Investigating writers’ attitudes by mining a large corpus of books: Preliminary research

Sayan Bhattacharya (sayan@illinois.edu), Post-doctoral research associate, Graduate School of Library and Information Science, UIUC
Rini Bhattacharya Mehta, Assistant professor, School of Literatures, Cultures and Linguistics, UIUC

Abstract
Given a large corpus of digitized books, such as the HathiTrust corpus, it is useful for researchers in history or literary studies for have an efficient discovery mechanism that can identify which texts from the large collection of texts is likely to repay close study, in connection with their specific research question. Also, many research questions that interest researchers can often be envisioned as compositional, made up of multiple distinct parameters. Our work-in-progress approaches the issue in a scalable way by utilizing the compositionality of research questions to treat the problem as one of searching for collocations using a list-based approach. This approach will be of use especially to researchers who have to work within the restrictive limitations on access on derived data likely to be imposed by copyright rules and their implications. Our use case involves investigating similarities and differences in attitudes of French language and English-language writers towards women’s work in the colonized world. Our approach is to use co-occurrences of the elements that constitute the decomposable content of this investigative question, after starting with available bibliographic metadata as an initial filter. We are also exploring methods of validation and tuning, as well as analytical comparison across defined segments of the dataset. Our use case involves investigating the attitudes of French-language and English-language writers towards women’s work, especially in the colonized world.

Rationale
With the digitization and online availability of large quantities of text from the contents of academic libraries, there has been much recent interest 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 that can help locate those few selected texts within the entire corpus that would repay close reading for the purpose of their research question, is much more useful. Analytic comparison between different subcorpora so discovered (e.g. by language, or by historical period, by gender of authors, etc.) is of much use to comparatist scholars.

Use Case: Investigating Writers’ Attitudes
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 and analytic comparison facilities 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 topics that can consist of constitutive elements (compositional units composed together), and lying within certain specified ranges of interest (e.g. language, time-period). For our use case, the languages are French and English, the genre of texts is travel-oriented writing, the time period is the nineteenth and early twentieth centuries, and the composable topic is writer’s attitudes towards the work performed by women in the colony and/or the metropolis.

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 repay 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 repay 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 repay 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 repay 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 repay close study in answering this question?

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

What is the attitude of writers satisfying the criteria \( \{A, B, C, \ldots\} \) about the intersection of the composable elements \( \{D, E, F, \ldots\} \), where \( A = \text{period of publication} \), \( B = \text{genre} \), \( C = \text{language of the text} \), etc., and \( D = \text{women} \), \( E = \text{work} \), \( F = \text{colony or metropolis} \), etc.?

Method
Our approach consists of the following:

• Generate separate lists of occurrences (indexed by 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, within the initial, metadata-based, selection.

• 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.

• Ideally, We would like to identify instances of co-occurrences of all three items within a determinate proximity window based on these lists.

Intended Workflow:
• Books initially selected on the basis of available metadata.
• Lists generated on the basis of book-level (volume-level) frequencies of words (and set of synonyms).
• Co-occurrences identified by comparing the index of each content item in a list against those in the other lists.
• Initially selected books ranked by order of number of co-occurrences each book is involved in.
• Top N books identified as the “most promising” books that are likely to repay close reading. These N books, thus, are “discovered” by the discovery mechanism. Domain expert (likely to be the scholarly researcher end-user in most cases) evaluates results to check for obvious discrepancies (expected, “canonical” texts not showing up) and tweaks sets of synonyms as needed.

Scalability
The discovery mechanism is reasonably scalable with regard to additional parameters (leading to additional lists) in the service of end-user’s research questions that are compositional in parameters. (In the example in this poster, there were three parameters, D,E,F.)

Any addition of a new parameter (the creation of an additional list) means that, during co-occurrence detection, a new list corresponding to that parameter must also be searched for collocations. 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.

References:

Acknowledgments
The research-in-progress described has been undertaken as part of the HathiTrust, which receives funding from the Sloan and Mellon foundations. We thank the HathiTrust Research Center team, and especially Loretta Auvil, for suggestions.