Exploring Gender Pronoun Distribution and Character Identification in 19th Century Novels

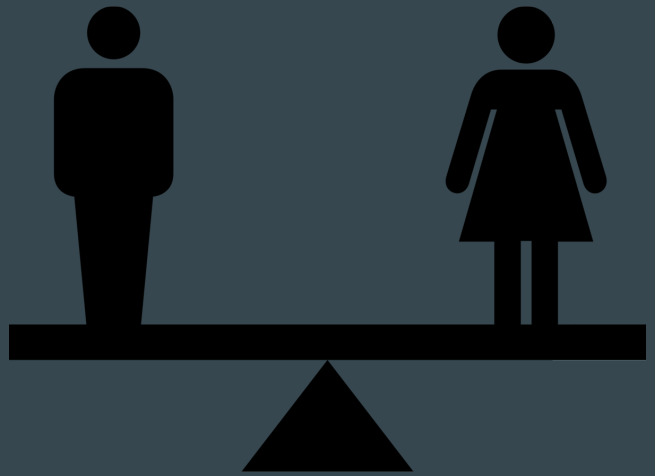
Cormac Dacker, Tyler Gomez Riddick, Avery Pike

Introduction

This study investigates two primary questions:

Is there a statistically significant difference between the frequencies of subject and object pronouns for feminine and masculine personal pronouns?

Is it possible to predict the gender of a character in a story using the pronouns associated with them?



Methodology

We used the following literary works from Calvin's Project Gutenberg repository:

- Pride and Prejudice by Jane Austen
- Frankenstein by Mary Shelley
- Wuthering Heights by Emily Brontë

Technologies

- SparkNLP for POS-tagging
- Spacy for Named-Entity Recognition (NER)
- Scipy for statistics
- R for visualization

Process

1. Corpus Preparation
2. Text Processing
3. Pronoun Counting
4. Dictionary Conversion
5. Character Filtering
6. Gender Prediction
7. Manual Verification
8. Accuracy Calculation

Hypotheses

1. Given that all books in the corpus are authored by women, we hypothesized no statistical difference between the counts of subjective male pronouns, subjective female pronouns, objective male pronouns, and objective female pronouns.

2. We hypothesized that gender prediction based on pronoun counts in relevant sentences would be more accurate than a random guess.

Evaluation

Character Gender Prediction

Accuracy was calculated by dividing the number of correct predictions by the total number of predictions. Results were split by book to evaluate model performance across different texts. We also measured accuracy by gender. Notably, *Frankenstein* achieved 100% accuracy, likely due to its smaller cast of characters.

Pronoun Frequency Analysis

We conducted statistical tests to compare the frequencies of male and female subjective and objective pronouns. Our findings indicated no statistically significant differences between the frequencies of male and female subjective pronouns (p-value = 0.74), nor between male and female objective pronouns (p-value = 0.09).

Conclusion

- Our corpus of documents is small, consisting of three books written by women
  - For future research we would want a corpus with a more diverse array of authors
- The p-value for objective pronouns is significantly smaller than for subjective pronouns, though not enough to reject the null hypothesis
- Our gender prediction model proved to be a more accurate predictor than random chance