NB! Preliminary manuscript, NOT to be quoted!
Submitted for inclusion in conference proceedings for "International Conference on Cultural research in the context of "Digital humanities"", Saint-Petersburg, Russia, 3–5 October 2013

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 Moretti'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

Introduction: digitizing the historian’s toolbox

There is a practical and methodological change underway in the historian’s craft in the form of ‘digital history’. (Weller 2013) This is not the first time that computer-based methods have been seen as having the potential to revolutionize historical studies. During the late 1960s and early 1970s the rise of ‘cliometrics’ and ‘quantitative history’ led within history to that a

> 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)

Despite the successes of this ‘cliometric revolution’ it never managed to revolutionize historical studies on the grand scale but instead added a valuable tool to the historian’s tool box. Whether the ongoing ‘digital history’ is going to be a revolution or not just yet another addition to the historian’s toolbox is too early to tell but it is nevertheless worth trying to sees its current status. To do this we in this paper are going to analyze the perhaps most central – and definitely the most topical – of the new methodological tools in the digital historian’s toolbox in the form of ‘topic modeling’.

Topic modeling is a prominent methodological example of literary historian Franco Moretti’s ‘distant reading’ approach to (literary) history which he has described as ‘, where “history will quickly become very different from what it is now: it will become ‘second hand’: a patchwork of other people’s research, *without a single direct textual reading*.“ (Moretti 2000: 57, emphasis in original, see also Moretti 2013) Distant reading rather than the data that can be gotten from ‘close reading’ of texts, depends on reading and analyzing aggregated ‘metadata’ of texts: titles, author names, publication years, affiliations and keywords. Another term for distant reading is ‘not-reading’ (Mueller 2007) with its connection to distant reading and metadata explained in the following way:

> 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)

This study give an overview of the history of topic modeling within digital humanities and survey its application within digital history as well as possible future methodological extensions. We will also analyze its uses in terms of various historical and methodological parameters: aims of investigations, what historical periods it has been applied to, languages, number of topics, kinds of texts, and kinds of publications.

Topic modeling as computer science: meaning, applications and potential

Topic modeling (TM) usually represents some form of a computer aided text processing tool that

> can be used to postulate complex latent structures responsible for a set of observations, making it possible to use statistical inference to recover this structure. This kind of approach is particularly useful with text, where the observed data (the words) are explicitly intended to communicate a latent structure (their meaning). (Griffiths & Steyvers 2004, 5228)

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.

The overwhelming benefit of TM is that it allows analysis of vastly larger quantities of data as compared to traditional approaches, allowing new ways of data mining. For example it would be practically impossible using traditional methods to summarize all publications of the journal *Science* 1990-1999 making up a corpus of 57 million words (Blei & Lafferty 2007). Therefore the structuring of textual data material, regardless of size probably represents TM's major advantage. Furthermore *TM can function as search tool* far superior to traditional single word searches (Mimno 2012). As TM potentially identifies themes within texts, it is possible to search for these within a corpus, turning it into a search function. And lastly, *TM can serve as quantitative check for intuition*. As TM identifies the most prominent 'themes' of a text it is possible to use it as indicators of which themes are (and maybe more interestingly which are not) addressed within a text. For example the rural development policy paper of the EU proclaimed itself to fundamentally break away from earlier policy efforts, by including quality of life aspects among others. However, only a few identified topics dealt with these new issues, compared to the traditional agricultural focus. So this 'break' appears to be primarily rhetorical (Brauer & Dymitrow 2013).

Today's topic modeling relies on the development of so called 'Latent Semantic Analysis' (LSA) within natural language processing and machine learning in the 1990s (Deerwester et al 1990). The version of topic modeling most commonly used by historians is 'latent Dirichlet allocation' (LDA) developed in the early 2000s by a group of researchers led by David Blei and presented 2003. The LDA algorithm works by first removing so called 'stop words' from the text, e.g. a, an, the, there, under, which etc. that only have relational meaning. This speeds up the processing and filtering for 'meaningful' topics. Then the algorithm assumes that each document represents a 'bag of words' where co-occurring words share some sort of meaning and based upon a statistical mean (e.g. Gibbs sampling) constructs topics. There are a myriad of different assumptions within LDA, but the three major assumptions (Blei 2012) are the following:

- the order of the words within an analyzed text is irrelevant
- the order of the documents from an analyzed corpus is irrelevant
- the number of topic is previously known

These are quite bold assumptions, however, it seems that even despite this LDA is able to identify meaningful topics (Mimno 2012). Another algorithm is the Correlated Topic Model (CTM) which is a further development of the LSA approach (Blei & Lafferty 2007) and that tries to address the issue of having to assume the number of topics prior to the analysis. CTM unlike LDA does not assume that topics are unrelated and tries to build 'correlations' between the individual topics (hence the name). CTM's advantage is that the number of topics does not have to be specified in advance, as these are a result of the correlation. The more technical side of TM research is constantly refining the algorithms involved (Asuncion et al. 2011; Baillie et al. 2011; Daud 2012; Huh & Feinberg 2012; Jianping et al. 2012).

The TM software used by the majority of researchers is the LDA-based MAchine Learning for LanguagE Toolkit (MALLET) developed by researchers at the University of Massachusetts-Amherst. MALLET works through a command line interface, making it a somewhat daunting for people just getting started with TM. MALLET requires two parameters to be defined before it can discover topics within a corpus: number of topics and the size of the document (*chunk*) within the corpus. (Jockers 2013, 133-34) However, there is currently no commonly agreed upon standard what these parameters should be. A 'rule of thumb' suggested by David Mimno is 100 topics with document chunks of 1000 words (Mimno in Jockers 2013, 134). However, this has to be adjusted to every individual analyzed corpus based upon the 'best fit' for the particular situation; therefore it represents an ongoing effort of improvement. There are also other LDA-implementations being developed such as the Paper Machines application (Johnson-Roberson 2012).

Additionally, another issue actively worked upon is finding the best possible way by researchers to interpret the meaning of the topics. Chang et al. (2009) discuses different statistical solutions to the problem involved in the interpretation of topics by humans; Jockers (2013) aid his interpretation by visualizing topics in a style akin to word clouds; while Heuser & Le-Khac (2012), among others efforts, identify the topic in combination with the original text by highlighting keywords on the document pages. Either way this represents an area of research, both in trying to identify and best visualize the topics (Blei 2012). The possible use of topic modeling as a search function is also a topical research effort. Mimno (2012) has developed a method where it is possible to identify topics within one corpus and search for them in another. Extending upon this idea, it becomes possible to use TM over several languages, using similar corpuses in different languages addressing the same issue (e.g. a Wikipedia article on the same term, Mimno et al. 2013). Other areas being explored is to expand TM from words to other representations such as images, sequencing of genes or scientific network structures (Li et. al. 2010; Chen et. al. 2012; Ding 2011). Last but not least are great efforts devoted to improving the TM user interface, making topic modeling more user friendly (Blei 2012).

Topic modeling as history: historians processing, modeling, and analyzing topics

This study of the emergence of topic modeling in historical studies take as its end points the first publication of a peer-reviewed journal article by an historian using topic modeling in 2006 and the 2013 publication of the first academic research monograph by an historian using topic modeling. The studies discussed here are those historical studies we have discovered from 2006 until and including 2012.

The first peer-reviewed academic article by an historian – an earlier historical study (Griffiths & Steyvers 2004) was written by two cognitive scientists – had the title "Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper" and were written by historian Sharon Block together with computer scientist David Newman and published in *Journal of the American Society for Information Science and Technology (JASIST)* in 2006. To give an example of and a feel for what historical topic modeling might entail we in the following give a more closer reading of this pioneering study than we in the following will give the other studies.

The article applied topic modeling to analyze the content of the major American colonial newspaper *Pennsylvania Gazette* between 1728-1800 consisting of roughly 80 000 texts in the form of articles and advertisements. TM was used exploratively to test if its application was feasible in order to structure the content of the newspaper. They discovered that most identified topics were trivial – representing common linguistically structures or attributes of particular aspects, or just noise – admitting that the interpretation was greatly helped by a historian familiar with the subject matter as the difference between what is 'trivial' and 'interesting' is sometimes not very easy to determine. By analyzing the types of advertisements over time, they could plot relative trends of over time. For example through the rise and subsequent demise of their CLOTH theme (including words like; 'silk', 'cotton', 'ditto', 'white', 'black', 'linen' etc.), they were able to strengthen a previous assumption that there was a rise and subsequent fall of the Pennsylvania cloth industry. Similar trends could be established for the expansion of government, religion and crime. Their conclusion was that TM could provide a quantitative measure for these initial more qualitative historical intuitions of the period. Their conclusion was that the main advantage was in the amount of documents that could be covered, as compared to more traditional methods and the possibility of using TM as a quantitative indicator of larger overall trends.

In their article Block & Newman also stated that there had "been a huge increase in the number of historical primary sources available online. Yet there has been little work done on processing, modeling, or analyzing these recently-available corpora." (Newman & Block 2006, 753) That was the situation then and that is still the situation as evidenced by literary historian Matthew Jockers who in 2013 in the first historical research monograph using topic modeling *Macroanalysis: Digital methods and literary history* (University of Illinois Press, 2013) laments the lack of scholarly work in digital humanities:

> To be sure, literary scholars have taken advantage of digitized textual material, but this use has been primarily in the arena of search, retrieval, and access. We have not yet seen the scaling of our scholarly questions in accordance with the massive scaling of digital content that is now held in twenty-first-century digital libraries. In this Google Books era, we can take for granted that some digital version of the text we need will be available somewhere online, but we have not yet fully articulated or explored the ways in which these massive corpora offer new avenues for research and new ways of thinking about our literary subject. (Jockers 2013, 16-17)

Our study confirms this view in regards to topic modeling as we during 2006-2012 found only some twenty historical studies using topic modeling of which the overwhelming majority either stayed at sketching possible uses or explored the method rather than used it to answer specific historical questions.

Distant reading of historical topic modeling

The texts using historical topic modeling included in this study could appear somewhat unreliable to traditional historians as they go beyond the standard academic texts. Following Toni Weller's observation that "the traditional forms of publication in history are not suited to the fast-changing discourses of the digital age – demonstrated by the fact that most pure digital history texts tend to be in the form of websites, blogs and online articles and journals rather than the traditional historical outlet of the monograph" (Weller 2013, 4) we have also included these kind of texts as well as conference proceedings if these texts contains historical studies of topic modeling.

	Publica	Author Information					Corpus			
ID	Year	Name	Based in	Discipline	Acad. seniority	Publ. type	Chronology	Location	Language	Type
1	2006	Sharon Block	USA	History	Assoc. Prof	journal article	1728 - 1800	USA	English	newspaper, magazine
		David Newman	USA	CS	PhD					
2	2006	Sharon Block	USA	History	Assoc. Prof	journal article	1728 - 1800	USA	English	newspaper, magazine
3	2007	David M. Blei	USA	CS	Assoc. Prof	journal article	1900 - 1999	global	English	scientific articles
		John D. Laffety	USA	CS	Professor					
4	2008	David Hall	USA	CS	Student	B. A. thesis	1978 - 2006	USA	English	scientific articles
5	2008	David Hall	USA	CS	PhD Student	conf. paper	1978 - 2006	USA	English	scientific articles
		Daniel Jurafsky	USA	Linguistics CS	Professor					
		Christopher Manning	USA	Linguistics CS	Professor					
6	2009	Qi He	USA	CS	PhD Student	conf. paper	1993 - 2008	USA	English	scientific articles
		Bi Chen	USA	IS	Student					
		Jian Pei	Canada	CS	Professor					
		Baojun Qiu	USA	CS	Post Doc					
		Prasenjit Mitra	USA	IS	Assoc. Prof.					
		C. Lee Giles	USA	IS	Professor					
7	2010	Cameron Blevis	USA	History	PhD Student	blogg, web.	1785 - 1812	USA	English	diaries, letters
8	2011	Lincoln Mullen	USA	History	PhD Student	conf. paper	1700 - 1735	USA	English	diaries, letters
9	2011	Robert K. Nelson	USA	History	PhD	newspaper	1860 - 1865	USA	English	newspaper, magazine
10	2011	Peter Wittek	Sweden	CS Physics	Assoc. Prof	conf. paper	1584 - 1650	Netherlan	Mutiple Languages	diaries, letter
		Walter Ravenek	Netherlan	Chemistry IS	PhD					
11	2011	Tze-I Yang	USA	CS	n/a	conf. paper	1836 - 2008	USA	English	newspaper, magazine
		Andrew J. Torget	USA	History	Post Doc					
		Rada Mihalcea	USA	CS	n/a					
12	2011	Sharon Block	USA	History	Assoc. Prof	journal article	1985 - 2005	USA	English	scientific articles
		David Newman	USA	CS	PhD					
13	2011	Robert K. Nelson	USA	History	PhD	blogg, web.	1860 - 1865	USA	English	newspaper, magazine
14	2011	Thomas C. Templeton	USA		n/a	conf. paper	1860 - 1865	USA	English	newspaper, magazine
		Travis Brown	USA	Language	M.A.					
		Sayan Battacharyya	USA	IS Engineering	M.A. PhD					
		Jordan Boyd-Graber	USA	IS CS Linguistics	Ass. Prof.					
15	2012	Ashton Anderson	USA	CS	PhD Student	conf. paper	1980 - 2008	USA	English	scientific articles
		Dan McFarland	USA	Sociology	Assoc. Prof.					
		Daniel Jurafsky	USA	Linguistics CS	Professor					
16	2012	Sophie Kushkuley	USA	Linguistics Mathematics	Ass. Prof.	conf. paper	1867 - 1899	USA	English	newspaper, magazine
17	2012	Matthew L. Jockers	USA	Language	Post Doc	faculty paper	1800 - 1899	USA	English	novels
		David Mimmo	USA	CS	Post Doc			UK		
18	2012	Ryan Heuser	USA		n/a	faculty paper	1790 - 1900	UK	English	novels
		Long Le-Khac	USA		PhD Student					
19	2012	David Mimmo	USA	CS	Post Doc	journal article	1911 - 2004	global	Mutiple Languages	scientific articles
20	2012	Andrew Piper	Canada	Literature Language	Assoc. Prof	conf. paper	1774/1787	Germany	German	novels
		Mark Algee-Hewitt	Canada	Literature	Post Doc					
21	2012	Matt Erlin	USA	Languages Literature	Professor	blogg, web.	1731 - 1864	Germany	German	novels
22	2012	Allen B. Riddell	USA	Literature	PhD Student	conf. paper	1928 - 2006	Germany	German	scientific articles
23	2012	Ching-man Au Yeung	China	CS	Post Doc	conf. paper	1990 - 2010	global	Mutiple Languages	newspaper, magazine
		Adam Jatowt	Japan	IS	Assoc. Prof					

Table 1. Texts with historical studies 2006-2012 using topic modeling. Full references can be found in the bibliography. CS stands for Computer Science and IS for Information Science.

The texts were found by first mining the by now canonical texts of historical studies of topic modeling literature – e.g. Newman & Block 2006, Block 2006, Blevins 2010, Mimno 2012, Nelson 2011 that were all referenced in texts using historical topic modeling – for authors, articles, references and citations connected to these studies. This was followed by searching through Google, Google Scholar and Google Books with keywords such as 'topic models', 'topic modeling' in combination with 'history' and 'historical' and then the authors, articles, references and citations that were connected to the studies found through this were followed up to find additional texts. We limited ourselves to texts in English.

The studies found were then skimmed through to discern whether they were actually using topic modeling in any major way leading to studies only mentioning topic modeling in passing to be sorted out.

The result was 23 texts using topic modeling and shown in Table 1 as well as included in the bibliography marked with a star (*). These texts were analyzed in a distant or not-so-close reading fashion in that we were primarily not analyzing the details of the topic modeling in the studies but rather more larger patterns regarding topic modeling's users and use. Although we intend to devote a extended study do a close(r) reading of the use of topic models in historical studies we can already now state that the majority of historical studies are primarily exploratory or prospective in that they are focused on developing, testing or assessing TM as a historical method rather than actually using it to solve an independent historical problem, much in line with Jockers' lament discussed above.

The texts' authors and corpora were characterized according to several parameters: authors' academic background, gender, rank and country of academic institution; corpora's type, language, chronology, and geographical focus of the analyzed corpus. In the following we provide a presentation of our results both in the form of summarizing discussions of the results and in the form of diagrams that are also discussed. Like most studies using topic models many of the results are not unsurprising to those that have been following the development of the field.

Two such unsurprising facts about the texts' *authors* are that the overwhelming majority of authors are men - with only two authors with recognizable female names; that an overwhelming majority (92%) are located at American (US and Canadian) academic institutions and the others are solitary researchers located in the Netherlands, Sweden, Japan and China. Somewhat perhaps more surprising is that judging from authors' academic seniority this appears to be a young man's – indeed – game with at least 56% untenured junior researchers of which almost a quarter undergraduate or graduate students and about a third of the authors being full or associate professors. (Fig. 1)

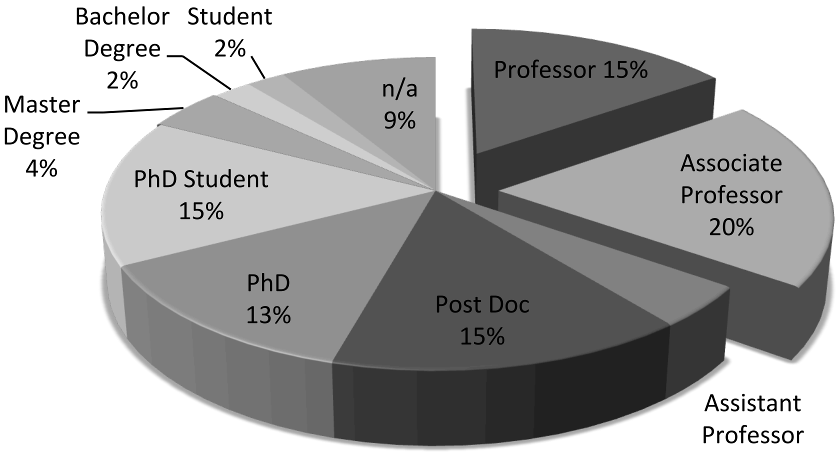

Fig. 1. Academic Background of Authors in Table 1

Furthermore another interesting finding is that it appears that this is a field that is very much still being technology driven in that only 30% of the authors have a disciplinary grounding in the humanities (history, literature and languages) and almost 60% belong to the technical and natural sciences (Fig. 2). This does not count the 9% of authors from linguistics who could be from either its humanist or technical side although it is the impression that most could be firmly placed in the technical camp. Finally one interesting finding is that such a relatively large part (13%) of the texts using topic modeling are non-standard academic publications such as blogs and websites.

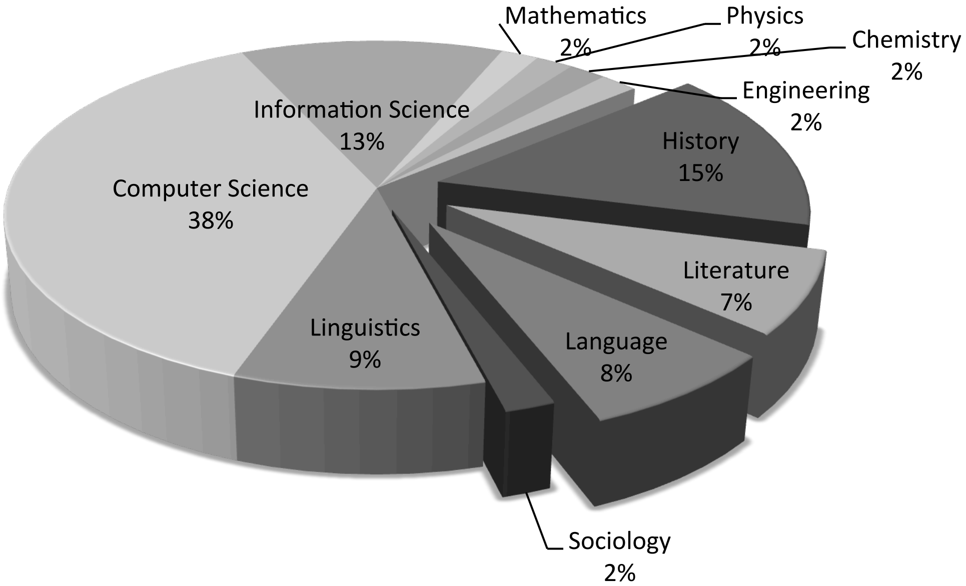

Fig. 2: Academic Background of Authors in Table 1

When it comes to the texts' *corpora* also here there are some expected results. The first is, as said above, the majority use LDA in its MALLET implementation and that the majority of corpora (74%) are in English followed by German (13%) and – perhaps more surprisingly –

13% in multiple languages. In line with this the geographical areas the corpora refers to are primarily the USA (62%) but interesting is that a substantial part (12%) are global in coverage. Each corpus' chronological span varies between 2-134 years but most (55%) are 2-30 years. One of the most interesting findings is that it is so contemporary focused. The different corpora cover texts between 1564-2010 (Fig. 3) with a focus on the near present with almost a third starting after 1977. This is also reflected in what kind of corpus that is studied with the two largest parts (70%) being scientific articles and newspapers and more traditional historical material such as novels and handwritten texts making up the minority.

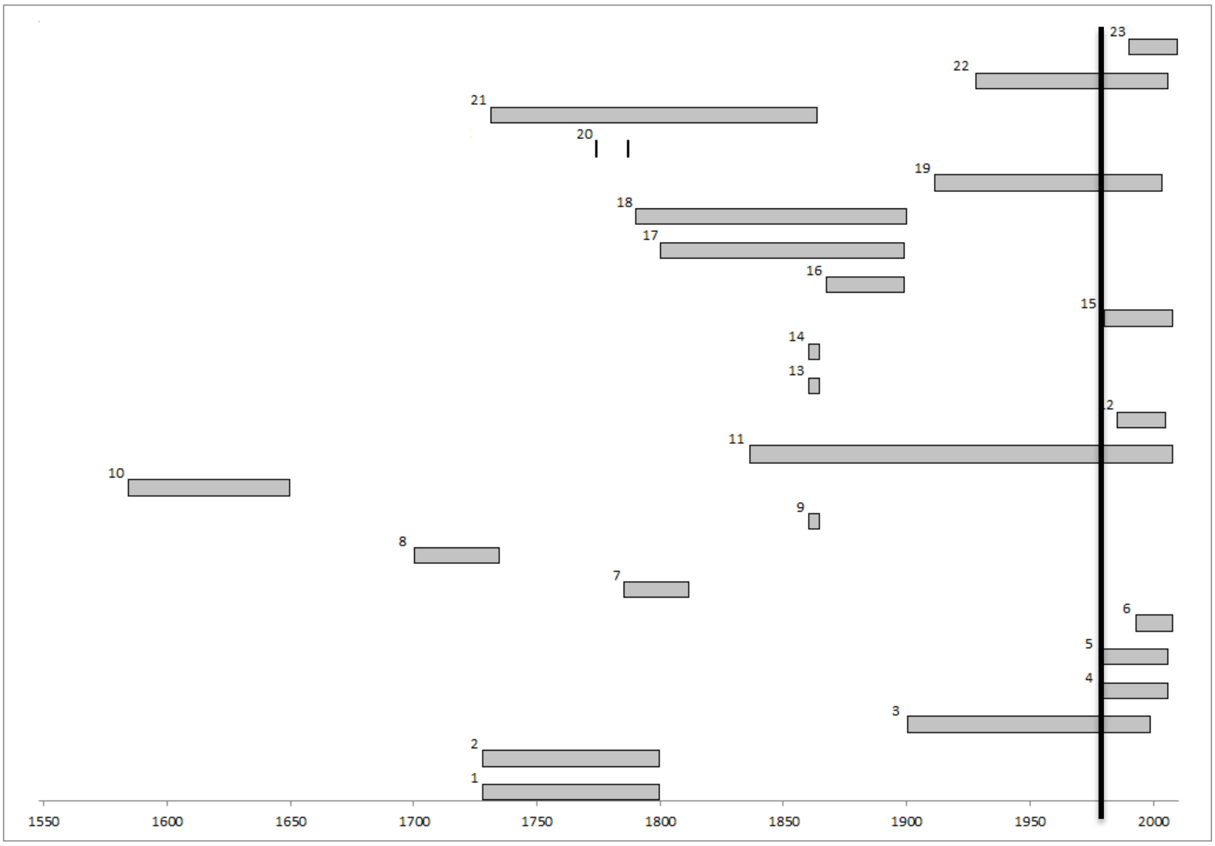

Fig. 3: Temporal coverage of corpora where the numbers refer to their ID in Table 1.

Concluding discussion: retooling history?

This have consisted in a first attempt towards systematically assessing the state of the art of the use of topic modeling within the prognosticated digital revolution of historical studies. The study has applied a distant reading approach towards a corpus of 23 texts consisting of historical studies of topic modeling 2006-2012. Although saving a closer reading of the use of topic modeling in the corpus for a future study what the study have shown is that very few in-depth and exhaustive historical studies of topic modeling can be found. The method is currently very much an emergent method in its infancy.

From a methodological point of view topic modeling has reached some stability in that there are primarily one method (LDA) and one implementation (MALLET) that is used by the majority of users. However, despite this there are a large interest both among computer scientists and historians in developing new variants and applications (such as Paper Machines) for topic modeling. TM also shows great potential in becoming used as search function and indexing method. Probably the best current application of TM is its application to quantitative check for intuitions. That most – although not all - of the work done upon developing TM is conducted by people from the computing disciplines are not surprising, what might be more surprising is that they are also in the majority developing it for historical studies.

When it comes to the historical survey many results are rather unsurprising and expected such as the US and English dominance. What is less expected is the dominance of technology and of junior researchers. Historical studies using topic modeling is in many ways following the model from natural sciences in that it is so far a young men's and computer scientist's game rather than the established historian's. This relative lack of experienced humanists might probably to a large degree explain why so many of the studies are focused on the near present and on method development. Contrary to the natural and technical sciences, in humanities new critical perspectives and questions are generally considered to be the fruits of experienced scholars. Perhaps what topic modeling is lacking more than more sophisticated models is the experience to ask the new unexpected questions.

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.

Bibliography

*__Anderson, A., McFarland, D. & Jurafsky, D.__ (2012). Towards a Computational History of the ACL: 1980-2008, *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries,* 13-21.

*__Andrew, P. & Algee-Hewitt, M.__ (2012). The Werther Effect I: Goethe Topologically. In *Distant Readings/Descriptive Turns: Topologies of German Culture in the Long Nineteenth Century*. Matt Erlin & Lynn Tatlock, ed. Rochester, NY: Camden House, forthcoming, 24 pp.

Asuncion, A., Smyth, P., & Welling, M. (2011). Asynchronous distributed estimation of topic models for document analysis, *Statistical Methodology*, 8:1, 3-17.

Baillie, M., Carman, M. & Crestani, F. (2011). A multi-collection latent topic model for federated search, *Information Retrieval*, 14:4, 390-412.

Blei, D.M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77-84.

*__Blei, D. M., & Lafferty, J.D.__ (2007). A correlated topic model of *Science*, *Annals of Applied Statistics*, 1:1, 17-35.

Blei, D.M. & Lafferty, J.D. (2007). A correlated topic model of *Science*, *Annals of Applied Statistics*, 1:2, 634-642.

*__Blevins, D.__ (2010). Topic Modeling Martha Ballard's Diary, *historying: thoughts on scholarship and history in a digital age,* blog available at: historying.org/ 2010/04/01/topic-modeling-martha-ballards-diary/ accessed 2013-07-29.

*__Block, S. & Newman, D.J.__ (2011). What, Where, When and Sometimes Why: Data Mining Twenty Years of Women's History Abstracts, *Journal of Women's History*, 23:1, 81-109.

*__Block, S__. (2006). Doing More with Digitization: An introduction to topic modeling of early American sources, *Common-place*: *The Interactive Journal of Early American Life* 6:2, www.common-place.org/vol-06/no-02/tales/, accessed 2013-07-29.

Brauer, R. & Dymitrow, M. (2013). Digitally modeling regional development in Europe: a new methodological approach to policy analysis, *Man-City-Nature* [forthcoming].

Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S. & Blei, D. M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models, available at: www.umiacs.umd.edu/~jbg/docs/ nips2009-rtl.pdf, accessed 2013-07-23.

Chen, X., Hu, X., Lim, T.Y., Shen, X., Park, E.K., & Rosen, G.L. (2012). Exploiting the functional and taxonomic structure of genomic data by probabilistic topic modeling, *IEEE/ACM transactions on computational biology and bioinformatics*, 9:4, 980-994.

Daud, A. (2012). Using time topic modeling for semantics-based dynamic research interest finding, *Knowledge-Based Systems*, 26, 154-163.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. & Harshman, R. (1990). Indexing by Latent Semantic Analysis, *Journal of the American Society for Information Science and Technology* 41:6, 391-407.

Ding, Y. (2011). Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks, *Journal of Informetrics*, 5:1, 187–203.

*__Erlin, M. (__2012). Rethinking the History of German Literature 1731-1864: A Statistical Approach, available at: hdw.artsci.wustl.edu/node/42, accessed 2013-08-15.

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.

*__Hall, D., Jurafsky, D., Manning, C. D. (__2008). Studying the History of Ideas Using Topic Models, *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 363–371.

*__Hall, D.L.W.__ (2008). *Tracking the Evolution of Science*, BSc Honors Thesis, Stanford University, 58 pp.

*__He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., Giles C.L. (__2009). Detecting topic evolution in scientific literature: how can citations help?, *The ACM Conference on Information and Knowledge Management,* 957-966.

*__Heuser, R., Le-Khac, L. (__2012). A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method, *Pamphlets of the Stanford Literary Lab 4*, 66 pp.

Huh, S., Feinberg, S.E. (2012). Discriminative Topic Modeling Based on Manifold Learning, *Transactions on Knowledge Discovery from Data*, 5:4, 1-25.

Jianping, Z., Jiangjiao, D., Wenjun, C., & Chengrong, W. (2012). Topics modeling based on selective Zipf distribution, *Expert Systems With Applications*, 39:7, 6541-6546.

Jockers, M.L. (2013). *Macroanalysis: Digital methods and literary history*. University of Illinois Press.

***Jockers, M.L., Mimno, D. (**2012). Significant Themes in 19th-Century Literature, *Faculty Publications – Department of English*. Paper 105. 24 pp.

Johnson-Roberson, C. (2012). Paper machines, *metaLAB (at) Harvard Blog*, available at metalab.harvard.edu/2012/07/paper-machines/, accessed 2012-07-12.

***Kushkuley, S. (**2012). Trend Analysis in Harper's Bazaar, *Workshop on Computational Linguistics for Literature*, 84–87.

Li, L.-J., Wang, C., Lim, Y., Blei, D., & Fei-Fei, L. (2010). Building and using a *semantivisual* image hierarchy. In *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3336 – 3343.

***Mimno, D. (**2012). Computational historiography: Data mining in a century of classics journals, *Journal on Computing and Cultural Heritage*, 5:1, 1-19.

Mimno, D., Wallach, H.M., Naradowsky, J., Smith, D.A., & McCallum, A. (2009). Polylingual Topic Models, *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 880-889.

Moretti, F. (2000), Conjectures on World Literature, *New Left Review* N.S., 1:1, 54-68.

Moretti, F. (2013), *Distant reading*, London: Verso.

Mueller, M. (2007), "Notes towards a user manual of Monk", *The MONK Project Wiki,* apps.lis.illinois.edu/wiki/display/MONK/Notes+towards+a+user+manual+of+Monk, May 06, 2007.

***Mullen, L. (**2011). The Vocabulary of Conversation: Text-Mining the East Windsor Conversation Relations, Paper presented at the annual meeting of the American Society for Church History, Boston, January 9, 2011, available at: lincolnmullen.com/downloads/ docs/Mullen.Vocabulary-of-Conversion.pdf, accessed 2013-07-29.

***Nelson R. K. (**2011). Of Monsters, Men — And Topic Modeling, *New York Times Disunion Blog*, available at: opinionator.blogs.nytimes.com/2011/05/29/of-monsters-men-and-topic-modeling/?_r=0 accessed 2013-07-29.

***Nelson, R.K. (**2011). Website *Mining the* Dispatch, available at: dsl.richmond.edu/ dispatch/pages/home accessed 2013-07-29.

***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.

***Riddell, A.B.** (2012). How to Read 22,198 Journal Articles: Studying the History of German Studies with Topic Models, Paper presented at USP Seminar, "Big Data and the Humanities," Tuesday, Oct. 9, University Scholars Program, Duke University, available at: ariddell.org/static/how-to-read-n-articles.pdf, accessed 2013-07-29, 24 pp.

***Templeton, T.C., Brown, T., Battacharyya, S., & Boyd-Graber, J. (**2012). Mining the Dispatch under Supervision: Using Casualty Counts to Guide Topics from the Richmond Daily Dispatch Corpus, *Chicago Colloquium on Digital Humanities and Computer Science*, available at: www.umiacs.umd.edu/~jbg/docs/ slda_civil_war.pdf accessed 2013-07-29.

Thomas, II, W.G. (2004). Computing and the historical imagination, In *A Companion to Digital Humanities*, Susan Schreibman, Ray Siemens & John Unsworth, eds., Oxford: Blackwell, 56-68.

Weller, T. (2013). Introduction: history in the digital age, In *History in the Digital Age*, Toni Weller, ed., London: Routledge, 1-19.

***Wittek, P., & Ravenek, W. (**2011). Supporting the Exploration of a Corpus of 17th-Century Scholarly Correspondences by Topic Modeling, *SDH 2011 Supporting Digital Humanities:*

Answering the unaskable, available at: www.clarin.nl/sites/default/files/sdh2011-wittek-ravenek.pdf, accessed 2013-08-14, 6 pp.

***Yang, T., Torget, A.J., Mihalcea, R. (**2011). Topic Modeling on Historical Newspapers, *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. The Association for Computational Linguistics*, 96–104

***Yeung, C.A., Jatowt, A. (**2012). Studying How the Past is Remembered: Towards Computational History through Large Scale Text Mining, CIKM’11, available at: www.dl.kuis.kyoto-u.ac.jp/~adam/cikm11a.pdf, accessed 2013-07-29.