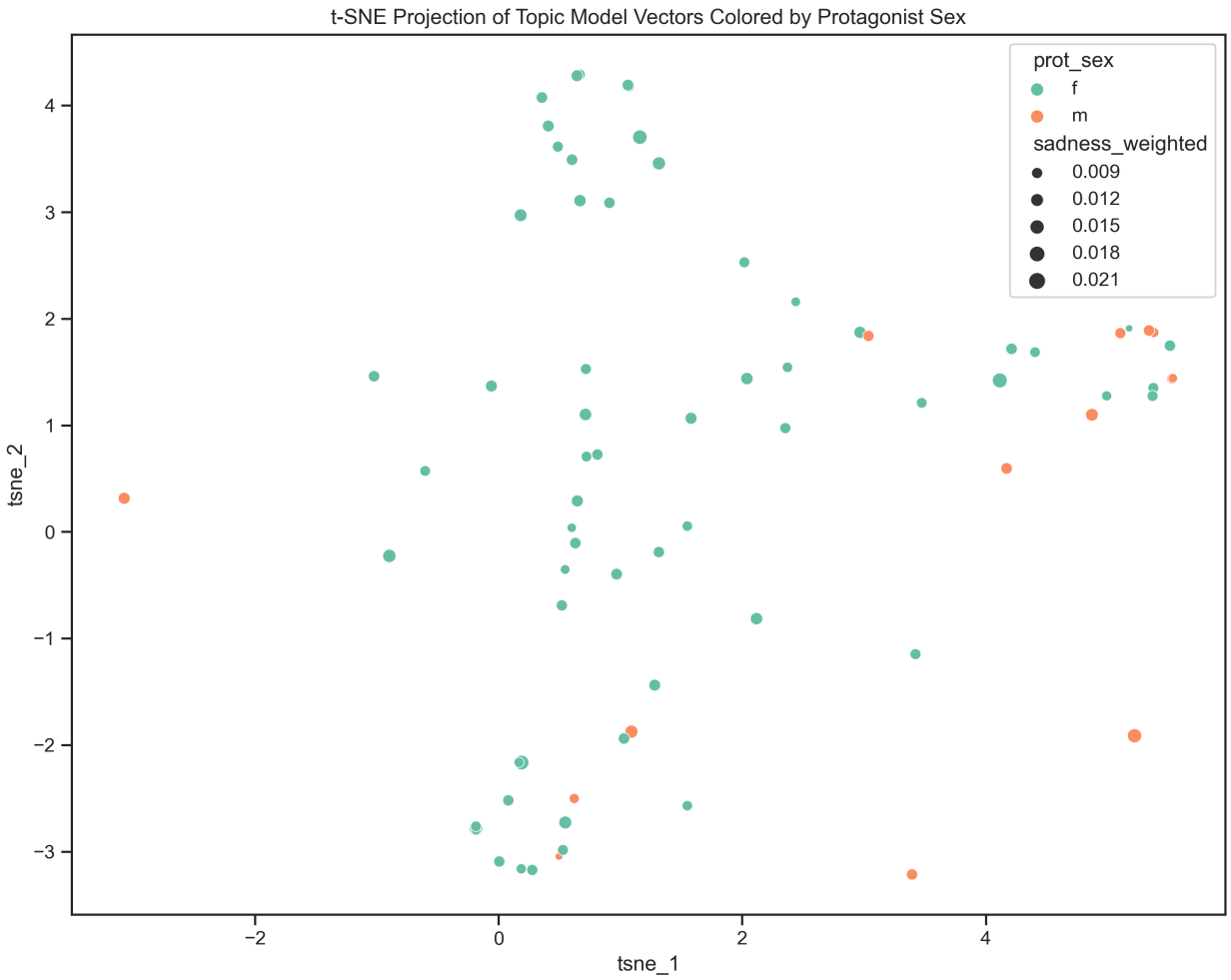
In [391…
```
display(Image(filename="sent_chap.png"))
```



Here I've plotted average overall sentiment throughout the chapters of each book. We can see a steep decline in sentiment at the end of book 3, which is Jacob's Room. This notable drop-off makes sense, as Jacob's Room has a very sad ending with the death of Jacob. We can also see some notable dips in sentiment towards the end of books 1 and 2, which are The Voyage Out and Night and Day, respectively. The Voyage Out also ends in the death of a major character, so this decline also makes sense.

Riff 2 (5)

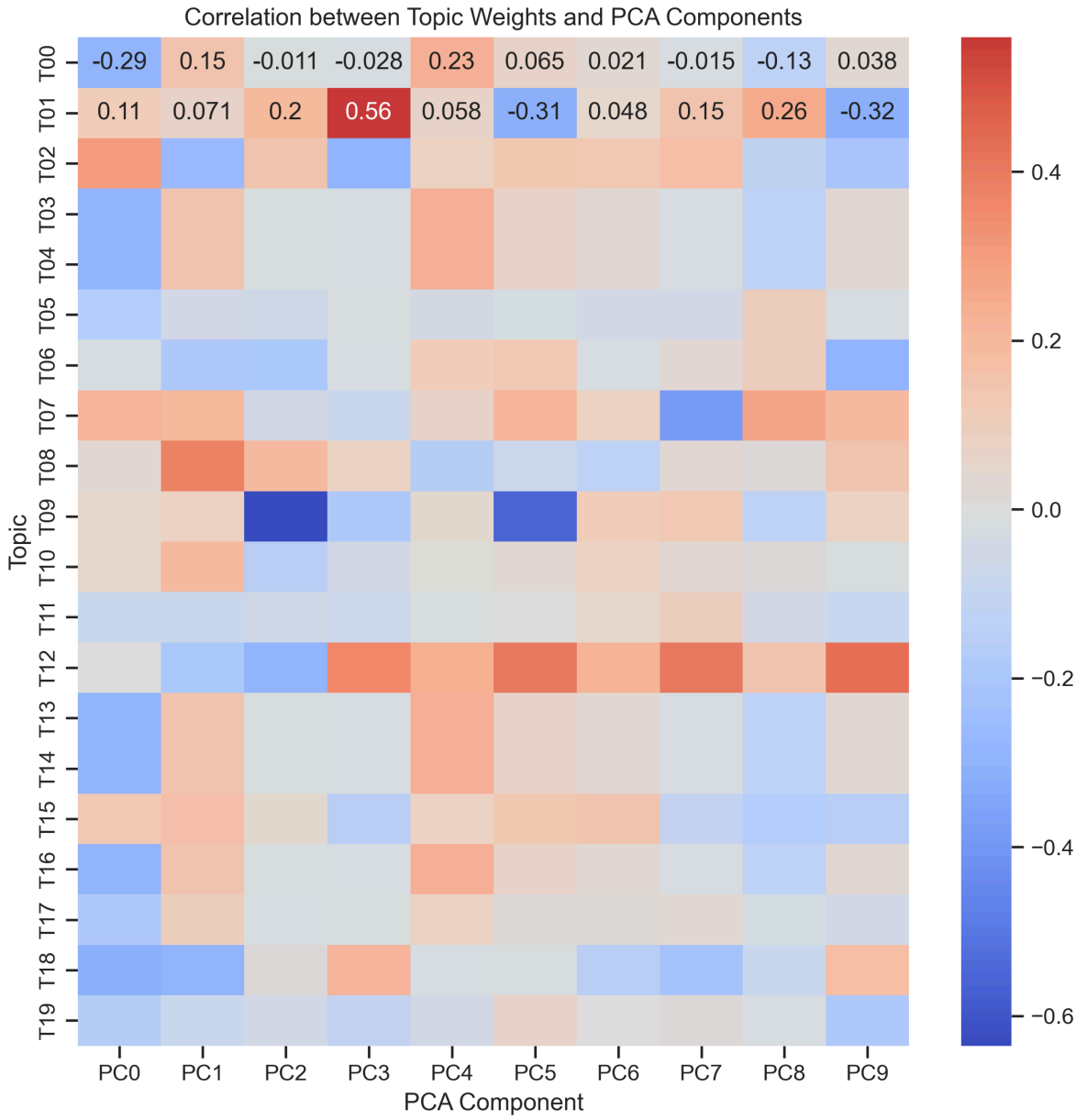
In [395…
```
display(Image(filename="tsne_topic_models.png"))
```



Here I've plotted the tsne projection of topic model vectors. I mapped protagonist sex and sadness score to color and size, respectively. Interestingly, most points with protagonist sex of male fall on the positive side of the first tsne dimension and the negative side of the second tsne dimension, with all other points spanning both positive and negative poles of each tsne dimension more equally. I'm not seeing any notable trends in terms of sadness scores.

Riff 3 (5)

In [399…
```
display(Image(filename="pca_lda_corr.png"))
```



Here I've generated a heatmap correlating topic weights and PCA components. It appears that PC2 and T09 are moderately correlated, as well as PC3 and T01.

The top terms associated with T09 are night, eyes, arm, moment, light, door and time; the top terms associated with the negative pole of PC2 are region, darkness, globe, swam and recollection. There seems to be a shared theme of darkness or liminality here. Woolf is known for themes of consciousness and the mind, so this topic/component correlation makes a lot of sense.

The top terms associated with T01 are mother, life, time, book, people, men and poet; the top terms associated with the positive pole of PC3 are poet, among, mothers, poets, task and books. There's definitely a shared theme of poetry and/or motherhood here, which also makes sense in the context of Woolf's characteristic themes.

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.

I found the results of topic modeling and PCA to be the most interesting part of this pipeline, as the themes and topics ended up being well-aligned with themes that Virginia Woolf is known for, even with a subset of only four of her works. Topics and principal components seemed to capture highly symbolic themes, distinguishing between themes of people, nature, darkness, and the human psyche. One LDA topic I found particularly interesting was associated with words like “truth,” “mind,” “care,” “suffering,” and “illusion.” This topic struck me as a clear embodiment of Woolf’s infatuation with the human condition and consciousness.

I also thought the clusters uncovered by the word2vec tSNE plot were really interesting. Even in a vector space occupied by thousands of words, we still see Woolf’s characteristic themes emerge amongst the noise. In one area I found terms like “beauty,” “ladies,” “made,” “daughter,” “observed,” and “heart,” suggesting a theme of womanhood, and in an area far away I discovered terms like “emotions,” “contemplate,” “kissing,” “softening,” “human,” and “invested,” suggesting a theme of relationships or interpersonal dynamics. It’s intriguing to come across concrete evidence via analytical techniques of themes that we know to be present in a given author’s work.

What stood out most was how these computational methods—topic modeling, embeddings, and sentiment analysis—converged to reinforce the known motifs of Woolf’s writing while also revealing subtle nuances across different texts. Incorporating additional metadata, such as publication date or protagonist characteristics, added a valuable lens for interpretation. I think it would be worth exploring her work further by incorporating essays and short stories and analyzing whether thematic or emotional shifts emerge across genres or time.