A Joyceless Stream-of-Consciousness Book*

Exploring Word Frequency, Sentiment Value and Mental Health Themes in the Works of Joyce, Woolf, Proust, Mansfield and Eliot from Project Gutenberg

Quang Mai

April 4, 2024

This project focuses on understanding the language used by renowned stream of consciousness (SOC) authors James Joyce, Virginia Woolf, Marcel Proust, Katherine Mansfield and T.S Eliot. By conducting word frequency analysis and sentiment analysis of these authors' nine novels, I aim to uncover shared linguistic patterns and gain insights into the authors' mental states. Through the analysis on SOC literature, this paper thus attempts to offer insights into themes of self-identity, anxiety, disassociation and existential contemplation within Western society's from late 19th to mid-20th century. (add one sentence on main results)

Table of contents

1 Introduction **2**

2 Data **4**
- 2.1 Measurement . . . 4
- 2.2 Source Data . . . 5
- 2.3 Data Cleaning and Word Tokenization . . . 5
  - 2.3.1 Word Count . . . 5
  - 2.3.2 Comparative Word Frequency . . . 6
  - 2.3.3 Generating Word Networks . . . 6

3 Model **6**
- 3.1 Model set-up . . . 6
  - 3.1.1 Model justification . . . 7

*Code and data are available at: https://github.com/ponolite/stream-consciousness-language.git

4 Results **7**
4.1 The Dominant Vocabulary of SOC Literature 7
4.2 Sentiment Analysis 9
4.3 Gendered Mental Landscape of Stream of Consciousness Novels 9
4.3.1 Comparing SOC Literature's Female Authors 9
4.3.2 Comparing SOC Literature's Male Authors 9
4.4 Transnational SOC Novels and Mental Health Themes 9
4.5 Combined Texts: Trends, Word Networks, Bigram and Trigram Analsis 11

5 Discussion **11**
5.1 Mental Health Vocabulary: Patterns and Trends 11
5.2 Insights into Socio-Political Landscape of the West's Modernist Era 11
5.3 Schizophrenic and Disassociative Tendencies in Female Stream of Consciousness 13
5.4 Weaknesses 13
5.4.1 Lack of Thorough Word Cleaning 13
5.4.2 Decontextualized Literature Works and Limiting Publication Editions . 13
5.4.3 Uneven Novel Length and Categorization of Authors 13
5.4.4 Project Gutenberg's Focus on the Canon 13
5.5 Moving Forward and Next Steps 13

6 Appendix **14**
6.1 Additional Data Details 14
6.1.1 Data Gathering 14
6.1.2 Data Cleaning 14
6.2 Model Details 14
6.3 Posterior predictive check 14
6.4 Diagnostics 14

References **15**

1 Introduction

Stream of consciousness (SOC) is a narrative technique that aims to capture the continuous flow of thoughts, feelings, and sensations experienced by a character without conventional organization or punctuation (Bernini and Fernyhough 2022). It mirrors the unpredictable and interconnected nature of human thought processes, often revealing the inner workings of the character's mind in an intimate and unfiltered manner (Long and So 2016). In literature, most scholars agree that stream of consciousness reveals the complexities of mental-scapes, shedding light on the nuances of characters' emotional well-being and psychological struggles (Nyongesa 2023). As such, this paper has mined the texts of a total of nine novels from the volunteer archive, Project Gutenberg, to examine the mental health themes of famous stream of consciousness authors, namely by Joyce, Woolf, Proust, Mansfield and Eliot, from

the modernist era of literature, spanning from late 19th century to mid-20th century ("Project Gutenberg," n.d.). (more stats and data mentioned here)

By analyzing these textual datasets through word frequency and sentiment analysis, I seek to pose and answer crucial questions: *What are some important factors contributing to this relationship between mental health, disassociation and stream of consciousness? Moreover, how does this relationship vary differently across differnet demographics of authors, for instance, authors with different geographical locations and genders?* Understanding these dynamics is crucial in having an informed understanding of the West's late 19th to mid-20th century socio-political landscape, especially in regards to how authors and creative writers navigate and deal with then-taboo topics such as existential angst, mental health issues and disabilities.

Thus, my estimand is the correlation between mental health-related words in SOC literature, their frequency and sentimental value as provided by Mohammad and Turney (2013). This is considered in terms of nine selected SOC novels only, namely Joyce's A Portrait of the Artist as a Young Man and Chamber Music; Woolf's Mrs Dalloway and Jacob's Room; Proust's Swann Way; Mansfield's Bliss and The Garden Party; and Eliot's The Waste Land and The Love Song of J. Alfred Prufrock. Through our analysis, we found that (percentage, number and data here, main results)…

To further understand the correlation between stream of consciousness novels and mental health themes, in Introduction, our paper briefly discusses the nature of stream of consciousness literature, relevant authors and the works that we've chosen to analyze. Subsequently, in Data and Results, we talk about the nature of the data obtained and analyze the results garnered from the data with suitable tables and charts. Next, Discussion provides further insights and future areas of study. Finally, [Conclusion] summarizes our main findings. To complete the paper, Appendix clarifies how each variable within each dataset is generated and tables to accordingly demonstrate this.

The novel texts used for analysis were sourced from Project Gutenberg under the library `gutenbergr` (Johnston and Robinson 2023) ("Project Gutenberg," n.d.). Data was generated, extracted and cleaned using the open-source statistical programming language R (R Core Team 2022), leveraging functions from `tidyverse` (Wickham et al. 2019), `tidytext` (Julia Silge and Robinson 2016), `rmarkdown` (Allaire et al. 2024), `dplyr` (Wickham et al. 2022), `ggplot2` (Wickham 2016), `scales` (Wickham, Pedersen, and Seidel 2023), `here` (Müller 2020), `igraph` (J. Silge and Robinson 2006), `widyr` (J. Silge and Robinson 2022), `ggraph` (Pedersen 2024), `textdata` (Hvitfeldt 2022), `tm` (Feinerer, Hornik, and Meyer 2008) and `knitr` (Xie 2014).

2 Data

2.1 Measurement

Out of two variables used, the first one, word frequency, is directly quantified through tokenizing the novel texts using the packages Julia Silge and Robinson (2016; Robertson 2021) (Henry 2021). To do this, I first downloaded all nine novel texts from "Project Gutenberg" (n.d.), and leveraged functions such as `unnest_tokens()` from Julia Silge and Robinson (2016) to mine the texts, or separating it into individual words. Finally, I used `count()` to quantify the word frequency. The second variable used, sentiment value, is based on Mohammad and Turney (2013)'s research, where questioning to accurately capture respondents' backgrounds. For instance, the General Social Survey (GSS) obtains information on marital status through questions such as, "Are you currently married, widowed, divorced, separated, or have you never been married?", with responses categorized into predefined groups for further analysis (**gssCodebook?**). On the other hand, instead of relying on income sources, socioeconomic class is subjectively assessed based on the respondents' answers to this question, "If you were asked to use one of four names for your social class, which would you say you belong in: the lower class, the working class, the middle class, or the upper class?". Some are continuous variables, like age. With respondents' various age groups documented, we can understand the generational disparities in childbearing patterns. Systematic and data-driven, these measurement methods ensure that all data faithfully reflects the American population's childbearing tendencies.

Even though considerable attention has been given to the polarity of words (positive and negative) and thecreation of large polarity lexicons, research in emotion analysis has had to rely on limited and small emotionlexicons. In this paper, we show how the combined strength and wisdom of the crowds can be used to generate alarge, high-quality, word–emotion and word–polarity association lexicon quickly and inexpensively. We enumeratethe challenges in emotion annotation in a crowdsourcing scenario and propose solutions to address them. Mostnotably, in addition to questions about emotions associated with terms, we show how the inclusion of a word choicequestion can discourage malicious data entry, help to identify instances where the annotator may not be familiarwith the target term (allowing us to reject such annotations), and help to obtain annotations at sense level (ratherthan at word level). We conducted experiments on how to formulate the emotion-annotation questions, and showthat asking if a term isassociatedwith an emotion leads to markedly higher interannotator agreement than thatobtained by asking if a termevokesan emotion

Table 1: Table of Number of Classes Students Considered for Regrade Requests by Students' Gender

2.2 Source Data

2.3 Data Cleaning and Word Tokenization

2.3.1 Word Count

Table 2: An Examplary Table Containing Unprocessed Novel Text (James Joyce)

Book ID	Text	Book	Author
2817	To deep and deeper blue,	Chamber Music	James Joyce
2817	NA	Chamber Music	James Joyce
2817	III At that hour when all things have repose,	Chamber Music	James Joyce
2817	O lonely watcher of the skies,	Chamber Music	James Joyce
2817	NA	Chamber Music	James Joyce

Table 3: An Examplary Table Containing Tokenzied Novel Text (James Joyce)

Book ID	Book	Author	Word
4217	A Portrait of the Artist as a Young Man	James Joyce	stead
4217	A Portrait of the Artist as a Young Man	James Joyce	dublin
4217	A Portrait of the Artist as a Young Man	James Joyce	1904
4217	A Portrait of the Artist as a Young Man	James Joyce	trieste
4217	A Portrait of the Artist as a Young Man	James Joyce	1914

Table 4: An Examplary Table Containing Word Count of Each Word within Novel Texts (James Joyce)

Word	Count
stephen	373
god	194
eyes	180
soul	178
father	151

2.3.2 Comparative Word Frequency

Table 5: Word Frequency of Stream of Consciousness Novels, A Comparison Between Five Authors

Word	James Joyce	Katherine Mansfield	Marcel Proust	T.S. Eliot	Virignia Woolf
abandon	0.000112	NA	8.19e-05	NA	NA
abandoned	0.000112	NA	8.19e-05	NA	NA
abandonment	0.000112	NA	8.19e-05	NA	0.0001421
abase	0.000112	NA	NA	NA	NA
abased	0.000112	NA	NA	NA	NA

2.3.3 Generating Word Networks

3 Model

The goal of our modelling strategy is twofold. Firstly,…

Here we briefly describe the Bayesian analysis model used to investigate… Background details and diagnostics are included in Appendix 6.2.

3.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \alpha + \beta_i + \gamma_i \tag{2}$$
$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\gamma \sim \text{Normal}(0, 2.5) \tag{5}$$
$$\sigma \sim \text{Exponential}(1) \tag{6}$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of (**rstanarm?**). We use the default priors from `rstanarm`.

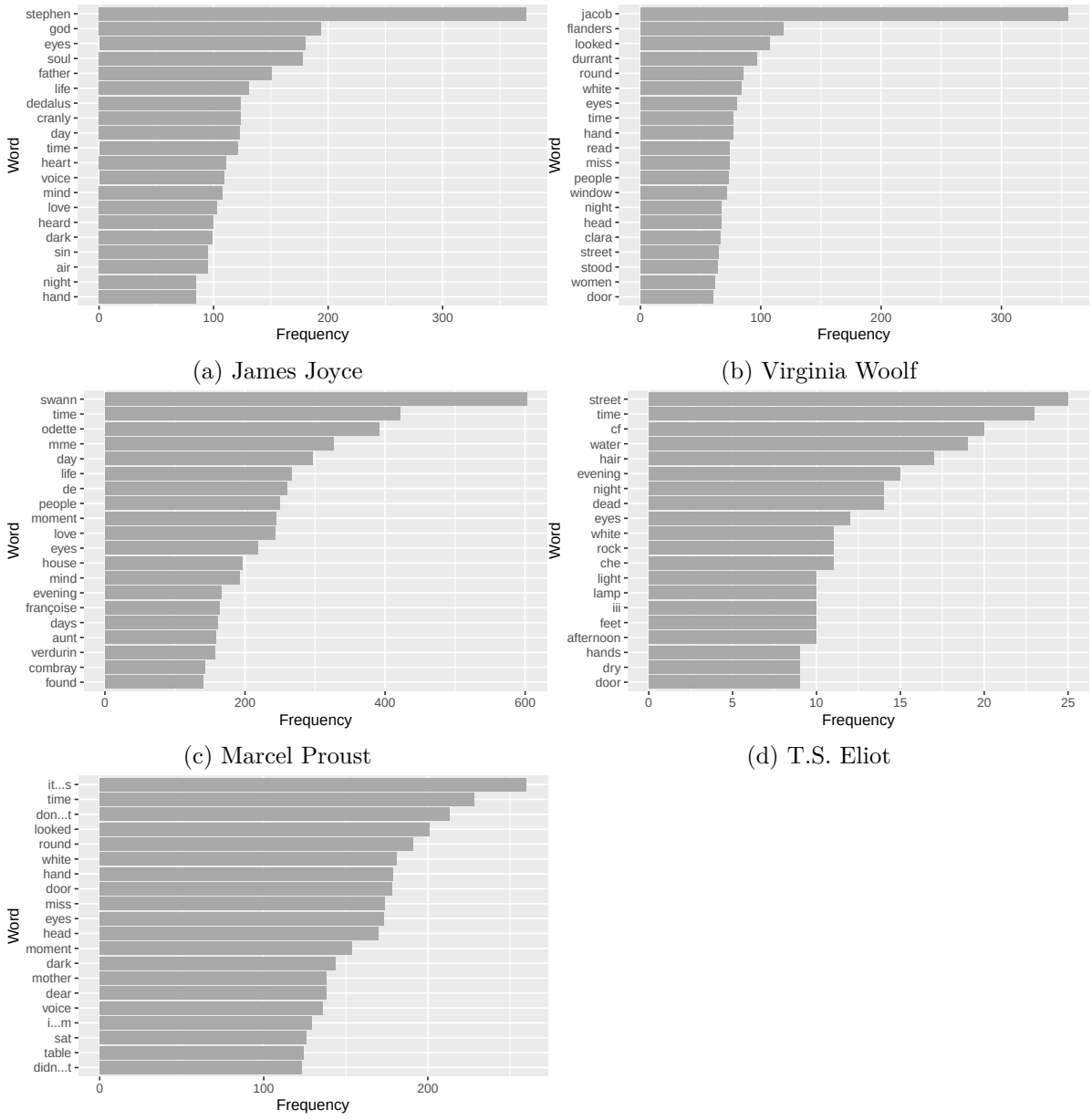
3.1.1 Model justification

We expect a negative relationship between average household income and the number of children per child care space by ward. In particular...

We can use maths by including latex between dollar signs, for instance θ.

4 Results

4.1 The Dominant Vocabulary of SOC Literature



(a) James Joyce

(b) Virginia Woolf

(c) Marcel Proust

(d) T.S. Eliot

(e) Katherine Mansfield

Figure 1: Comparative Analysis of Top 20 Word Frequencies from Famous SOC Authors

4.2 Sentiment Analysis

Leveraging setiment analysis from Mohammad and Turney (2013)

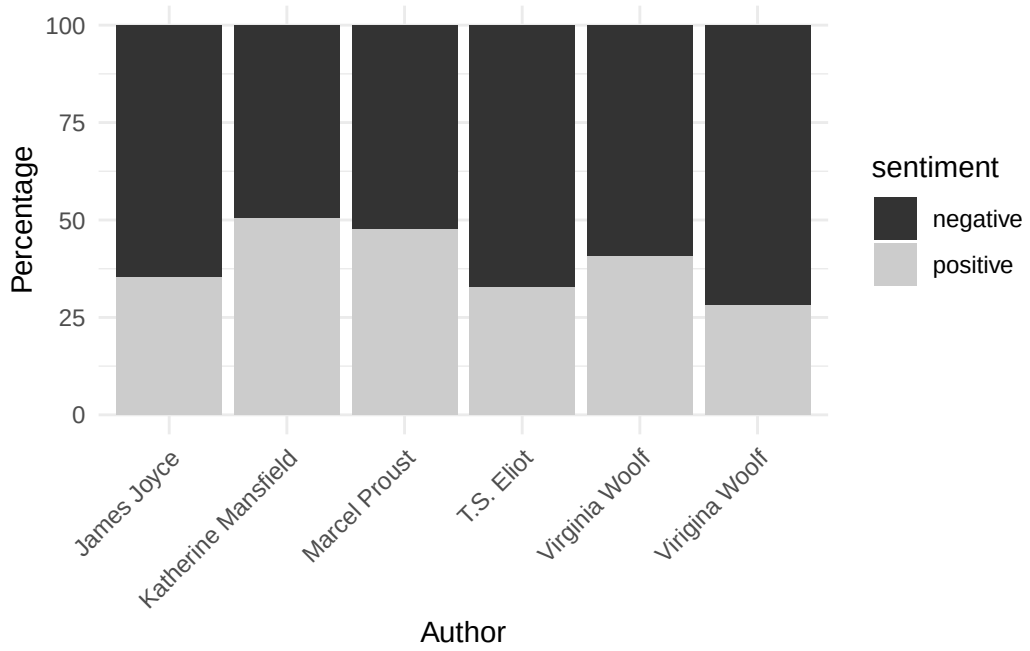


Figure 2: Sentiment Analysis of All SOC Novels

4.3 Gendered Mental Landscape of Stream of Consciousness Novels

4.3.1 Comparing SOC Literature's Female Authors

4.3.2 Comparing SOC Literature's Male Authors

4.4 Transnational SOC Novels and Mental Health Themes

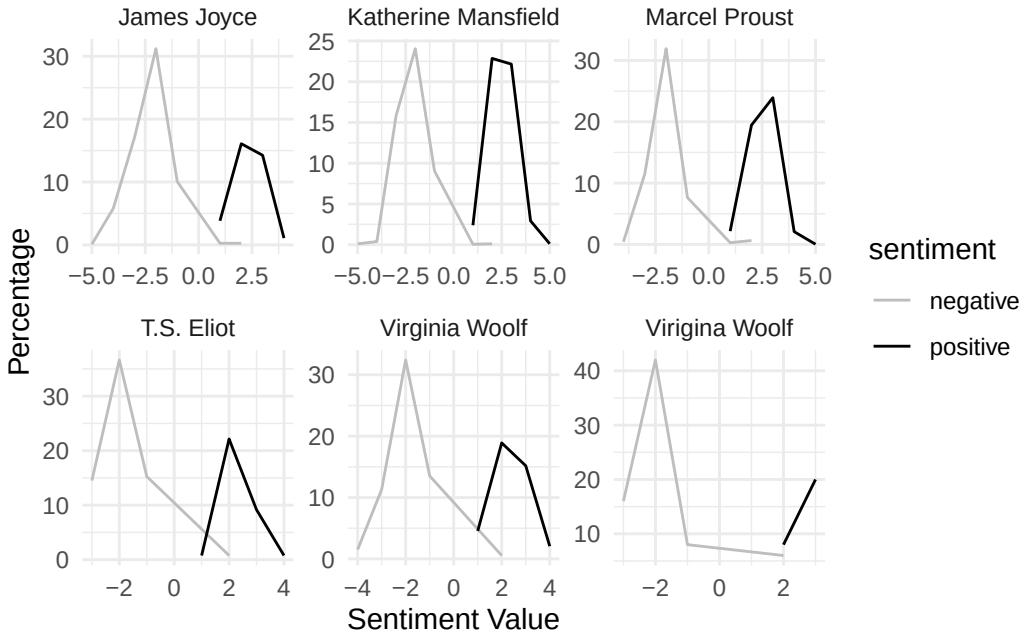
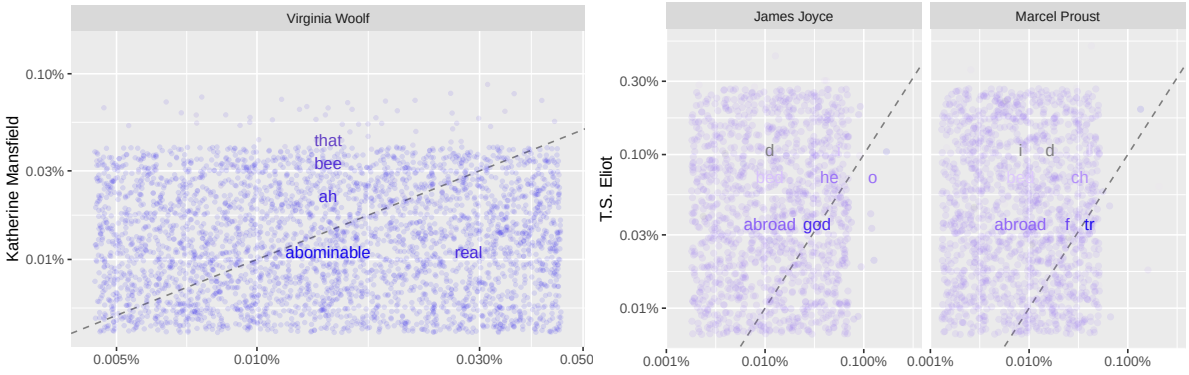


Figure 3: Sentiment Analysis of All SOC Novels by Sentiment Value



(a) Female Authors

(b) Male Authors

Figure 4: Comparative Analysis of Word Frequency in Female and Male Stream of Consciousness Authors

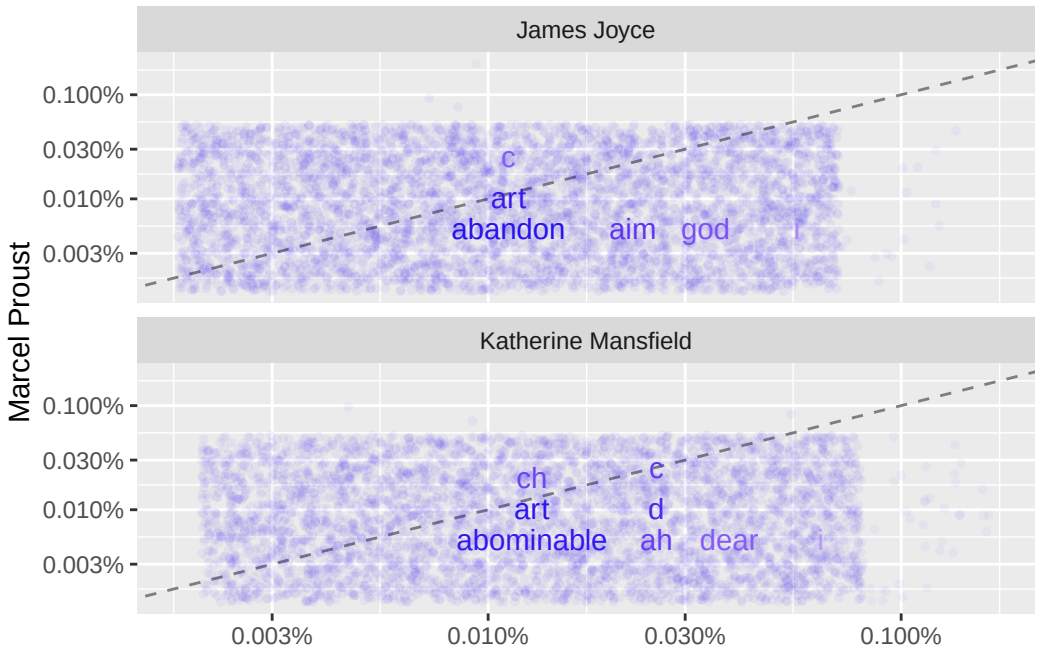


Figure 5: Comparative Analysis of Word Frequency in Transnational Stream of Consciousness Authors

4.5 Combined Texts: Trends, Word Networks, Bigram and Trigram Analsis

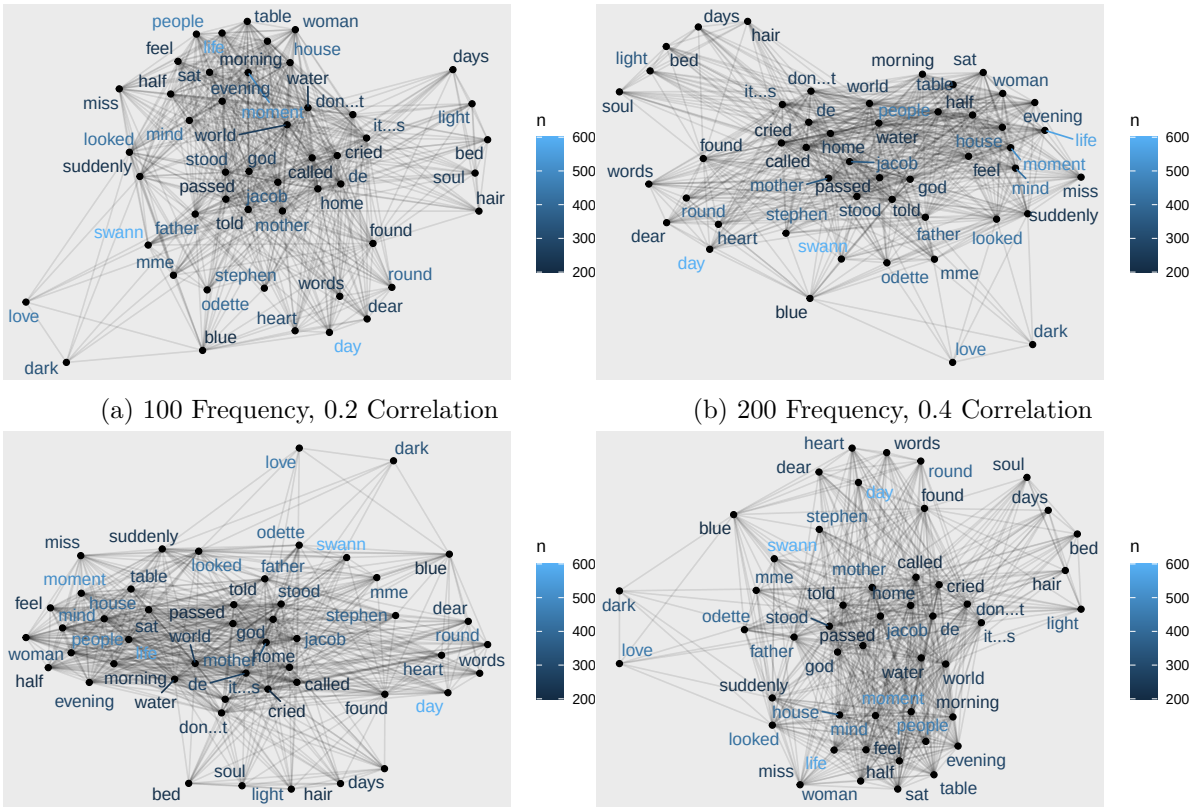
5 Discussion

5.1 Mental Health Vocabulary: Patterns and Trends

Discuss vocabulary patterns and word trends

5.2 Insights into Socio-Political Landscape of the West's Modernist Era

The novels' linguistic patterns reflect the socio-political landscape of the Western hemisphere, from late 19th century to the mid-20th century.



(a) 100 Frequency, 0.2 Correlation

(b) 200 Frequency, 0.4 Correlation

(c) 200 Frequency, 0.8 Correlation

(d) 400 Frequency, 1 Correlation

Figure 6: Word Networks Measured by Frequency and Correlation when Combining All Stream of Consciousness Novels

5.3 Schizophrenic and Disassociative Tendencies in Female Stream of Consciousness

5.4 Weaknesses

5.4.1 Lack of Thorough Word Cleaning

5.4.2 Decontextualized Literature Works and Limiting Publication Editions

5.4.3 Uneven Novel Length and Categorization of Authors

5.4.4 Project Gutenberg's Focus on the Canon

5.5 Moving Forward and Next Steps

6 Appendix

6.1 Additional Data Details

```
##| eval: true
##| echo: false
##| message: false
##| warning: false
#combined_books
```

6.1.1 Data Gathering

```
##| echo: false
##| message: false
##| label: tbl-reasons-strip-search
##| tbl-cap:
```

6.1.2 Data Cleaning

```
##| echo: false
##| message: false
##| label: tbl-items-strip-search
##| tbl-cap:
```

6.2 Model Details

6.3 Posterior predictive check

6.4 Diagnostics

References

Allaire, J., Y. Xie, C. Dervieux, J. McPherson, J. Luraschi, K. Ushey, A. Atkins, et al. 2024. *Rmarkdown: Dynamic Documents for r*. R package version 2.26. https://github.com/rstudio/rmarkdown.

Bernini, M., and C. Fernyhough. 2022. "Resampling (Narrative) Stream of Consciousness: Mind Wandering, Inner Speech, and Reading as Reversed Introspection." *Modern Fiction Studies* 68 (4): 639–67. https://doi.org/https://doi.org/10.1353/mfs.2022.0045.

Feinerer, I., K. Hornik, and D. Meyer. 2008. "Text Mining Infrastructure in r." *Journal of Statistical Software* 25 (5): 1–54. https://doi.org/10.18637/jss.v025.i05.

Henry, T. 2021. "Https://Www.youtube.com/Watch?v=ae_XVhjHd_o." https://www.youtube.com/watch?v=ae_XVhjHd_o.

Hvitfeldt, Emil. 2022. *Textdata: Download and Load Various Text Datasets*. https://github.com/EmilHvitfeldt/textdata.

Johnston, Myfanwy, and David Robinson. 2023. *Gutenbergr: Download and Process Public Domain Works from Project Gutenberg*. https://docs.ropensci.org/gutenbergr/.

Long, H., and J. So R. 2016. "Turbulent Flow: A Computational Model of World Literature." *Modern Language Quarterly* 77 (3): 345–67. https://doi.org/https://doi.org/10.1215/00267929-3570656.

Mohammad, Saif M., and Peter D. Turney. 2013. "Crowdsourcing a Word-Emotion Association Lexicon." *Computational Intelligence* 29 (3): 436–65. https://doi.org/10.1111/j.1467-8640.2012.00460.x.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Nyongesa, A. 2023. "The Centre and Pathology: Postmodernist Reading of Madness in the Oppressor in Contemporary Fiction." *Cogent Arts & Humanities* 10 (1): 1–12. https://doi.org/https://doi.org/10.1080/23311983.2023.2249280.

Pedersen, L., T. 2024. *Ggraph: An Implementation of Grammar of Graphics for Graphs and Networks*. https://ggraph.data-imaginist.com.

"Project Gutenberg." n.d. https://www.gutenberg.org.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

———. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robertson, C. 2021. "Text Tokenization in r." https://www.youtube.com/watch?v=9T-hr3jinTw.

Silge, J., and D. Robinson. 2006. "The Igraph Software Package for Complex Network Research." *InterJournal,*Complex Systems**, 1695. https://igraph.org.

———. 2022. *Widyr: Widen, Process, Then Re-Tidy Data*. https://github.com/juliasilge/widyr.

Silge, Julia, and David Robinson. 2016. "Tidytext: Text Mining and Analysis Using Tidy Data Principles in r." *JOSS* 1 (3). https://doi.org/10.21105/joss.00037.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2022. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Thomas Lin Pedersen, and Dana Seidel. 2023. *Scales: Scale Functions for Visualization*. https://scales.r-lib.org.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.