Datasheet for 'Pride and Prejudice'*

Sindhu Priya Mallavarapu Jena Shah

March 16, 2024

This paper explores the distribution of the letter 'e' in Jane Austen's Pride and Prejudice to understand if its frequency of occurrence correlates with the amount of words in the text. The methodology involves counting the occurrences of the letter 'e' in the first ten lines of each chapter gathered from the dataset obtained from Project Gutenberg. We found that the frequency of the letter "e" increases as the amount of words in the text increases. This analysis can be of interest to linguists as it can help them to identify patterns and characteristics of the English language, better understand the writing style of Jane Austen, and understand the evolution of the English language over time.

Extract of the questions from Gebru et al. (2021).

Motivation

1. *For what purpose was the dataset created?*
    - The dataset was created to analyse the frequency of the letter 'e' in the book "Pride and Prejudice" (Austen 1813) and its relationship with the number of words per line.
2. *What does this study want to help with/achieve?*
    - The study aimed to contribute to the understanding of the English language in classic literature, specifically in terms of letter frequency and textual structure.

Composition

1. *What datasets are included?*
    - The raw dataset includes the text of "Pride and Prejudice" obtained using the `gutenbergr` (Johnston and Robinson 2023) package in R (R Core Team 2022). The cleaned dataset consists of all lower cases, no whitespace, and the first ten lines of all chapters only.

*Code and data are available at: https://github.com/MSindhuPriya/letters_jane_austen

Data Collection

1. *Where is the data collected from?*
    - The text of "Pride and Prejudice" was retrieved using the `gutenbergr` package (Johnston and Robinson 2023) in R (R Core Team 2022), which provides access to Project Gutenberg texts.
2. *How much of the data is used?*
    - The first 23 chapters were used and the first ten lines of each chapter were extracted for analysis.

Data Cleaning and Analysis

1. *How was the raw data altered and changed?*
    - All text was converted to lowercase to ensure consistent counting of the letter 'e'. Whitespace characters were removed to accurately count words per line.
2. *How was the cleaned data analysed?*
    - The frequency of the letter 'e' was counted in each line, and the average number of words per line was calculated. These were then analysed to determine if there was a relationship between the frequency of 'e' and the number of words per line.

Conclusion

1. *What did the study conclude?*
    - The study concluded that there is a positive relationship between the number of words per line and the frequency of the letter 'e' in "Pride and Prejudice".

Applications

1. *How can I use this dataset?*
    - The dataset can be used to analyse the distribution of the letter 'e' in "Pride and Prejudice" and its relationship with the textual structure.
2. *How is this useful in terms of application?*
    - Researchers interested in linguistics or literary analysis can use the dataset to explore patterns of letter frequency and textual organisation in classic literature.

Distribution (Reproducibility)

1. *Can I access the dataset used in this study and/or use it to reproduce?*
    - Yes, the dataset is available for download at https://github.com/MSindhuPriya/letters_jane_austen. Researchers interested in using the dataset for their own analyses can access it too.

Maintenance

Note - Any updates or corrections to the dataset will be documented and made available to users if necessary. The creators of this paper can be contacted for further information at sindhupriya.mallavarapu@mail.utoronto.ca.

References

Austen, Jane. 1813. *Pride and Prejudice*. Penguin Classics.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.

Johnston, Myfanwy, and David Robinson. 2023. *Gutenbergr: Download and Process Public Domain Works from Project Gutenberg*. https://CRAN.R-project.org/package=gutenbergr.

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