# Pirating Texts

Grant Glass, Gabriel Guedes, Lucian Li, Orgil Batzaya

## Objective

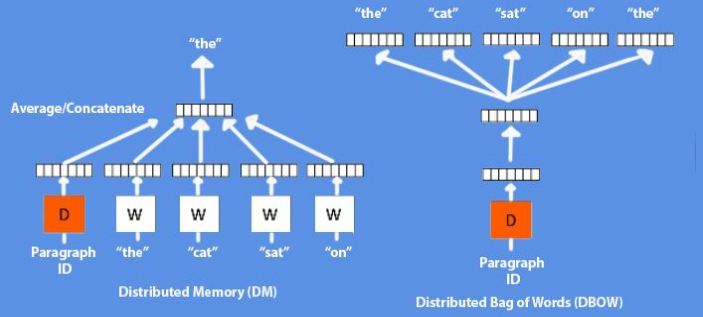
- Trace the geographic spread of "pirated" copies of Daniel Defoe's *Robinson Crusoe* over time and draw conclusions about how the historical context of publication impact the content of the copies
- Identify the most important parts of the *Crusoe* story that persist despite differences between volumes

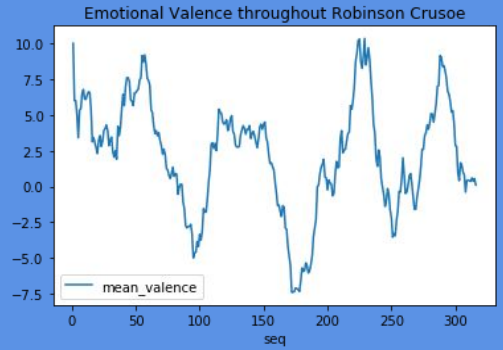
## Methods

**Web scrape** editions & metadata, create **Doc2Vec** (unsupervised algorithm which uses 2 approaches to embed documents as vectors) models linking similar editions, track **sentiment**, and build **map** with CartoDB

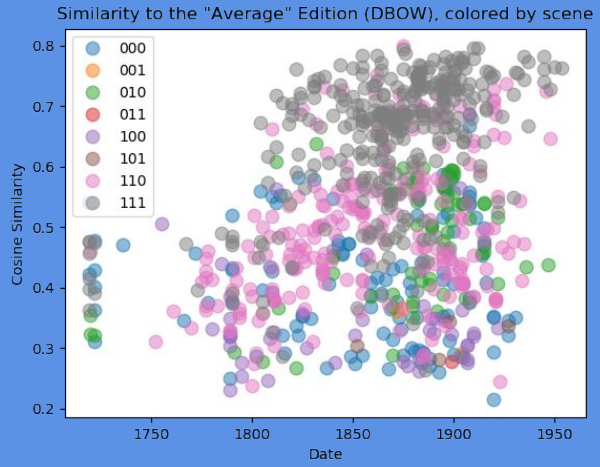
## Data Collection

16,319 metadata entries (city, year, language, country, etc.) & 1,482 full-text editions from University of Florida, HathiTrust, and the Internet Archive



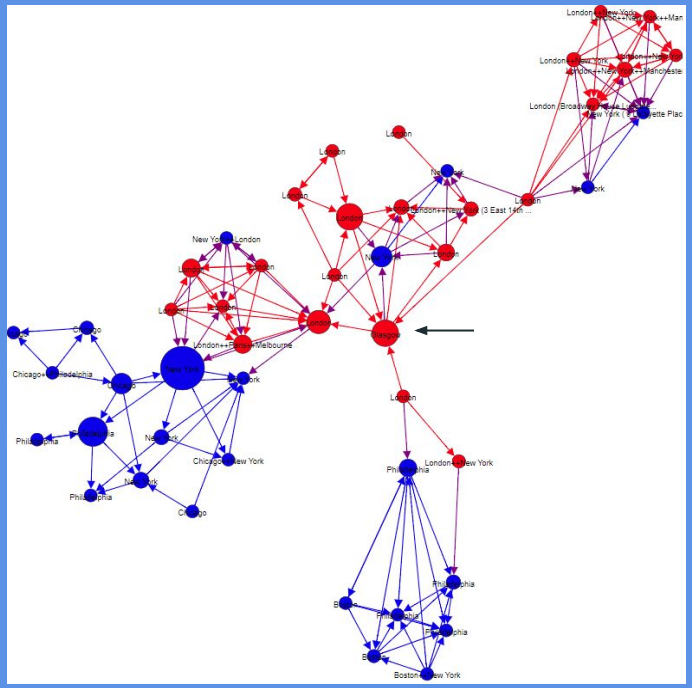



| | |
|---|---|
| 1st digit | Enslavement in Morocco |
| 2nd digit | Discovery of the footprint |
| 3rd digit | Post-rescue Pyrenean adventure |

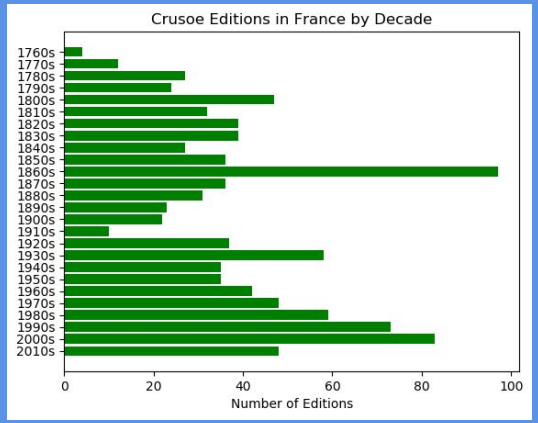


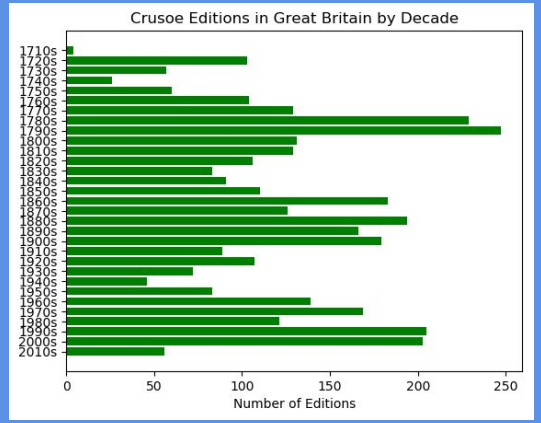

| Label | date | betweeness centrality |
|---|---|---|
| New York | 1897 | 39.5 |
| Philadelphia | 1908 | 21 |
| London | 1869 | 17.5 |
| Glasgow | 1875 | 17.5 |
| London | 1910 | 13.5 |

The 1875 Glasgow edition was the most similar to the average (DBOW) and also the most centrally located node in the largest connected component



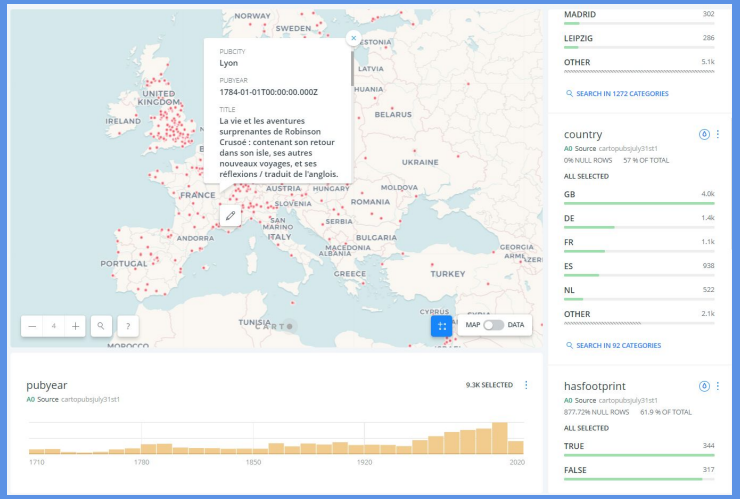
We used the Doc2Vec results to generate a vector representing the average qualities of the corpus. This enabled us to analyze the relative importance of the inclusion of specific plot points in influencing a book's proximity to the average.



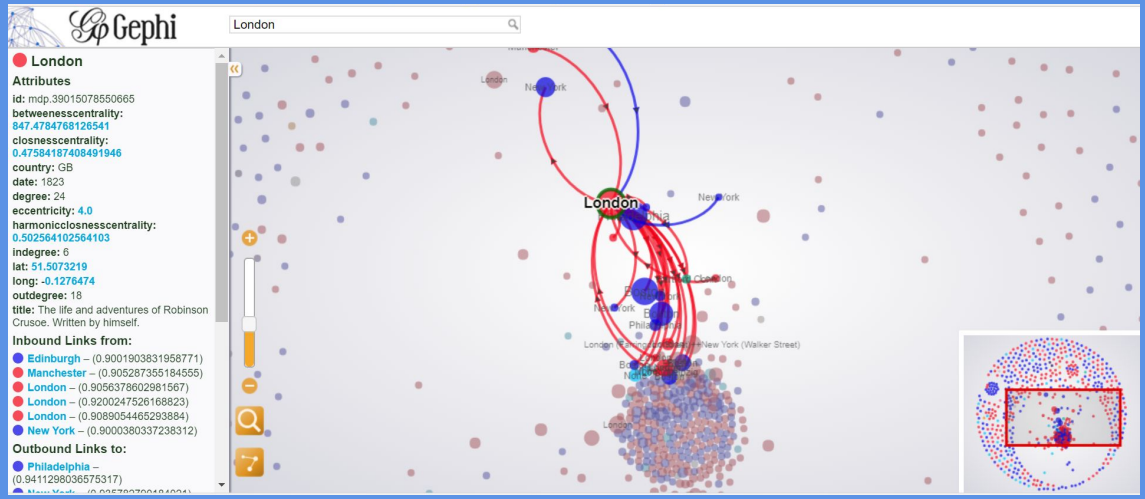




The peaks in FR and GB correspond to periods of elevated jingoism and imperial confidence. In France, the 1860s saw a wave of colonial adventurism in Algeria, Mexico, and Indochina. At the end of the 1860s, French prestige suffered a massive blow with the disastrous Franco-Prussian war, and did not recover until WW1, a trend reflected in the graph. In Britain, we see peaks in the 1780s and 1880s, corresponding to the First and Second Industrial Revolutions respectively. These two decades also marked peaks in colonial power with centralization of power in India and rapid expansion in the Scramble for Africa.

**Product 1**
Interactive map built with Carto and populated with our database of processed metadata.
Includes time-series, filtering, and widgets .




**Product 2**
Interactive network visualizer built from Gephi & Doc2Vec embeddings in which an edge is created if cosine similarity > 0.90. Using this tool, researchers can trace lineages of text transmission and "piracy" through date-based directed edges




**Product 3**
Pipeline and integrated database of books tying together texts, metadata, scene info, and Doc2Vec metrics. Allows for relatively simple and flexible analysis across different metrics and datasets and provides a starting point for future researchers to carry out cleaning, segmentation, analysis, and visualization.

**Future Research:**

- Compare Crusoe publication patterns per country with overall publication data and other books
- Improve segmentation techniques to automatically categorize presence or absence of specific scenes across the entire corpus
- Conduct more complex graph analysis on the networks using Python NetworkX package (maximal cliques, shortest path, network flows, bridge identification, and topological sorting)