Investigating writers' attitudes by mining a large corpus of books

Sayan Bhattacharyya (sayan@illinois.edu), **Postdoctoral Research Associate, Graduate School of Library and Information Science**

Rini Bhattacharya Mehta, Assistant Professor, Program in Comparative and World Literature, University of Illinois at Urbana-Champaign

GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE
The iSchool at Illinois

Introduction

Researchers in history or literary studies are often interested in the question of attitudes of writers to some specific subject matter, and seek an efficient discovery mechanism that can identify which texts in a large collection of texts would repay close study for the purpose of exploring this question. Our work-in-progress approaches this problem in a scalable way by combining a search for collocations within the corpus using a list-based approach, with filtering using available bibliographic metadata.

Rationale:

With the digitization and online availability of large quantities of text from the contents of academic libraries, much recent interest has focused on algorithmic analysis of large aggregate collections of text, which has come to be termed "distant reading" (Moretti, 2013). However, only a relatively few scholars in the humanities are likely to pursue this kind of large-scale algorithmic analysis. Most scholars perform "close reading" of selected texts. For such scholars, a **discovery mechanism** for those few selected texts within the entire corpus that would repay close reading for the purpose of their research question, is much more useful.

Use Case: Investigating Writers' Attitudes

Our use case involves investigating the attitudes of French-language and English-language writers towards women's work, especially in the colonized world.

Corpus: The corpus for our use case is the HathiTrust corpus, which consists of digitized text from the collections of numerous academic research libraries that are part of the HathiTrust consortium.

We are building a tool which will provide a discovery mechanism for researchers trying to identify those volumes within the corpus that will repay closer study when the object is to discover particular kinds of writers' attitudes about specific composable topics, within a certain range of interest (e.g. language, time-period).

The Use Case's Exemplars of Research Questions

A historian or a scholar in literary studies may be interested in the following kinds of questions, all of which would involve identifying texts of interest that can throw light on the questions. Below are some example research questions for our use case:

1. What was the attitude of French writers (from metropolitan France) about women's work in the French colonies in the nineteenth century? Which books from that period will most reay close study in answering this question?
2. What was the attitude of British writers (from metropolitan Britain) about women's work in the British colonies in the nineteenth century? Which books from that period will most reay close study in answering this question?
3. What was the attitude of writers from the French colonies in the nineteenth century, about women's work in the French colonies in the nineteenth century? Which books from that period will most reay close study in answering this question?
4. What was the attitude of writers from the British colonies in the nineteenth century, about women's work in the British colonies in the nineteenth century? Which books from that period will most reay close study in answering this question?
5. For each of the cases 1-4 above, what were the corresponding attitudes in the early twentieth century? Which books from that period will most reay close study in answering this question?

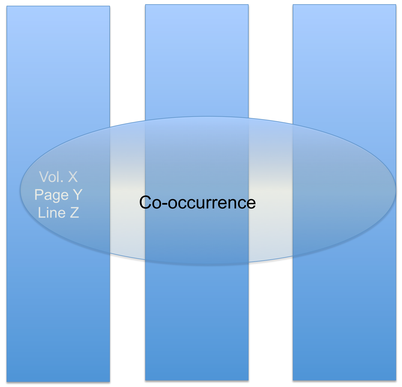
Generalization: A generalization of this use case can answer the more general question of the form:

What was the attitude of writers satisfying the criteria {A,B,C…} about the intersection of interests {D,E,F,…}. (Here, D=women, E=work, F= (colony) or (metropolis), etc.)

Method

Our approach consists of the following:

- We generate separate lists of occurrences (indexed by line number, page number, and book id number) of all occurrences of words relating to womanhood (and a set of close synonyms), of all occurrences of words relating to work (and a set of close synonyms), and of all occurrences of words expressive of attitudes.
- We then identify, based on these separate lists, instances of co-occurrences of all three items within a determinate proximity window. This list of co-occurrences serves as the basis for the discovery mechanism for identification of relevant texts, as well as for aggregate-level analysis enabling comparative measures.



List1: List of all occurrences of words relating to womanhood (and a set of close synonyms)

List2: List of all occurrences of words relating to work (and a set of close synonyms)

List3: List of all occurrences of words expressive of attitudes.

All occurrences in the lists are indexed by line number, page number, and book id number.

Discovery Process:

Books (volumes) initially selected on the basis of available metadata.

Lists generated by scanning through each page content of the selected volumes

Co-occurrences identified by comparing the index of each content item in a list against those in the other lists.

Books (volumes) ranked by order of how many co-occurrences they are involved in.

Top *N* books then furnished to researcher as the "most promising" books that are likely to repay close reading. These *N* books, thus, are "discovered" by the discovery mechanism.

Scalability

The discovery mechanism is reasonably scalable with regard to additional parameters (leading to additional lists). The addition of a parameter (the creation of an additional list) means that, during co-occurrence detection, that list must also be scanned. While this makes co-occurrence detection theoretically exponential in the number of parameters, in practice the number of parameters is unlikely to exceed 3 or 4. Additionally, we are examining mechanisms and heuristics that could lead to further efficiency.

Acknowledgments

The research described here is generously supported by the Mellon Foundation and by the HathiTrust Consortium.

ILLINOIS