corpuscomp

a package for analyzing characters in literary texts & trends in literary representation

STATS 290 Final Project

Motivations

- Granularity (close reading) vs breadth (computation)
- Tracking development in literary style within & across texts
- Quantifying trends in racial representation across time

Data Processing

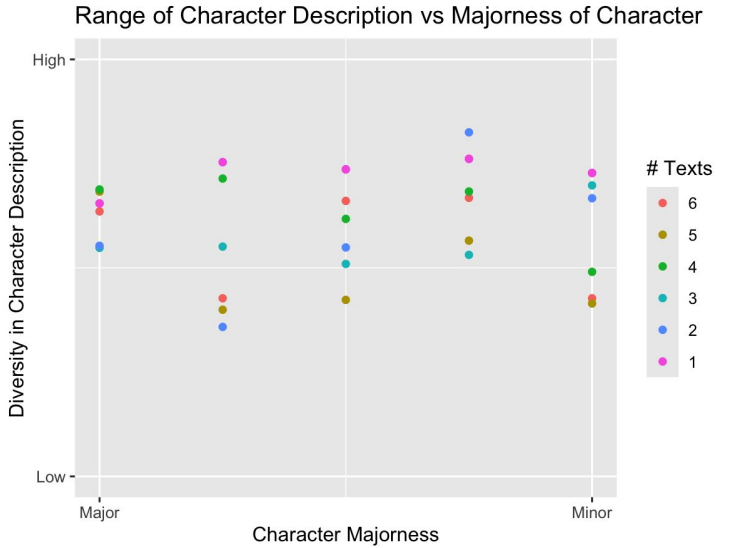
- Wrangle .json files produced by BookNLP
- Convert outputs from dependency parsing to character-level statistics
- Ready data for downstream analysis

Racial & Ethnographic Representation

- What terms do we deem “racializing?” What terms are neutral? Where do we draw this data?
- Cross references RSDB (ad hoc data source)

Sentiment Analysis

- Analyzes differences in characterization depending on character majorness



Analyzing Literary Representation Using TF-IDF and Collocation Analysis

To systematically understanding how representation evolves, we turn to computational tools like TF-IDF and collocation analysis, each offering unique insights into the nuances of literary style and representation.

TF-IDF: Capturing Frequency Over Time

TF-IDF, which stands for **Term Frequency-Inverse Document Frequency**, can help us identify terms that are significant within a specific text or character's dialogue compared to a broader corpus.

For example: By calculating TF-IDF scores across texts from different periods, we can track how often certain "racializing" terms or descriptors appear. This allows us to quantify trends in racial and ethnographic representation across time.

Collocation Analysis: Understanding Substitutability

While TF-IDF tells us what is important, collocation analysis reveals relationships between words—identifying which terms often appear together or within a defined context (the "horizon"). This is key to understanding how words function in specific narratives:

By identifying which terms co-occur with racial descriptors, we can explore how language choices reflect or reinforce stereotypes. Collocation analysis helps analyze how characters are portrayed differently depending on their narrative importance. For instance, do major characters exhibit more diverse or nuanced associations compared to minor ones?

Conclusion and Future Work

Larger-Scale Analysis: Applying the package to larger corpora, such as entire literary movements or cross-cultural datasets, would enhance its ability to uncover trends across time and space. For example, analyzing global literary traditions could reveal how representation varies across regions and historical contexts.

Diverse Datasets: Incorporating more diverse datasets—including texts from underrepresented authors, genres, or languages—would broaden the applicability of the package. This would also allow for cross-comparisons of how different cultures and traditions approach themes of race, identity, and characterization.

Integrating Advanced NLP Models: As NLP technologies continue to evolve, integrating state-of-the-art language models with the package could allow for more sophisticated analyses, such as sentiment analysis, figurative language detection, or the automatic identification of implicit racialized language.