Problem: It is difficult to find publicly-available collections of poetry. ... but... we have high-quality digital libraries of millions of books.

Observation #1: Search Engines do not solve our problem; they only find poems about poetry, or whole books of poetry.

Observation #2: Even books that are not about poetry may have poems in them. This poetry may be under-studied!

Existing Poetry Extraction Work:

- In Digital Libraries:
  - Book-level genre-labeling:
    - Content features, using summary of page-level labels (Underwood 2013)
    - Page-level labeling of many genres, including poetry (Underwood, 2014)
    - 847k books > 61k books with poetry (est. 70% recall)
  - Performance dominated by terms for selecting Genre.
- In Newspapers:
  - Image Processing (Lorang et al, 2015).
  - Found Underwood’s content approach to be low recall.

What does the model get wrong?

- Plays or Dialogue are hard for both the model & humans.
- Interviews “look like Poetry” from far away...
- Advertisements may be poetic, but may not be of interest.

Acknowledgements

This project began at the University of Massachusetts with the help of REU students Simon Everhead and Caroline Buck. The classifier and relevant datasets were designed as part of my dissertation there. Work on this project has continued with the help of my students from Smith College, including Ananda Montoly, Mayeline Peña, Eindra Kiy, and Sivan Nachum. Work continues as I begin my position as an Assistant Professor of Computer Science at Middlebury College.

References


A deeper discussion of related work available in the abstract & (Foley 2019).

Live Access to the Poetry50k Dataset:

Search Poetry: https://poetry.jf Foley.me/ or Tweet @poetrysearchbot

A "visual model" of Poetry

We explored several strategies for detecting which pages are poetry and which are prose (or other-genres) within each book; the most successful were pseudo-visual features, similar to work done by Kilner & Fitch (2017). A reviewer identified this area as digital layout parsing.

After labeling 2,700 pages, we were able to train a Random Forest model over these features; we wrote our feature extraction code in Rust for efficiency. Estimated 0.89-0.94 AUC-ROC.

Extracting the Poetry50k dataset

Running this model on a set of 50,000 books gives us a collection of about 600,000 poems, after processing and removal of duplicates (popular poems were found to be printed in many volumes).

What can we study with this poetry?

- Studying term evolution & usages over time.
- Studying popularity of poetry in the eyes of different publishers.
- Studying poetic devices such as allusion, simile, and metaphor.

Get in touch! The dataset is public and I’m happy to help you get started.