Distant Reading Topic Modeling in Historical Research

Mats Fridlund & René Brauer
History of Industrialization & Innovation Group
Aalto University

Historicizing Topic Models

A distant reading of topic modeling texts within historical studies

René Brauer* & Mats Fridlund**

*Department of Earth Sciences
University of Gothenburg
SE-405 30 Gothenburg, Sweden

**Department of Engineering Design and Production
Aalto University
FI-00076 Aalto, Finland
Corresponding author:
mats.fridlund@aalto.fi

Abstract

Topic modeling (TM) is a method used within the new 'digital history' that represents a data driven methodology that might be closest to fulfilling literary historian Franco Morettl's promise of making possible 'distant reading' of large text quantities. Inspired by this promise, TM has been used for historical studies since the early 2000s and this study provides a survey of the state of the art of TM among historical studies by giving a historical and methodological introduction into the use of TM within historical minded research.

TM's was first being developed for data mining within natural language processing and machine learning in the 1990s and had as its overwhelming benefit its ability to cover magnitudes more of data as compared to traditional methods. The primary topic model used is the Latent Dirichlet Allocation that allows TM to be used as a search function, a quantitative check of intuition or as a summarization tool for large corpora of texts. Having many competing theories and assumptions that are constantly being challenged and developed TM in itself currently represents a very active area of research within computer science.

The survey of historical texts take its starting point as the first peer-reviewed historical article in 2006 and end point the publication of the firs research monograph in 2013 and identified 23 historical studies employing TM. To provide a general overview of the field the studies were examined using a distant reading quantitative approach and analyzed according to authors' academic background, gender, academic seniority and country of academic institution; corpora's type, language, chronology, and geographical focus. The results showed most authors being junior untenured male researchers, primarily affiliated with US-universities and the texts consisting of a substantial number of non-standard online texts. Despite the application within historical studies TM still comes across as a technology driven approach with majority of authors having a background in technical disciplines. Corpora where primarily focused on English texts with a US or global focus and with an emphasis on recent history. All in all TM appear to an emergent rather than established historical methodology.

Keywords: topic modeling, digital history, digital humanities, historical methodology, Latent Dirichlet Allocation

matsfridlund.com/digital-humanities-finland/

Overview

– History's second digital revolution?

– Distant reading & 'not-reading'

– Topic Modeling

– The history of Topic Modeling in history, 2006-2013

– A 'Distant Reading' of historical TM articles 2006-2012

– Conclusion

History's second digital revolution?

'quantitative revolution' & 'digital history'

kind of culture war broke out in the profession and a flurry of tense conference panels, public arguments, and roundtables took place with subtitles, such as "The Muse and Her Doctors" and "The New and the Old History." This culture war pitted the "new" history, largely influenced by social science theory and methodology, against the more traditional practices of narrative historians. The "new" historians used computers to make calculations and connections never before undertaken, and their results were, at times, breathtaking. Giddy with success, perhaps simply enthusiastic to the point of overconfidence, these historians saw little purpose in anyone offering resistance to their findings or their techniques. When challenged at a conference, more than one historian responded with nothing more than a mathematical equation as the answer. (Thomas 2004, 56)

Distant reading & 'not-reading'

As long as there have been books there have been more books than you could read. In the life of a professional or scholar, reading in the strong sense of "close reading" almost certainly takes a back-seat to finding out what is in a book without actually reading all or even any of it. There are age-old techniques for doing this, some more respectable than others, and they include skimming or eyeballing the text, reading a bibliography or following what somebody else says or writes about it. Knowing how to "not-read" is just as important as knowing how to read. ... A provisional answer to the question what metadata are good for, then, might say that metadata ... let you condense not only a single text, but in a sufficiently ample environment they let you condense arbitrarily large sets of texts. And if you employ visualization techniques - an increasingly powerful digital tool - these condensed representations can be displayed as if they were locations on some map. Just as white space in a book with good layout maps the terrain of the pages and orients readers before they actually "read", so metadata, when "laid out" in the right way can provide readers with a simultaneous overview of many books and direct their attention to areas where it would pay to read closely. That is the promise of Franco Moretti's "distant reading." (Mueller 2007)

Topic Modeling

Put in simpler terms, a *topic model is a computer aided program that from a text generates 'topics' or 'themes': strings of words that are supposed to be indicative of themes addresses within the text*. The basic idea is that words that cluster 'closely' share a meaningful connection, i.e. a 'topic', 'theme' or 'motif' of a text, which in lay terms could be understood as the 'important' or 'significant' key words of shared theme.

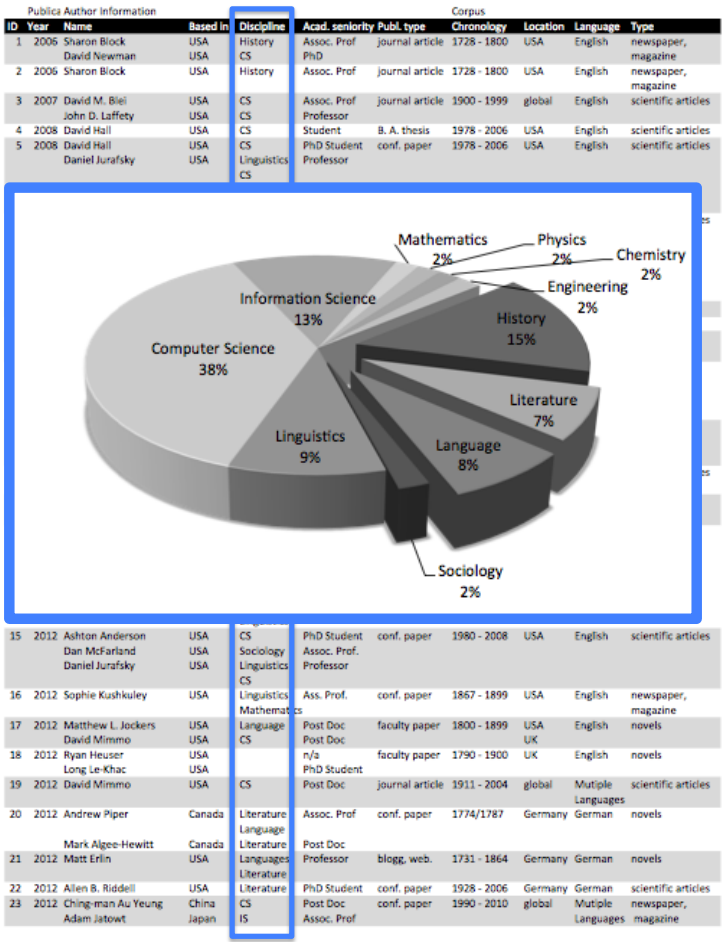
- Latent Semantic Analysis (Deerwester 1990)
- **Latent Dirichlet Allocation (LDA)** (David Blei 2003)

- **MALLET – Machine Learning for Language Toolkit**
  - Number of topics
  - Chunk size
  - **Rule of thumb:** 100 topics & document chunks of 1.000 words

- Best fit to corpus
- Interpreting meaningfull or not

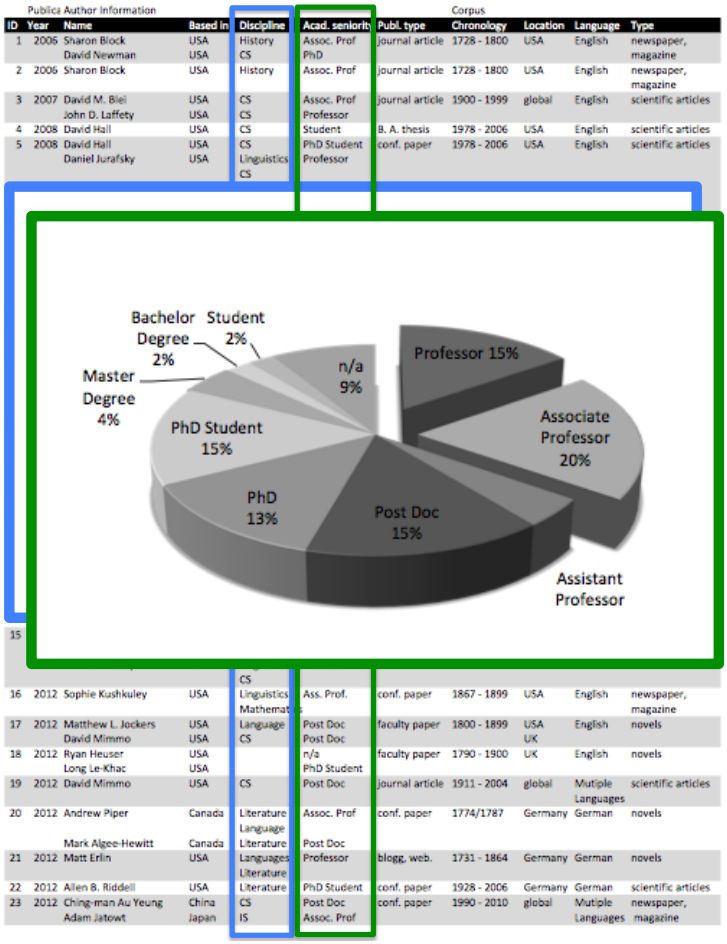
The history of TM in history, 2006-2013

- **2006:** First peer-reviewed article by historian using TM
  - **Newman, D.J., & Block, S. (**2006). Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper, ***Journal of the American Society for Information Science and Technology***, 57:6, 753-767.
    - **(Griffiths, T. L., Steyvers, M. (**2004). Finding scientific topics, *Proceedings of the National Academy of Sciences of the United States of America*, 101:1, 5228-5235.)
- **2006-12: 22 historical publications using TM**
- **2013:** First historical monograph using TM
  - **Jockers, M.L.** (2013). ***Macroanalysis: Digital methods and literary history***. University of Illinois Press.



Patterns

- Men
- N. American (92%)
- **Tech-driven (60%)**

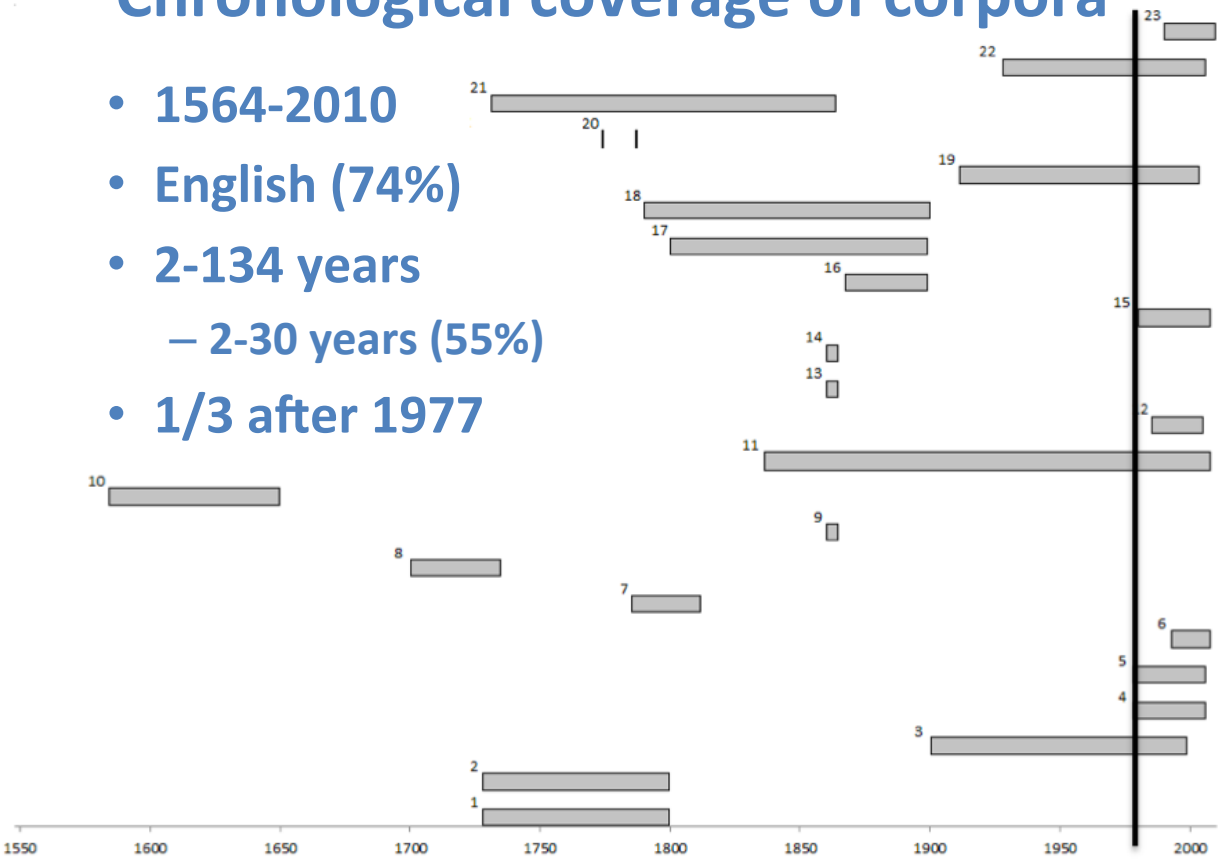


Patterns

- Men
- N. American (92%)
- **Tech-driven (60%)**
- **Junior (56%)**

Chronological coverage of corpora

- 1564-2010
- English (74%)
- 2-134 years
  - 2-30 years (55%)
- 1/3 after 1977

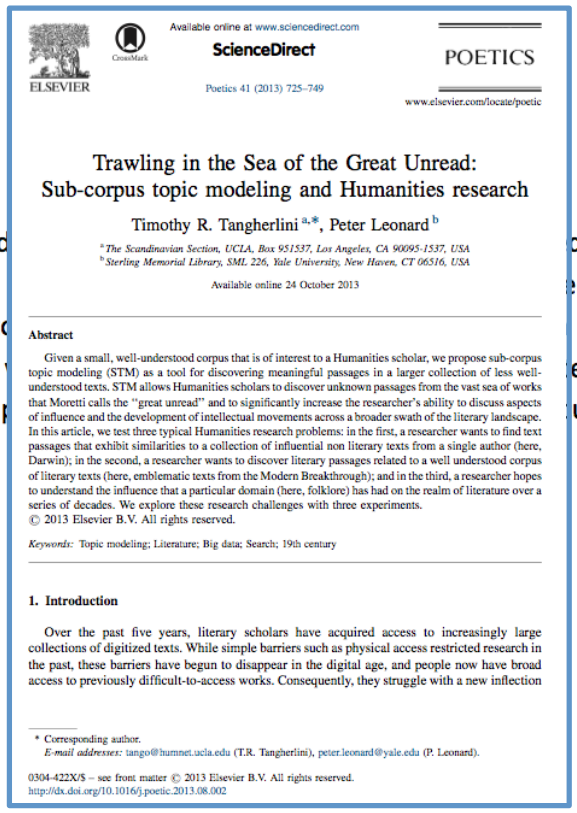


Concluding discusion: Retooling history?

As it is now topic modeling is primarily being developed and explored rather than utilized as a reliable historical method. And although representing an interesting and promising methodology for historical applications is still very much a solution in search of its perfect problem to prove its value to historians. Or perhaps better, it is a technology in search for the historical killer app that will make it into a necessary sharp cutting-edge tool in the historian's toolbox.

Concluding discusion: Retooling history?

As it is now topic mod d rather than utilized as a reliable historical eresting and promising methodology for hist in search of its perfect problem to prove its echnology in search for the historical killer a utting-edge tool in the historian's toolbox.



Available online at www.sciencedirect.com
ScienceDirect
Poetics 41 (2013) 725–749
POETICS
www.elsevier.com/locate/poetic

Trawling in the Sea of the Great Unread:
Sub-corpus topic modeling and Humanities research

Timothy R. Tangherlini [a,*], Peter Leonard [b]
[a] The Scandinavian Section, UCLA, Box 951537, Los Angeles, CA 90095-1537, USA
[b] Sterling Memorial Library, SML 226, Yale University, New Haven, CT 06516, USA

Available online 24 October 2013

Abstract

Given a small, well-understood corpus that is of interest to a Humanities scholar, we propose sub-corpus topic modeling (STM) as a tool for discovering meaningful passages in a larger collection of less well-understood texts. STM allows Humanities scholars to discover unknown passages from the vast sea of works that Moretti calls the "great unread" and to significantly increase the researcher's ability to discuss aspects of influence and the development of intellectual movements across a broader swath of the literary landscape. In this article, we test three typical Humanities research problems: in the first, a researcher wants to find text passages that exhibit similarities to a collection of influential non literary texts from a single author (here, Darwin); in the second, a researcher wants to discover literary passages related to a well understood corpus of literary texts (here, emblematic texts from the Modern Breakthrough); and in the third, a researcher hopes to understand the influence that a particular domain (here, folklore) has had on the realm of literature over a series of decades. We explore these research challenges with three experiments.

Keywords: Topic modeling; Literature; Big data; Search; 19th century

1. Introduction

Over the past five years, literary scholars have acquired access to increasingly large collections of digitized texts. While simple barriers such as physical access restricted research in the past, these barriers have begun to disappear in the digital age, and people now have broad access to previously difficult-to-access works. Consequently, they struggle with a new inflection

* Corresponding author.
E-mail addresses: tango@humnet.ucla.edu (T.R. Tangherlini), peter.leonard@yale.edu (P. Leonard).

0304-422X/$ – see front matter
http://dx.doi.org/10.1016/j.poetic.2013.08.002