

Data Networks of Asian American Literature, 1974-2024.

Eyad Alfaifi, Zhiliang Hu, Ryan Si, Disty Mahmud, Joyce Li

Supervised by:

Jerry Zou

Abstract:

Asian American short-form fiction remains understudied despite its rapid growth, with over 60% of works published in the past three decades. The brevity and complexity of short stories make aspects such as prose, theme, and circulation challenging to evaluate. This project applies computational methods from the Digital Humanities to trace literary trends in Asian American literature from 1974 to 2024. We construct a corpus of short stories, employ natural language processing (NLP) tools to analyze syntax, style, and themes, and produce annotated code and documentation to support future scholars in conducting quantitative literary studies.

Project objectives and goals:

- Construct a comprehensive corpus of Asian American short stories spanning 1974–2024 for computational analysis.
- Trace literary trends over five decades to understand shifts in themes, styles, and cultural representation.
- Apply NLP techniques to analyze syntax, linguistic style, and thematic elements.
- Document methods and workflows through annotated code and structured documentation to facilitate future quantitative literary studies.

Current progress and milestones achieved:

Over the ten weeks, we accomplished the following tasks:

- **Corpus Development**
    - Built a structured spreadsheet for literary magazine entries with clearly defined attributes and metadata fields.
    - Expanded the database from a single entry to over 525 short stories, covering multiple literary journals (e.g., AGNI, Georgia Review, Kenyon Review).
    - Established clear criteria for defining “Asian American short story” to ensure consistency in the dataset.
    - Converted collected works into standardized .txt files for computational analysis.

- **Automation & Data Processing**
    - Developed multiple Python scripts to calculate metrics such as average word length, words per sentence, Flesch Reading Ease, quotation percentage, and sentiment ratios.
    - Created scripts for Named Entity Recognition (NER) to extract names of people, organizations, and geographic entities.
    - Merged individual scripts into a unified processing pipeline for efficiency and reproducibility.
    - Integrated Google Sheets Apps Script and Python to autofill over **27 columns per story**, significantly accelerating metadata entry.
    - Trained a model on around 500 samples of places that were manually annotated with categories *(public, private, transitional) for facilities mentioned in each story*. After training, the model could predict whether a new input place was public, private, transitional, or should be classified as "other.", this will help for easier analysis.
- **NLP & Analytical Workflows**
    - Tested pre-trained models including Sentence-BERT, BookNLP, spaCy, and BART (MNLI-based zero-shot classification) for tasks such as POV detection, historical figure recognition, multilingual detection, and theme classification.
    - Applied topic modeling and clustering techniques including LDA, BERTopic, KMeans, HDBSCAN, and KeyBERT keyword extraction to identify themes and stylistic trends.
    - Experimented with reference story comparison for themes such as immigration, romance, and historical fiction, though results varied depending on labeling consistency and representativeness.
    - Explored embeddings at word, sentence, and paragraph-level to measure similarity between stories and refine classification strategies.
    - Conducted manual labeling of stories into core themes to create ground truth for evaluating and fine-tuning models.
    - Created exploratory visualizations, including geographic mapping of story settings and preliminary clustering plots.

Methodology and approach:

We employed a combination of Python-based NLP and data processing methods to analyze the corpus. Feature extraction was automated using spaCy and BookNLP (run in parallel for comparison) for Named Entity Recognition and character/entity analysis, while Sentence-BERT and embeddings at the word, sentence, and paragraph level were used for semantic similarity tasks. BART-large-MNLI was applied for zero-shot theme classification, and a small custom model was trained on ~500 manually annotated samples of places to classify them into public, private, transitional, or other categories. NLTK supported text preprocessing and readability metrics, while pandas and Google Sheets Apps Script facilitated data structuring and integration. For thematic and stylistic analysis, we applied topic modeling and clustering using LDA, BERTopic, UMAP, HDBSCAN, KMeans, KeyBERT, and also attempted Top2Vec. Manual labeling of a subset of stories provided ground truth for evaluating and fine-tuning models. Visualization of results was conducted with Tableau and matplotlib to identify patterns and trends, including exploratory mapping of story settings and clustering of facilities and historical events.

Preliminary results:

Early analysis of 500+ stories shows recurring themes of identity, migration, and family, with recent works also exploring technology and globalization. Over time, prose has become simpler, and story settings span both U.S. cities and international locations. Most Asian American literature in our dataset is written by women, who use emotional language more often than male authors. Additional analysis indicates that facilities mentioned in stories can be grouped into three broad categories: public, private, and transitional spaces. Zero-shot theme classification using BART performed reasonably well but lost accuracy as the number of themes increased, highlighting the need for a reduced theme list and manual ground truth labeling. Embedding comparisons showed that sentence-level vectors captured differences between stories more effectively than word or paragraph-level embeddings, though noise in unusual texts sometimes confused models. Finally, topic modeling suggested that while automatic clusters were often difficult to interpret, combining manual labeling with computational methods provided more reliable thematic insights.

Next steps and future work:

- Expand the corpus by adding short stories from additional literary magazines and archives.
- Refine theme classification models to improve accuracy and account for multiple valid interpretations of a story.

- Enhance NER pipelines to reduce false positives and improve recognition of locations, characters, and historical figures.
- Conduct deeper linguistic analysis to compare stylistic trends across decades and author demographics.
- Integrate more advanced visualization tools to map trends in themes, settings, and author backgrounds.
- Explore the dataset further to uncover additional patterns and insights that may not be immediately visible.

Resources and tools used:

- **Programming Languages & Libraries**
    - Python
    - spaCy
    - BookNLP
    - Sentence-BERT
    - NLTK
    - pandas
    - BERTopic
    - BART-large-MNLI
    - KMeans
- **Tableau**
- **Collaboration & Infrastructure**
    - GitHub
    - Google Sheets
    - Duke Compute Cluster (DCC)
- Literary magazines and journals (e.g., AGNI, Georgia Review, Kenyon Review, Shenandoah Literary)