SEMI-SUPERVISED TEXTUAL ANALYSIS AND HISTORICAL RESEARCH HELPING EACH OTHER: SOME THOUGHTS AND OBSERVATIONS

FEDERICO NANNI, HIRAM KUMPER AND SIMONE PAOLO PONZETTO

Abstract *Future historians will describe the rise of the World Wide Web as the turning point of their academic profession. As a matter of fact, thanks to an unprecedented amount of digitization projects and to the preservation of born-digital sources, for the first time they have at their disposal a gigantic collection of traces of our past. However, to understand trends and obtain useful insights from these very large amounts of data, historians will need more and more fine-grained techniques. This will be especially true if their objective will turn to hypothesis-testing studies, in order to build arguments by employing their deep in-domain expertise.*

For this reason, we focus our paper on a set of computational techniques, namely semi-supervised computational methods, which could potentially provide us with a methodological turning point for this change. As a matter of fact these approaches, due to their potential of affirming themselves as both knowledge and data driven at the same time, could become a solid alternative to some of the today most employed unsupervised techniques.

However, historians who intend to employ them as evidences for supporting a claim, have to use computational methods not anymore as black boxes but as a series of well known methodological approaches. For this reason, we believe that if developing computational skills will be important for them, a solid background knowledge on the most important data analysis and results evaluation procedures will become far more capital.

International Journal of Humanities and Arts Computing 10.1 (2016): 63–77
DOI: 10.3366/ijhac.2016.0160

www.euppublishing.com/journal/ijhac

Keywords: semi-supervised methods; historical studies; data analysis; born-digital archives

1. INTRODUCTION

In December 2010 Google presented a service called 'Google Ngram Viewer'.[1] This tool allows us to look at the occurrence of single words or sentences in specific subsets of the immense corpus digitized by the Google Books project.

A few weeks after, Erez Lieberman Aiden and Jean-Baptiste Michel, team leaders of the prototype Viewer, offered a demonstration of the tool at the annual meeting of the American Historical Association in Boston.[2] In front of around 25 curious historians, they noted the enormous potential of conducting historical research by extracting information from large corpora. In particular, they revealed a way to deal with one of the biggest issues for historians who are exploring large datasets, namely rapidly detecting the distribution of specific words in the corpus.[3]

Interestingly, the development and the functionalities of this tool demonstrate some of the most relevant characteristics of the current interactions between the practice of historical research and the use of computational methods:

> Firstly, no historian has been directly involved in any step of the development of this project.[4] This is particularly significant, given that they would likely be the primary targets of a tool able to process information from a corpus spanning five hundred years. As Aiden and Michel remarked, this is due to two well-known reasons: historians traditionally do not have solid computational skills and they are usually skeptical about the development of quantitative approaches for the analysis of sources.[5]
>
> Secondly, others have noted that the Ngram Viewer offers an over-simplified research tool, which usually leads to general coarse-grained explorative analyses and to few simple historical discoveries.[6]
>
> Finally, the way in which the Ngram Viewer has been presented and identified outside academia as a representative tool of the digital humanities also reveals the growing enthusiasm for methodology studies and big-data driven researches in this community.[7] However, as already remarked, researchers in digital humanities need to bear in mind their long-term purpose, that is to use the computer in order to answer specific and relevant research questions and not simply to build tools.[8]

But while the Ngram Viewer symbolises a current widespread way of employing computational methods for studying historical corpora, namely for data-exploration and general hypothesis-confirmation analyses, we believe that a change is about to come. In fact, in our opinion new generations of historians

will need more and more fine-grained techniques to conduct inspections of large datasets. This will be especially true if their objective turns from exploratory analyses to hypothesis-testing studies, in order to build arguments by employing their deep in-domain expertise.

For this reason, we focus here on a set of computational techniques, namely semi-supervised computational methods, which could potentially provide us with a methodological turning point for this change.[9] As a matter of fact these approaches make it possible to actively include the human expert in the computational process. Therefore, due to their potential of affirming themselves as both *knowledge-* and *data-driven* at the same time, they could become a solid alternative to some of the most common unsupervised techniques currently used.

However, historians who intend to employ computational methods as evidence for supporting a claim, have to use them as a series of well known methodological approaches rather than as 'black boxes' whose workings are unknown.[10]

For this reason, if developing computational skills will be important for historians, a solid background knowledge of the most important data analysis and results evaluation procedures will become far more necessary.

Starting from all these assumptions, this paper is organized as follows: firstly, a few basic concepts of machine learning methods are introduced. Then, a diachronic description of the use of computational methods in historical research is presented. Following this, our focus on a specific technique, namely Latent Dirichlet Allocation topic modeling, is defined. Next, the advantages and the consequences of the use of semi-supervised topic modeling approaches on the historian's craft are described. Finally, a future project on the use of these methodological frameworks for the analysis of the different semantic dimensions of specific concepts in a collection of around 1,000 French legal books from the 17th and 18th century is introduced.

Our essay is focused on a precise potentiality of the complex datasets of sources that historians have now at their disposal. This is the possibility of exploiting the results of fine-grained analyses as historical evidence through the combination of specific in-domain research interests and the scientifically correct employment of computational methods. This will help researchers to deal with the abundance of digital materials by extracting precise information from them, and to move from exploratory studies to hypothesis testing analyses. However, now that both large datasets and text mining methods are at our disposal, other challenges are emerging, such as multilingual corpora or the evolution of languages in diachronic extended datasets. In the near future this will raise other issues for the new generations of historians, increasing the need for advanced computational approaches (i.e. specific language models for machine translation) and demanding always more advanced competencies of the humanities researcher.

2. SUPERVISED AND UNSUPERVISED TEXT ANALYSES

Before going into the details of how these methods have previously been employed in historical research and how they could be used in the near future, it is important to clarify a few key concepts in data analysis and machine learning that have already been mentioned in the previous paragraphs.[11] As described earlier, an initial requirement of many historical studies is to identify semantic similarities and recurrent lexical patterns in a collection of documents. In machine learning there are two main different kinds of approaches that allow us to do this.

The first one consists of *supervised* learning methods, which focus on classification tasks. In classification tasks, humans identify a specific property of a subset of elements in the dataset (for example articles about foreign policy in a newspaper archive) and then guide the computer, by means of an algorithm, to learn how to find other elements with that characteristic. This is done by providing the machine with a dataset of labeled examples ('this is an article about foreign policy', 'this is not'), called a 'gold standard', which are described by a set of other 'features' (for instance, the frequency of each word in each document). Moreover, the learning process is typically divided into two main phases, namely: i) a training phase, in which the predictive model is learnt from the labeled data; ii) a testing phase, in which the previously learnt model is applied to unseen, unlabeled data in order to quantity its predictive power, specifically its ability to generalize to data other than the labeled ones seen during training. Additionally, a validation phase can take place to fine-tune the model's parameters for the specific task or domain at hand – e.g., classifying foreign policy articles from newspaper sources, as opposed to websites.

The potential of a good classifier is immense, in that it offers a model that generalizes from labeled to (a potentially very large set of) unlabeled data. However, building such models can also be extremely time-consuming. In fact, researchers not only need a dataset with specific annotated examples to train the classifier but, perhaps even more fundamentally, they need to have extremely clear sense of what they are looking for, since this leads them to define the annotation guidelines and learning task itself. For this reason, it is evident that classification methods are arguably not the most convenient approaches for conducting data exploration in those situations where a researcher sets out to investigate the dataset with no clear goal in mind other than searching for any phenomenon they deem interesting *a posteriori*.

The second class of methods is *unsupervised*, and addresses the problem of clustering. In a nutshell, clustering methods aim at grouping elements from a dataset on the basis of their similarity, as computed from their set of features (for example by looking at patterns in the frequency of words in different documents). This is achieved by computing likenesses across features without

relying on labeled examples, unsupervised by humans. Crucially for digital humanities scholars, researchers can study the resulting clusters in order to understand what the (latent) semantic meaning of the similarities between the elements is.

Clustering techniques are extremely useful for analyzing large corpora of unlabeled data (i.e., consisting of 'just text'), since they rapidly offer researchers a tool to get a first idea of their content in a structured way (i.e., as clusters of similar elements, which can be optionally hierarchically arranged by using so-called hierarchical clustering methods). This is primarily because, as they do not require labeled data, they can be applied without having in mind a specific phenomenon or characteristics of the dataset to mine (i.e., learn). However, even if scholars noted their potential, for example by creating serendipity, and different metrics have been proposed for evaluating the number and correctness of these clusters, this is still an extremely challenging task, typically due to the difficulties of interpreting the clusters output by the algorithms.[12]

3. STUDYING THE PAST, IN THE DIGITAL WORLD

The potential of computational methods for the study of primary sources has been a recurrent topic in the humanities. As Thomas remarked, already in 1945 Vannevar Bush, in his famous essay 'As We May Think', pointed out that technology could be the solution that will enable us to manage the abundance of scientific and humanistic data; in his vision the Memex could become an extremely useful instrument for historians.[13] The use of the computer in historical researches consolidated between the Sixties and the Seventies with its application to the analysis of economic and census data. The advent of cliometrics gave birth to a long discussion on the use of the results of quantitative analysis as evidence in the study of the past.[14]

Due in part to this long debate on the application of quantitative methods in historical research and in part to the new potentials of the Web as a platform for the collection, presentation, and dissemination of material, during the Nineties a different research focus emerged in what was already at that time identified as digital history.[15] As Robertson recently pointed out, this specific attention on the more 'communicative aspects' of doing research in the humanities could be recognized as one of the main differences between the ways in which historians have been interpreting the digital turn compared to their colleagues in literary studies over the last twenty years.[16]

However, regardless of whether historians of the 21st Century are interested in employing computational methods for analysis textual documents or not, it is evident that the never-ending increase of digitized and born digital sources is no longer manageable with traditional close reading hermeneutic approaches alone.[17] For this reason, two different activities have consolidated in the digital

humanities community during the last decade. On one side digital historians started creating tools in order to help other traditionally trained colleagues in employing computational methods.[18] On the other side, more recently a small but strongly connected community of historians has decided to focus their efforts on teaching the basic of programming languages and the potential of different textual analyses techniques for conducting exploratory studies of their datasets. As Turkel remarked: 'My priority is to help train a generation of programming historians. I acknowledge the wonderful work that my colleagues are doing by presenting history on the Web and by building digital tools for people who can't build their own. I know that the investment of time and energy that programming requires will make sense only for one historian in a hundred'.[19]

a. Computational History

The works conducted by Willam J. Turkel at the University of Western Ontario, with particular attention to his blog 'Digital History Hacks' and his project 'The programming historian', could be identified as a starting point of these digital interactions.[20] Following Turkel's approaches and advice, a group of historians has begun experimenting with these different computational methods to explore large historical corpora.[21] The use of Natural Language Processing and Information Retrieval methods, combined with network analysis techniques and a solid set of visualization tools, are the points around which this new wave of quantitative methods in historiography has consolidated.

During recent years several interesting examples of these interactions between historical research and computational approaches have been presented.[22] In addition, thanks to the collaborations with other digital humanities colleagues (i.e. literary studies researchers and digital archivists), the words 'text mining' and 'distant reading' have become buzzwords of this new trend in digital history. If we were to look more closely at how these techniques have been applied, we could notice that the first objective of the digital humanities researchers has been to show the exploratory potential of these methods and to confirm their accuracy by re-evaluating already well-known historical facts.[23] As we will remark in the next sections, this is due to the unsupervised nature of the specific textual analysis techniques most widely used in historical research (e.g., topic modeling), which do not need (but at the same time cannot obtain benefit from) human supervision and in-domain knowledge during the computational process.

b. Topic modeling

Topic modeling is arguably the most popular text mining technique in digital humanities.[24] Its success is due to its ability to address one of the deepest need of a historian, namely to automatically identify with as little human supervision

as possible (none, ideally) a list of topics in a collections of documents, and how these are intertwined with specific document sources in the collection. At a first sight this technique seems to be the methodological future of historical research.

However, as researchers rapidly discovered, working with topic modeling toolboxes is neither easy nor always yielding satisfactory results. First of all, Latent Dirichlet Allocation (LDA - the main topic models algorithm), like other unsupervised techniques, needs to be told in advance the number of topics (resp. clusters) that the researcher is interested in.[25] However, knowing the number of topics is itself a non-trivial issue, which leads researchers to a chicken-and-egg-problem in which they use LDA to find *some* interesting topics, while being required to explicitly state the *exact number* of such topics they are after. Moreover, as this technique looks at the distribution of topics by document, the results will be extremely different in relation to the number of topics chosen.

Thus, topic modeling highlights both advantages and limitations of unsupervised techniques. In fact, the obtained topics are, as others noticed, usually difficult to decode; each of them is presented as a list of words, and being able to identify it with a specific concept generally depends on the intuitions of the researcher.[26]

The first paper on LDA was published in 2003, however before 2010 there were just a few publications on humanities topics where this technique was employed.[27] We could identify a turning point in the digital humanities community between 2011 and 2012, when suddenly a remarkable number of blogposts, online discussions, workshops and then publications been focused on how to deal and employ this technique.[28] As we will describe later, in the same period Owens observed the risks for humanists of using topic modeling results as justification for a theory and in general suggested limiting its use to exploratory studies.[29]

4. SEMI-SUPERVISED TEXTUAL ANALYSIS

Today, if there is something more criticized than the use of quantitative methods in the humanities, this is data-driven research.[30] More specifically, we agree that the practice of employing unsupervised computational approaches to analyse a dataset and then relying on their automatically generated results to build a scholarly argument could reduce the role of the humanist in the research process. This is due to two main reasons: firstly, since even the more technically skilled historian does not have a solid statistical background as computational linguists, computer scientists or other kinds of researchers that currently are implementing these methods; this will consequently limit their understanding of both the techniques and the obtained results.[31] Secondly, because by employing unsupervised techniques, historians will not draw on their background knowledge, and will not directly use these methods for

answering specific research questions they have in mind. This is because, since unsupervised methods do not rely on human supervision and are mainly targeted at generating serendipity, they do not, and are not meant to include human feedback to guide the process of model creation.

However, on the other side of the spectrum, supervised classification approaches are particularly time-consuming to build, and their usefulness depends on specific research purposes (i.e., what is the scholar trying to discover by classifying documents in different categories?). Therefore, it is evident that for historians interested in performing more fine-grained explorations, a different computational technique is needed that is able to stake out a middle ground between explicit human supervision and serendipitous searching and exploration; a method that could help the researcher switching from general exploration analyses to more specific ones, from getting a first idea of the contents of a corpus to start evaluating theories by employing her/his domain expertise.

For this purpose, we argue that a series of semi-supervised topic modeling algorithms, adopted in recent years in the fields of machine learning and natural language processing, could also become established research methods in digital history.

The first one is Supervised LDA, originally presented by Mcauliffe and Blei.[32] This method makes it possible to derive distribution of topics by considering a set of labels, each one associated with each document. In their paper the authors note the potential of this method when the prediction of a specific value is the ultimate goal; to this end, they combine movie ratings and text reviews to predict the score of unrated reviews. However, as remarked by Travis Brown, historians could also experiment with this technique, to, for example identify the relation between topics and labels (i.e. to find the most relevant topics for 'economics' articles).[33]

A conceptual extension of this technique is Labeled LDA, developed by Ramage et al.[34] This method makes it possible to highlight the distribution of labeled topics in a set of multi-labeled documents. If we imagine a corpus where every document is described by a set of meta-tags (for example a newspaper archive with articles associated with both 'economics', 'foreign policy', and so on), Labeled LDA will identify the relation between topics, documents and tags, and its output will consist of a list of topics, one for each tag. This, in turn, could be used to identify which part of each document is associated with each tag.

Another relevant approach is Dirichlet-multinomial regression, proposed by Mimno and McCallum.[35] As the authors describe, rather than generating metadata (as for example the ratings in Supervised LDA) or estimating topical densities for metadata elements (as the topics related to metadata, like Labeled LDA), this method learns topic assignments by considering a set of pre-assigned document-features. In their paper the researchers show how authors,

paper-citations and date of publications could be useful features of external knowledge to improve the topic model representation on a dataset of academic publications.

Finally, a last method is Seeded LDA.[36] Instead of using a prior set of descriptive labels for each document or topic, as in previous approaches, Seeded LDA offers the possibility of manually defining a list of seed words for the topics the researcher is interested in. Let us imagine, for instance, that we are after a specific topic within the corpus of interest (e.g., news related to the relations between USA and Cuba in a newspaper archive): using Seeded LDA the researcher could guide the topic model in a specific direction, receiving as output the distribution of topics that she/he is interested in.

A thorough comparison of these different semi-supervised topic modeling techniques is beyond the scope of this paper. However, the fact that all methods make it possible to include the human (i.e., the humanities scholar) in the loop (i.e., the learning process) by requiring the expert to provide either labeled metadata, or a set of initial seed words to guide the topic acquisition process is crucial for out argument. We argue that this last option, in particular, is very attractive for digital historians in that it forces them to explicitly state the lexical components of the specific topics they are after, while requiring a minimal amount of supervision. That is, the scholar has to input a small set of seed words he/she deems important on the basis of her/his expertise, as opposed to merely labeling documents with a pre-compiled set of class labels.

5. HOW DATA BECOMES EVIDENCE

In the previous section we gave a brief overview of different semi-supervised topic modeling techniques, and argued that they could help historians exploit different sources like metadata and seed words, stemming from their *human expertise as scholars*, in order to perform fine-grained exploration analyses.

Topic modeling is a fascinating way of navigating through large corpora, and it could become even more interesting for the researcher by making the tool consider specific labels or seed-words. Regarding this, Owens remarked: 'If you shove a bunch of text through MALLET and see some strange clumps clumping that make you think differently about the sources and go back to work with them, great'.[37] Then, he continues: 'If you aren't using the results of a digital tool as evidence then anything goes'.

In the second sentence Owens perfectly describes the current main problem of digital humanities scholars employing text mining methods. As others already remarked, on the one hand the research community wants to see the humanistic relevance of these analyses, and not only the computational benefits.[38] On the other hand, digital humanists are aware that they cannot present the results of

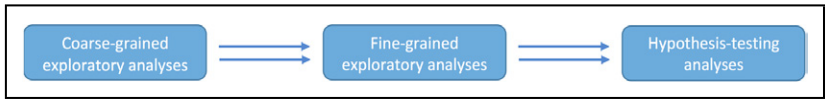

Figure 1. In this figure the methodological framework we suggest for analysing large historical corpora is summarized. Both the in-domain knowledge of the researcher and a solid expertise in data analysis are key components.

their studies as evidence without a solid evaluation of the performance of the methods.

For instance, if the purpose is to detect articles related to a specific subject (i.e. the relations between USA and Cuba), the documents obtained by looking at the distribution of specific (LDA-derived) topics are nothing more than an innovative way of searching through the dataset. Thus, it is important to keep in mind that these documents are not the only articles about the subject, and that maybe they are not even about that specific subject at all – due to the errors in the automatic learning process. Therefore, if we want to transform our data into evidences for supporting a specific argument or for confirming a hypothesis, we always have to evaluate our approach first.

It is interesting to notice that this specific process would sound perfectly ordinary if we were not talking about machine learning methods, computers and algorithms. When a researcher wants to be sure that a viewpoint is correct ('I believe this article is focused on the relations between USA and Cuba'), she/he will ask other colleagues.[39] The process described here is the same: we need human annotations (for example articles marked as 'being focused on the relations between USA and Cuba' or not) in order to confirm that our hypothesis (what the machine is showing to me are articles related to the relations between USA and Cuba) is correct.

Moreover, since humanists are working on extremely specific in-domain research tasks, they cannot rely on Amazon Mechanical Turk annotations as others usually do.[40] For solving this specific issue, they cannot even rely on computer scientists or data mining experts: they need the help of their peers.

Therefore, we believe that future advances in historical research on large corpora will be essentially achieved by *exploiting deep human expertise*, such as that provided by history scholars, as key components *within weakly-supervised computational methods* in two different ways.

In our vision (Fig. 1), a first stage will still consist of exploratory studies, which are extremely useful to develop an initial idea of arbitrary datasets. During this process, both standard LDA and especially the semi-supervised methods presented earlier could be particularly useful, as they will help researchers manage the vastness of digital data at their disposal. Following the exploratory phase, when the interest on a specific phenomenon has been established,

we envision researchers moving on and developing models to quantify such phenomenon in text, and creating a gold standard for evaluation based on human ground truth judgments – again, based on input from domain experts, i.e., scholars. During this second part of the study it might be that useful methods for exploratory studies (such as LDA) are not always as helpful when the task is to precisely identify specific phenomena. For this reason, the new generation of historians needs to learn how to employ text classification algorithms and have to become more and more confident with data analysis evaluation procedures.[41] As a matter of fact, these practices have the potential to sustain and improve our comprehension of the past, when dealing with digital sources.

6. Case study: applying these procedures in a well-defined historical research

In this final section we describe how we intend to employ the methodological framework presented before in an interdisciplinary research project that, in the near future, will bring together researchers from the Historical Institute and the Data and Web Science Group of the University of Mannheim.

Our cases study will be focused on circa 1,000 legal books from the 17th and 18th century, comprising over 310,000 pages of text. This is of course a large corpus for a historian, but only a small one for current research in computerized text analysis. Therefore, testing computational methods for specific analyses may proof insightful for both disciplines.

These volumes form the 'Juridica' part of a book collection brought to Mannheim by the learned Jesuit François-Joseph Terrasse Desbillons (1711–1798) in the 1770s. They cover a broad variety of legal matters with a special, but not very surprising interests in canon law, and another, little more surprising interest in legal history, or more precisely: the old (French) law.

Based on this corpus, we want to know more about this old French law, the 'ancien droit'. Yet, we do not trace legal institutions, ideas, or regulations. Rather we ask for the fundamental terms that old French law rested upon. These terms lay the conceptual groundwork upon which concrete institutions, rules, and distinctions of legal thinking were built. Hence, they are usually not technical in a stricter sense (i.e. not exclusively legal), or bear multiple semantic dimensions largely depending upon their uses in specific contexts, e.g. terms like *volonté* ('will'), *origin* ('origin'), or *liberté* ('liberty'). We aim to find these terms and their specific contexts, cluster together similar contexts, and weight them against each other, iteratively reaching a broad, yet precise spectrum of their meanings.

Traditionally, dictionaries like these are compiled by domain experts (i.e. historians) by reading large amounts of contemporary texts, and by analysing these texts in what we, broadly speaking, term a 'hermeneutical' fashion. The selection of texts rests upon the researcher and his or her scope of reach, its

amount on what he or she can physically read/bear, and its results rest largely on what he or she can find by physically reading either line by line or hastily flipping through the texts. This is not to say that this traditional method cannot or will not lead to fruitful conclusions.[42]

In the end, however, these projects are largely based on the presuppositions of the researcher about what she/he can (or will) actually find in the texts, and which texts will be more likely to give fruitful results. In other words, the researcher predefines both search terms and contexts. Our approach, in contrast, will also start with presuppositions, but iteratively enlarge them by finding both new contexts and probably even new search terms.

It could, for instance, well be that notions of 'will' (*volonté*) and its faculties will be discussed in contexts of compulsion (*contrainte*, *compulsion*, *coercition*) without even using a word deriving from *volonté*. Term-based textual analysis will not find such instances, but concept-based analysis will – even in far less obvious examples than the one given here.

As described before, our work will proceed through different steps. In the beginning, coarse-grained exploratory analyses (i.e. using standard LDA) will offer us a general idea of the content of the volumes and their similarities. Then, by combining different weakly-supervised techniques like Supervised LDA and Seeded LDA we will exploit domain expert knowledge to identify the semantic contexts in which these relevant concepts appear and to detect other similar patterns in the corpus. Finally, in order to use the results of these analyses as historical evidences, we will test, compare and improve our methods on a gold standard that it will be built with this specific purpose.

7. CONCLUSIONS

In this paper, we have discussed the applicability of a set of computational techniques for conducting fine-grained analyses on historical corpora. Furthermore, we have remarked the importance of an evaluation step when the data are exploited as evidence to support specific hypotheses. We believe that these practises will allow us to deepen our understanding of historical information embedded in digital data.

ACKNOWLEDGEMENTS

The authors want to thank Laura Dietz (Data and Web Science Group) and Charlotte Colding Smith (Historical Institute) for their precious methodological advice.

END NOTES

[1] https://books.google.com/ngrams; all the URLs mentioned in this research were lastly checked on November 13th 2015.

[2] J.B. Michel et al., 'Quantitative analysis of culture using millions of digitized books', *Science*, 331.6014 (2011), 176–182; A. Grafton. 'Loneliness and Freedom', *Perspectives on History*, online edition, March 2011, http://www.historians.org/publications-and-directories/perspectives-on-history/march-2011/loneliness-and-freedom.

[3] G. Crane, 'What do you do with a million books?', *D-Lib magazine*, 12.3 (2006).

[4] Grafton, 'Loneliness and Freedom'.

[5] See: http:// www . culturomics . org / Resources / faq / thoughts - clarifications - on - grafton-s-loneliness-and-freedom; F. Gibbs and T. Owens, 'The hermeneutics of data and historical writing', in J. Dougherty and K. Nawrotzki ed., *Writing History in the Digital Age* (Ann Arbor, MI, 2013).

[6] D. Cohen, 'Initial Thoughts on the Google Books Ngram Viewer and Datasets', *Dan Cohen's Digital Humanities Blog,* 19/10/2010, http://www.dancohen.org/2010/12/19/initial-thoughts-on-the-google-books-ngram-viewer-and-datasets/.

[7] See the answer to 'How does this relate to 'humanities computing' and 'digital humanities'?' in Culturomics FAQ section: http://www.culturomics.org/Resources/faq; C. S. Fisher, 'Digital Humanities, Big Data, and Ngrams, Boston Review, 20/06/2013, http://www.bostonreview.net/blog/digital-humanities-big-data-and-ngrams; C. Blevins, 'The Perpetual Sunrise of Methodology', 05/01/2015, http://www.cameronblevins.org/posts/perpetual-sunrise-methodology/

[8] I. Gregory, 'Challenges and opportunities for digital history', *Frontiers in Digital Humanities,* 1 (2014); M. Thaller, 'Controversies around the Digital Humanities: An Agenda', *Historical Social Research/Historische Sozialforschung* (2012), 7–23.

[9] O. Chapelle et al. (edited by), *Semi-Supervised Learning* (Cambridge, MA, 2006).

[10] T. Owens, 'Discovery and justification are different: Notes on science-ing the humanities', 19/11/2012, http://www.trevorowens.org/2012/11/discovery-and-justification-are-different-notes-on-sciencing-the-humanities/; D. Sculley and B. M. Pasanek. 'Meaning and mining: the impact of implicit assumptions in data mining for the humanities', *Literary and Linguistic Computing,* 23.4 (2008), 409–424.

[11] R. S. Michalski, J. G. Carbonell and T. M. Mitchell, *Machine learning: An artificial intelligence approach* (Heidelberg, 1983).

[12] E. Alexander et al. 'Serendip: Topic model-driven visual exploration of text corpora', *Proceedings of IEEE Conference on Visual Analytics Science and Technology* (Paris, 2014); M. Steinbach, G. Karypis, and V. Kumar, 'A comparison of document clustering techniques', *KDD workshop on text mining*. 400–1 (2000), 525–526.

[13] W. G. Thomas III, 'Computing and the historical imagination', in S. Schreibman, R. Siemens and J. Unsworth, ed., *A companion to digital humanities* (Oxford, 2004), 56–68.

[14] D. N. McCloskey, 'The achievements of the cliometric school', *The Journal of Economic History,* 38.01 (1978), 13–28.

[15] D. J. Cohen, and R. Rosenzweig. *Digital history: a guide to gathering, preserving, and presenting the past on the web* (Philadelphia, 2006).

[16] S. Robertson, *The differences between digital history and digital humanities*, 23/05/2014, http://drstephenrobertson.com/ blogpost/ the-differences-between-digital-history-and-digital-humanities/.

[17] S. Graham, I. Milligan and S. Weingart. *The Historian's Macroscope* - working title, Open Draft Version, Autumn 2013, http://themacroscope.org.

[18] For example the TAPoR project: http://www.tapor.ca/.

[19] In D. J. Cohen et al, 'Interchange: The promise of digital history', *The Journal of American History* (2008), 452–491.

[20] Willam J. Turkel' blog: http://digitalhistoryhacks.blogspot.com/; The Programming Historian: http://programminghistorian.org/.

[21] For example, I. Milligan, 'Mining the 'Internet Graveyard': Rethinking the Historians' Toolkit', *Journal of the Canadian Historical Association/Revue de la Société historique du Canada,* 23.2 (2012), 21–64.

[22] For instance, C. Blevins, 'Space, Nation, and the Triumph of Region: A View of the World from Houston', *Journal of American History,* 101.1 (2014), 122–147 and M. Kaufman, 'Everything on Paper Will Be Used Against Me: Quantifying Kissinger', 2014, http://blog.quantifyingkissinger.com/.

[23] For example, C. Au Yeung and A. Jatowt. 'Studying how the past is remembered: towards computational history through large scale text mining', *Proceedings of the 20th ACM international conference on Information and knowledge management* (Glasgow, 2011).

[24] E. Meeks and S. Weingart, 'The digital humanities contribution to topic modeling', *Journal of Digital Humanities,* 2.1 (2012), 1–6.

[25] D. M. Blei, A. Y. Ng and M. I. Jordan, 'Latent dirichlet allocation', *the Journal of machine Learning research*, 3 (2003), 993–1022.

[26] J. Chang et al., 'Reading tea leaves: How humans interpret topic models', *Advances in neural information processing systems*, 2009.

[27] R. Brauer, M. Dymitrow and M. Fridlund, 'The digital shaping of humanities research: The emergence of Topic Modeling within historical studies', *Enacting Futures: DASTS 2014* (Roskilde, 2014).

[28] T. Underwood, 'Topic modeling made just simple enough', *The Stone and Shell*, 07/04/2012, http://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/; Storify of the DH Topic Modeling Workshop: https://storify.com/sekleinman/dh-topic-modeling-seminar; Meeks and Weingart, 'The digital humanities contribution to topic modeling'.

[29] Owens, 'Discovery and justification are different: Notes on science-ing the humanities'.

[30] S. Marche, 'Literature is not data: Against digital humanities', *LA Review of Books* (2012); L. Wieseltier, 'Crimes against humanities', *New Republic,* 244.15 (2013), 32–39.

[31] D. Hall, D. Jurafsky and C. D. Manning, 'Studying the history of ideas using topic models', *Proceedings of the conference on empirical methods in natural language processing* (Honolulu, 2008); D. Mimno, 'Computational historiography: Data mining in a century of classics journals', *Journal on Computing and Cultural Heritage*, 5.1 (2012); M. Schich et al., 'A network framework of cultural history', *Science,* 345.6196 (2014), 558–562.

[32] J. D. Mcauliffe, and D. M. Blei, 'Supervised topic models', *Advances in neural information processing systems* (2008).

[33] T. Brown, 'Telling New Stories about our Texts: Next Steps for Topic Modeling in the Humanities', DH2012: Topic Modeling the Past, http://rlskoeser.github.io/2012/08/10/dh2012-topic-modeling-past/

[34] D. Ramage et al., 'Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora', *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (*Singapore, 2009).

[35] D. Mimno and A. McCallum 'Topic models conditioned on arbitrary features with Dirichlet multinomial regression', *Uncertainty in Artificial Intelligence*, 2008.

[36] J. Jagarlamudi, H. Daumé III and R. Udupa, 'Incorporating lexical priors into topic models', *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (Avignon, 2012).

[37] Owens, 'Discovery and justification are different: Notes on science-ing the humanities'.

[38] M. Thaller, 'Controversies around the Digital Humanities: An Agenda'.

[39] The examples presented here describe an over simplified case study. However, the complexity of the evaluation process can easily be shown by turning to more complex, realistic tasks like, for example, to identify how the different meanings of 'will' evolve within a reasonably sized historical corpus.

[40] In computational linguistics and natural language processing during last decade the use of human non-expert annotators for the construction of labeled datasets has become an established practice. To know more about the online labor market Amazon Mechanical Turk: https:// www.mturk.com/mturk/welcome.

[41] F. Sebastiani, 'Machine learning in automated text categorization', *ACM computing surveys,* 34.1 (2002), 1–47.

[42] For example R. Koselleck, W. Conze and O. Brunner ed. by, *Geschichtliche Grundbegriffe*, 8 vols. (Stuttgart, 1972–1997) and R. Rolf, E. Schmitt, and H. J. Lüsebrinck, *Handbuch politisch-sozialer Grundbegriffe in Frankreich, 1680–1820* (Berlin et al, 1985ff).