fryetag

natural literature processing

STEPHEN BRAND | BO CHENG | JONATHAN COMPTON | COREY HAVERDA | DANIEL KIM | WILL TAYLOR

The Problem

Project Gutenberg (PG) today is a collection of 60k+ e-books across 100s of 'bookshelves' with limited ability to see connections between works, especially based on the textual contents of a work. Book lovers can quickly find authors they know about, but there isn't a feasible way to know which other authors have similar word and topic choice. Plus, PG's site is simply a catalog and fails to visualize its library with the kind of fresh, sleek, or interactive interface needed to invite casual readers.

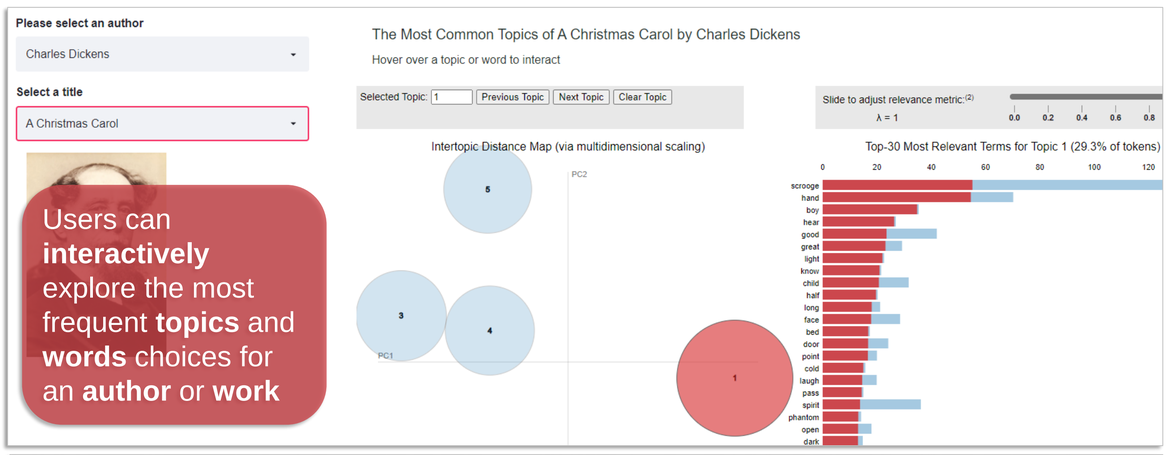
Great… but who cares?

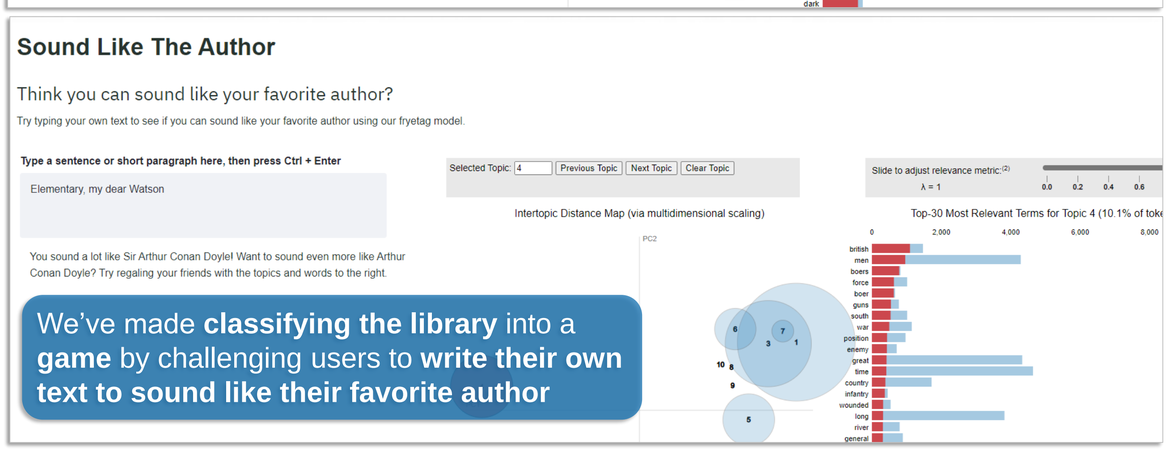
Book Lovers → … can **discover books & authors** they enjoy, potentially **creating social benefits**

Educators → … can **share literature** in a more meaningful, engaging way

Students → … can **explore literature** in an interactive experience

Academics & Librarians → … may uncover **word and topic selections** via visualization of hidden content characteristics





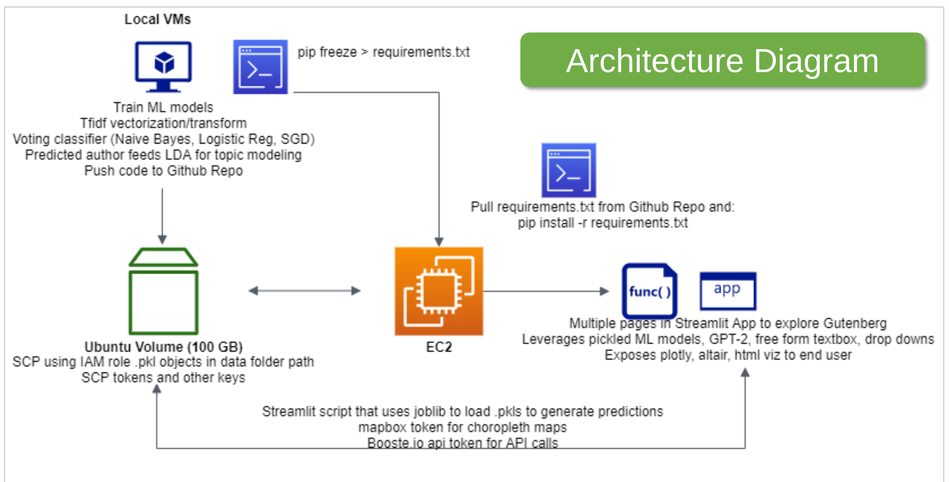
Our Approach

Our literature review research gave us the intuition for the methods described below. These have been used for other text-based problems but are new for exploring literature on the scale of the PG library.

- ***regex*** and ***nltk*** to clean the data, tokenize, and identify & remove stop words
- ***sklearn*** pipeline to prevent data leakage and tfidf vectorization to transform text into a numerical space.
- Logistic regression, multinomial naïve bayes, and stochastic gradient descent were tested on their own, but we found that using a voting classifier that incorporates all these models to be most successful for predicting authors based on snippets of text.
- Reduced the sample size from the entire Project Gutenberg library to the top 20 authors in order to facilitate further analysis.
- For each of these 20 authors, we used a latent dirichlet allocation (LDA) model to identify key topics & frequently used terms, then output the analysis using ***ldaViz***

How does this solve the problem?

With these core building blocks, we were able to use the streamlit app to create an interactive visualization tool (shown above) that helps users explore the library based on authors, topics, and frequently used words. This allows users to explore the PG library in new and intuitive ways and gives them a peek into the types of worlds authors create before committing to the book. These methods are the first steps to extracting more features from the text such as characters, locations, and themes, which we hope to continue to pursue, but found to be beyond the scope of this class. Eventually, these methods can be applied to other libraries and collections.



The Data in Focus

464 MB
Size of downloaded University of Michigan Project Gutenberg cleaned data set

SPARQL
scraping from Wikidata for author's places of birth and images

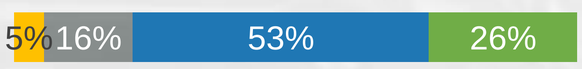
20
Authors analyzed

1312
Full-length novels & essays analyzed

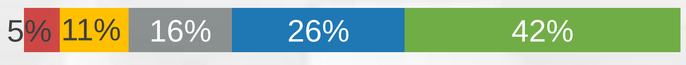
Experiment & Results

We evaluated our approach by asking participants (*n=19)* to spend time on the Project Gutenberg website and our application before answering a 13-question survey. Participants answered the questions according to a five-point Likert scale (*1 – Strongly Disagree* to *5 – Strongly Agree).* Compared to existing methods, the majority felt Fryetag allowed them to explore authors in a new way *(Figure 1)* and that Fryetag's new analytically driven approach would help them find books they'd love *(Figure 2)*. Sixty-eight percent said our visualizations balanced depth and simplicity and that our 'Sound Like The Author' was a unique interactive tool they could keep coming back to. Finally, most participants managed to find 1 or 2 topics and 10-12 common words per book on Gutenberg's website, while our application immediately supplied 5+ topics and 30+ words per book.

(1) Fryetag allows me to explore authors in new ways that were previously unavailable



(2) The analytically driven approach to library exploration offered by Fryetag will help me find books I love.



Scale: Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree