From Close to Distant and Back: How to Read with the Help of Machines

Rudi Bonfiglioli[1(✉)] and Federico Nanni[2]

[1] Textkernel, Amsterdam, The Netherlands
bonfiglioli@textkernel.nl
[2] University of Bologna, Bologna, Italy
federico.nanni8@unibo.it

Abstract. In recent years a common trend characterised by the adoption of text mining methods for the study of digital sources emerged in digital humanities, often in opposition to traditional hermeneutic approaches. In our paper, we intend to show how text mining methods will always need a strong support from the humanist. On the one hand we remark how humanities research involving computational techniques should be thought of as a three steps process: from close reading (identification of a specific case study, initial feature selection) to distant reading (text mining analysis) to close reading again (evaluation of the results, interpretation, use of the results). Moreover, we highlight how failing to understand the importance of all the three steps is a major cause for the mistrust in text mining techniques developed around the humanities. On the other hand we observe that text mining techniques could be a very promising tool for the humanities and that researchers should not renounce to such approaches, but should instead experiment with advanced methods such as the ones belonging to the family of deep learning. In this sense we remark that, especially in the field of digital humanities, exploiting complementarity between computational methods and humans will be the most advantageous research direction.

Keywords: Digital humanities · Text mining · Deep learning · Distant reading · Machine learning

1 Introduction

Digital humanities, originally known as humanities computing [1], is a diverse field of study that combines a humongous number of different interactions between humanities disciplines and the use of the computer. From the edition of manuscripts in digital form to the use of geographical information system in historical research, from man-computer interactions in media studies to the development of digital libraries, this field of study has gradually attracted the attention of the entire humanities community [2].

F. Gadducci and M. Tavosanis (Eds.): HaPoC 2015, IFIP AICT 487, pp. 87–100, 2016.
DOI: 10.1007/978-3-319-47286-7_6

Among these different applications of computational tools, researchers have in recent years consistently noticed the growth of a specific tendency in this field, characterised by the adoption of quantitative text mining methods for the study of digital sources [3].

Franco Moretti identified this practice with the concept of "distant reading" [4], namely the use of computational methods for the analysis of large collections of documents, usually adopted in opposition to traditional hermeneutic approaches. The notoriety gained by Moretti's works even outside academia [5] and the consistent growth in the adoption of these methods [6] have brought to the rise of two opposite factions in the humanities community [7,8]. Central to this division is the idea that computational methods seem to move in the direction of making the work of the humanist irrelevant for the production of insights, which could be obtained just by employing statistics and machine learning [9,10].

Starting with these assumptions, the purpose of our paper is twofold: first, we intend to stress how text mining methods will always need a strong support from the humanist, and second, we argue about the usefulness and necessity of advanced text mining approaches in the digital humanities.

In our study, we would like to think of humanities research involving computational techniques as a three steps process. The first step is a "close reading", which includes selecting a specific case study, crafting the initial features, and labelling of the training corpus. The second step is a "distant reading" since it involves performing a computational analysis. The third step is another "close reading", which consists of evaluation and interpretation of the results and the use of these results in a humanities research.

At the same time, we think that researchers should not renounce text mining approaches, but should instead experiment with advanced methods such as the ones belonging to the family of deep learning [11]. Deep learning techniques essentially perform representation learning, and therefore allow the automatic analysis of text as a multilayered set of encoded features.

This paper is organised as follows: firstly, the debate on the use of text mining methods in humanities research is introduced. Subsequently, our analysis schema is described. Then, we present a few existing advanced computational approaches and show how they could be beneficially employed in the digital humanities. Finally, we discuss the impact of the use of more advanced algorithmic approaches on the interaction between humanities research and the use of computers.

2 Text Mining Methods in Humanities Research

The interactions between humanities studies and the use of the computer have a long history [1]. Father Roberto Busa's Index Thomisticus [12], a complete lemmatisation of the works of Saint Thomas Aquinas developed in collaboration with IBM, is generally considered the starting point of the field originally called humanities computing [13]. In the following decades, different humanities

disciplines have approached computational methods for different purposes: from conducting stylistic analyses [14] to the realisation of geographical representation of events [15], from the digitisation [16] and encoding [17] of analogue sources to the dissemination of them through digital libraries [18].

In the same years, computational linguistics was also establishing its position in the academic environment [19]. Additionally, during the Eighties, this "close" field of study has dealt with a fundamental turning point in its methodology [20] with far reaching consequences for humanities computing as well. Previously, the most popular approaches in the field were characterised by the idea that language-knowledge was not predominantly derived by senses but already in the human mind [21]. This assumption made researchers orient their approaches towards the hand-crafting of knowledge and reasoning mechanisms in "intelligent systems". Due to several reasons, as the continuous advancement of computers, in the Eighties the mindset of researchers shifted more towards empiricism, which gave birth to the statistical approach that is still predominant in computational linguistics [22]. Following this methodology, knowledge regarding linguistic phenomena is extracted through the automatic analysis of large amounts of texts (corpora) and through the construction of predictive models.

In more recent years, the application of computational linguistics statistical methods has become a contradistinctive trait of a specific sub-group of digital humanities researches, for example stylometric tasks such as authorship attributions [23]. It's not until the last decade, however, that the application and discussion on the use of computational methods for the analysis of text contents has attracted the attention of the majority of the research community involved [24].

Franco Moretti has been identified as the scholar that brought this debate to the main public [25]. On the one hand, his publications on the use of computational techniques in order to extract quantifiable information from large amount of texts [4,26] attracted the attention of traditional humanities scholars [27] and of mainstream newspapers [5]. On the other hand, his "scientification" of literary studies practices [28], from the definition of "distant reading" to the creation of the "Stanford Literary Lab", suggested a completely different way of conceiving research in the humanities.

In his works, Moretti addresses in particular traditional close-reading approaches used in literary criticism, which are characterised by a careful interpretation of brief passages. In his vision, literature could and should be understood "not by studying particular texts, but by aggregating and analysing massive amounts of data" [5]. Several digital humanities scholars agree with Moretti's position [29,30]. They point out how computational methods could represent a solid alternative to traditional hermeneutic approaches, both in literary studies and in historical research, in order to deal with huge amount of sources in digital form.

Distant reading approaches have attracted great enthusiasm in digital humanities so far, but they have also received a series of specific critiques. First of all, it has been pointed out that these methods try to automatise an acquisition process of knowledge [31]. This might make the humanist scholar and his/her

background knowledge irrelevant to the production of insights and transform every aspect of these studies in the identification of quantitative features, aspects and evidences [10]. Secondly, its has been remarked how, for the moment, distant reading studies have developed an immense number of new tools, methods and techniques but produced so little in terms of new humanities knowledge [32].

As others have already pointed out [33], digital humanities seem to be often too easily seduced by the "big data" rhetoric of "making the data speak for itself". This is particularly evident by looking at one of the most adopted computational techniques for the study of text in digital humanities, Latent Dirichlet allocation (LDA) [34]. LDA is a statistical model that, given a corpus of documents, automatically identifies a pre-defined number of topics. Studying the distribution of these topics (effectively sets of words) in the corpus has been adopted for several different purposes in digital humanities scholarships [35], from exploring large corpora [36] to highlight content difference in scientific publications [37].

However, the use of LDA in digital humanities also highlights many of the flaws related to the use of computational methods in the discipline. Scholars seem to be attracted to it because it "yields intuitive results, generating what really feels like topics as we know them, with virtually no effort on the human side" [38]: being an example of an unsupervised learning technique, it requires no labelling of data, therefore little prior work from the humanist. However, although LDA can help to categorise big amounts of data, it can also generate ambiguous topics which makes it hard to draw deeper conclusion about the corpus [39], often calling for a lot of additional work for evaluating the quality of the results [40,41]. Producing valuable insights using LDA is difficult because the representations it learns for keywords/topics are judged semantically inferior to the ones achievable with more modern methods [42].

3 From Close to Distant and Back

Close reading practices in literary studies have a long and consolidated tradition [43]. Following this hermeneutic approach, scholars reach insights by considering a multitude of different factors, such as the choice of the vocabulary, the syntactical constructions employed, or knowledge of the author background or cultural and historical context. The attention of the researcher would be therefore focused on understanding the deeper meaning of representative passages, the choice of a specific word in a context, or the role of a rhyme in a poem. Through this process, humanities scholars discuss and define for instance how a specific combinations of values can signal "pathos" or "Victorian writing style"; then they reach insights by generalisation, recognising and further discussing the patterns of those combinations of values in other texts.

Ideally, we would like computational methods to be able to work in the same way: recognise those patterns, understand the relations among them, and then generalise them, allowing inference to be used to generate new insights. This would allow them to efficiently study corpora of large dimensions. In the language of computing and artificial intelligence, this means being able to learn

a good representation of our input through "features", which can encode the combinations of values ("pathos") expressed above. A perfectly trained machine would be able to recognise an already "read" Victorian novel, or to discriminate whether an unknown novel might be part of the Victorian movement, or even to be able to answer questions on whether new textual elements (syntactic constructions, use of words belonging to certain semantic field) might signal that we are reading a Victorian novel.

The main theoretical obstacle to create such a machine is that, for the purposes of humanities research, many additional more hermeneutic layers of "meaning" (and thus learnable relationships) might be added to the already complex, multi-layered medium we work with (text). On top of the standard syntactic and basic semantic layers, and maybe of the sentimental connotations, analyses in this field need to deal also, for example, with the layer capturing the cultural value of some words, or the layer that relates to the known historical background of the authors.

Since the quality of a machine learning approach can heavily depend on the choice of features [44], a first consequence of the observation above is that digital humanists are asked to encode hermeneutic layers of meaning into the features, a task that clearly requires solid domain knowledge obtained through close reading analyses. A second consequence is that digital humanities practitioners must be able to choose and adapt computational methods capable of learning complex representations: we will dedicate most of the next chapter to this issue, but what is clear is that it can require expertise in both the domain of artificial intelligence and humanities. In general, the first step of a research work in the domain of digital humanities must deal with formalising the research task, with adapting a chosen computational technique and with encoding the layers of meaning into a representation algorithms can understand. Such a step can be generally labeled as a close reading.

The second step is to run the computational analysis. Digital humanities is a fairly diverse field with research works aiming at different goals. For this reason, the output of the different computational tasks is also diverse, but in general the researcher is returned with some kind of organisation (more or less explicit) of (part of) the input which highlights some properties of it. For example, LDA returns explicit sets of words selected from the input, and can be used to query the distributions of those sets in the input documents. A Support Vector Machine [45] used for a classification task (e.g. authorship attribution) [46] returns a less explicit re-elaboration of the input (function expressing the decision boundaries) and can be used to query the label (author) corresponding to novel inputs.

The third step consists in drawing insights useful for humanities research from the output of the analysis [47]. While a computational method could capture additional relationships in the corpus, it is still the job of the humanist to query the right ones and then either validate them directly as new insights, use them to draw new conclusions, or discard them. Therefore, in this step the humanist should understand whether there is causation behind correlations, or decide to

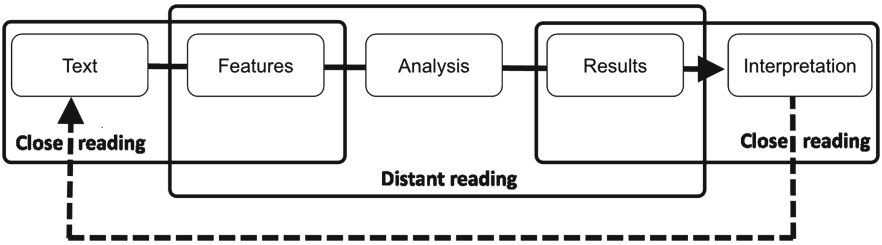

Fig. 1. The schema-model proposed in this paper.

go back to the first step and tune the model or the features (feature engineering) by looking at the current results. Once again, strong domain knowledge is clearly important in this step.

Figure 1 summarises the approach described above. By trying to formalise the practice of carrying on research in the digital humanities, it becomes clear how some points of distance between the two predominant positions in the field disappear. Being able to perform high-quality close readings is critical to succeed since it is very important in two steps out three: the role of the humanist is still essential for the production of insights.

4 Deep Reading

In the first step of the approach described in the previous chapter the researcher is facing the issue of crafting the features which encode the combinations of values essential to capture the layers of meaning we are interested in our analysis. At the same time, the computational method that we choose or craft often requires to be capable of learning multiple levels of representation, since we know that this is critical in order to capture interesting relations, and poor performance is often the reason why the results can not be used to reach valuable insights.

Recently, research in the field of machine learning turned heavily in the direction of deep learning, a family of algorithms that aims at learning automatically both good features or representations and an output from the input [48]. We believe that deep learning techniques could be extremely beneficial to a field such as digital humanities for a number of reasons. First, because they can decrease the cost of the feature engineering and annotation parts, since they can sometimes learn from un-labeled inputs. Additionally, they seem to learn features that result more general, adaptable and transferable when compared to the often over-specified,"manually" crafted ones. Finally, they fit the mental model of crafting a method that should capture different layers of meaning, resulting in an easier arrangement of computational analysis. In general, deep learning techniques work increasingly well with increasingly big input corpora, and this partners well with the current state of digital humanities, which has produced a large amount of digitalised sources from previous research works [49].

A good starting point for understanding the introduction of deep learning in the field of text analysis are the works related to word vector spaces, starting from the *word2vec* project [50]. Given each word, such methods compute a vector of high dimensionality that expresses and quantifies the relation between that single word and the rest of the text. All the word vectors form a (vector) space in which vectors representing similar words are located close to each other. *word2vec* models employ neural networks that try to capture linear regularities among words while being at the same time efficient to train, so that word vectors at high dimensionsionality (300–600) can be computed from "raw" un-labeled inputs of large size (few billions of words). Results [51] show how trained vector spaces seem to capture both grammatical (articles or verbs clustered together) and semantic (words for fruits clustered together) properties, being sensible to multiple degrees of similarity. Moreover, such word vector spaces appear capable of capturing relationships of a certain complexity; a famous example is the fact that, on a particular data set, subtracting the vector for "man" from the vector for "king" and then adding the vector for "woman" returned the closest vector to the one representing "queen". *word2vec* works by trying to predict either the probability of a single word appearing by knowing its neighbours, or the probability of certain neighbouring words appearing by knowing a pre-selected central word. Consequently, it estimates the vectors from such probabilities of "word embeddings". Given that language is "never, ever random", this seems to lead to representations that are sensible to multiple features of language and text, let them be syntactic or semantic: as already mentioned, such word vectors seem to capture relationships between words better than LDA, while being more efficient than LDA to train on large data sets [50]. Therefore, they can be used to trace relationships between concepts and characters outlined in a big corpus and subsequently derive valuable conclusions, as Bjerva and Praet [52] do by measuring proximity between latin historical figures and important concepts in texts spanning 2000 years of latin literature. Additionally, by using simple vector operations such as sum, researchers could query the space and check what is returned, for example, by taking the vector for "emperor", removing the vector for "compassionate" and adding the vector for "contentious": if the closest vector to this end-point "belongs" to an historical character, researchers may proceed with discussing whether it is likely that this character has been perceived as a fighting emperor.

Although *word2vec* generates word vectors from unlabelled data in the same way a neural network would do, it is a rather simple model that learns representation of words from a fairly basic feature: the way language positions words next to each other in complex texts. This potentially accounts for all the multilayered features we would like to learn a representation for, at the same time. We would like computational approaches to be able to learn representations for multiple features both by analysing them separately in depth and by analysing how they interact with each other. A popular model that seems to fit this is setting up a neural network composed of multiple, non-linear, interconnected layers: in case of representational learning, each layer will learn the representation of a

particular feature, and the entire neural network will learn the complex interactions between all the representations, which can be thought as hierarchical features. The learned features can be used by other computational analyses for example to classify inputs. When the input is text, the first level of the neural network usually works with word vectors computed in some of the ways examined above instead of simple words or sets of words, because they are a more effective representation of the meaning of each word.

Multi-layered neural networks seem to be a natural algorithmic counterpart to the close reading humanists perform, since they incorporate the idea that the distant reading must capture the contribution of many, complex features (e.g.: syntactic constructions, meaning related to the particular historical period) which influence meaning in non-trivial ways. In fact, despite in the domain of neural networks the neurobiological terminology is often erroneously used, such models do seem to mimic how the human brain works in some cases: for example, when it comes to vision, the first hierarchy of neurones that receives information in the visual cortex is sensitive only to specific edges or blobs while the following regions of the visual pipeline are sensitive to more complex structures like faces. Perhaps unsurprisingly, deep learning using multilayered (convolutional) neural networks saw its biggest successes when dealing with images, in problems such as image classification [53], but it has been successfully employed also in the field of natural language processing. For example, Socher et al. [54] craft a deep (recursive) neural network to perform a fine-grained sentiment classification of movie reviews excerpts, assigning not only the labels "positive" or "negative" but also "somewhat positive/negative" or "neutral". What is interesting is that the model learns for example that a negated negative sentence should be classified as "less negative" than a negative sentence although not necessary positive ("The movie was not terrible" mostly means the movie was less bad than a terrible one, but not necessary terrible, as the authors remark) without having any part of the system that has the explicit goal of recognising this complex (due to both syntactic an semantic reasons) pattern. Another interesting work [55] aims at generating high-quality word vectors that can learn more semantic, less syntactic relationships. It employs a multi-layered neural network with one layer trying to learn a representation ("global semantic vectors") from a global, document-wide context: the same architecture could be employed to train a vector space which could be more sensitive to hermeneutic (e.g.: stylistic) traits of texts by simply changing the way the "global semantic vectors" are computed, making it an interesting solution for distant reading for digital humanities. It is worth noticing that a multi-layered model can also leverage on existing knowledge: for example, Trask et al. [56] introduce a model that tries to learn less ambiguous word vectors (where "apple" is split into multiple tokens, one of them clustering close to "pear" and "banana" and the other close to "samsung" and "google") by replacing an unsupervised cluster with a Part-of-Speech tagger, therefore learning word representations that benefit from established methods capable of recognising certain features (parts of speech, such as nouns, adjectives).

To our knowledge, digital humanities currently lack examples of works that successfully incorporate deep learning techniques, thus performing what we could call a "deep distant reading". This is probably because such methods are still subject of state of the art research in machine learning and artificial intelligence, and in order to become part of the toolset of a (digital) humanist, they should become part of easy-to-use toolboxes (such as MALLET[1] for LDA). However, we think that such approaches could be beneficial to digital humanities in the future because they mimic the way we approach the analysis of a text as humans and because they offer an alternative to the difficult hand-crafting of features, learning instead representations which quality not only seems to scale well with the amount of available input but that are also easier to "transfer" from task to task.

5 A New Humanist

5.1 A Generation of Humanists - Machine Learning Experienced Users

In this paper, while describing the different aspects of our analysis-schema and introducing the usefulness of deep learning methods, we relied on a clear assumption: that humanities scholars must be able to conduct distant reading analyses. However, this is in most cases not true. In particular, traditional humanities curricula usually foster qualitative hermeneutic approaches over quantitative statistical analyses and tend to adopt the computer only as an advanced typewriter. While it is not the aim of this paper to discuss the pros and cons of this situation, it is important to remark that the lack of a "scientific/computational background" can be a real issue when conducting a distant reading analysis.

First of all, the absence of a solid knowledge of data analysis has serious consequences for the humanities scholar who intends to use text mining methodologies, since it can limit both his/her capacity to engineer/re-adjust features and to adapt the chosen computational technique (as also remarked in [57]). Moreover, his/her understanding of quantitative results will be always partial, compared to the one exhibited by other researchers from other disciplines (such as computational linguistics [37], natural language processing [58] or information retrieval [36]) that are currently also experimenting with text mining methods to solve humanities tasks.

Secondly, the traditional lack of programming skills and algorithmic thinking of humanities researchers will always force them to establish collaborations with computer scientists or software engineers. However, even if these interactions have led to a number of successful joint research projects [59], it is also known that these interdisciplinary collaborations could be difficult to conduct (and expensive for the humanities research [60]), as different backgrounds, approaches and expectations have to continuously focus on a common goal. During the last decade this knowledge gap on computational methods has guided digital

[1] http://mallet.cs.umass.edu/.

humanists on preferring exploratory studies (employing easy-to-use toolboxes) over quantitative hypothesis-testing research projects. This has in turn limited the potential of text mining in digital humanities studies so far, especially in attempting knowledge discovery [61].

In this complex scenario, we believe that a solution may exists. As we described in this paper, digital humanities scholarships that focus on the use of advanced computational methods need a solid research focus and expertise both in advanced computational techniques and in data-analysis practices. For this reason we think that, especially for improving the usefulness of distant reading approaches in humanities scholarships, this knowledge has to be consistently integrated in educational programs focusing on digital humanities. In our opinion, this field not only needs a generation of programmers, as Turkel once suggested [62], it needs a generation of humanities scholars that are also machine learning experienced users.

5.2 Complementarity Is the Key

As we have already mentioned, the debate on the effectiveness and usefulness of computational methods in the humanities seems sometimes to raise the question of whether the use of computers might substitute, even partially, the contributions of humans. This seems to happen at a time when scholars of different disciplines are discussing the implications of artificial intelligence in various domains, and their impact on society. Observing some recent milestones of artificial intelligence, the question of whether machines could substitute humans in performing many different tasks has been raised, and various arguments supporting a positive answer have been proposed.

For what we have observed as practitioners of digital humanities, we believe that understanding and pursuing complementarity between the humanist and the "machine" is the key to achieve great results in the field, in the same way it might be a way to also keep society prosperous with the advance of automation [63]. As it is clear by the framework we propose, the domain specific knowledge of the humanist is still fundamental, for example in tailoring the computational analysis and interpreting the results, and the adoption of advanced algorithms simply augments the possibilities of the humanist, who can use machines to perform a meaning-aware heavy-lifting on large corpora that can expose certain patterns. In fact, the tech-industry is following this path too, with companies like Palantir Technologies developing advanced data analysis products explicitly made to work with humans and to help them making critical decisions (e.g. in counter-terrorism situations) [64]. Therefore, we think that the new humanist should be aware of the importance of his role, capable of understanding how it can complement the machine to achieve the best results, and should be open to participate in the development of tools and technologies that could augment his/her capabilities.

6 Conclusions

Having observed the emerging factions in digital humanities, we proposed a three-steps framework to conduct research using text mining techniques, and showed how the framework helps, reasoning at a deeper philosophical level, to blur the contrasts present in the field. We think that the use of advanced computational methods is an important area of research that must be pursued, and argue that deep learning could be beneficial. Moreover, we stressed the importance of understanding that qualitative knowledge rooted in the domain of humanities is essential and can not be ignored by works focused on computational methods. In this sense, we believe that, especially in the field of digital humanities, exploiting complementarity between advanced computational methods and humans will be the most advantageous research direction.

References

1. Hockey, S.: The history of humanities computing. In: A Companion to Digital Humanities, pp. 3–19 (2004)
2. Svensson, P.: The landscape of digital humanities. Digit. Humanit. (2010)
3. Berry, D.M.: The computational turn: thinking about the digital humanities. Cult. Mach. **12**, 2 (2011)
4. Moretti, F.: Distant Reading. Verso Books, London (2013)
5. Schulz, K.: What is distant reading. The New York Times 24 (2011)
6. Weingart, S.: Submissions to DH2016 (pt. 1) (2016). http://www.scottbot.net/HIAL/?p=41533
7. Underwood, T.: Why digital humanities isn't actually 'the next thing in literary studies'. The Stone and the Shell 27 (2011)
8. Underwood, T.: The literary uses of high-dimensional space. Big Data Soc. **2**(2) (2015)
9. Marche, S.: Literature is not data: against digital humanities. LA Review of Books 28 (2012)
10. Posner, M.: Humanities data: a necessary contradiction (2015). http://miriamposner.com/blog/humanities-data-a-necessary-contradiction/
11. Bengio, Y.: Learning deep architectures for AI. Found. Trends Mach. Learn. **2**(1), 1–127 (2009)
12. Busa, R.: Index Thomisticus Sancti Thomae Aquinatis Operum Omnium Indices Et Concordantiae in Quibus Verborum Omnium Et Singulorum Formae Et Lemmata Cum Suis Frequentiis Et Contextibus Variis Modis Referuntur (1974)
13. Dalbello, M.: A genealogy of digital humanities. J. Documentation **67**(3), 480–506 (2011)
14. Stamatatos, E.: A survey of modern authorship attribution methods. J. Am. Soc. Inform. Sci. Technol. **60**(3), 538–556 (2009)
15. Knowles, A.K.: GIS and history. In: Placing History: How Maps, Spatial Data, and GIS are Changing Historical Scholarship. Esri Press (2008)
16. Boschetti, F., Romanello, M., Babeu, A., Bamman, D., Crane, G.: Improving OCR accuracy for classical critical editions. In: Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonas, G. (eds.) ECDL 2009. LNCS, vol. 5714, pp. 156–167. Springer, Heidelberg (2009). doi:10.1007/978-3-642-04346-8_17

17. Ide, N., Veronis, J.: Text Encoding Initiative: Background and Contexts, vol. 29. Springer Science & Business Media, Dordrecht (1995)
18. Rydberg-Cox, J.: Digital Libraries and the Challenges of Digital Humanities. Elsevier, Boston (2005)
19. Mitkov, R.: The Oxford Handbook of Computational Linguistics. Oxford University Press, New York (2005)
20. Manning, C.D., Schutze, H.: Foundations of Statistical Natural Language Processing, vol. 999. MIT Press, Cambridge (1999)
21. Lenneberg, E.H., Chomsky, N., Marx, O.: Biological Foundations of Language, vol. 68. Wiley, New York (1967)
22. Nadkarni, P.M., Ohno-Machado, L., Chapman, W.W.: Natural language processing: an introduction. J. Am. Med. Inform. Assoc. **18**(5), 544–551 (2011)
23. Juola, P.: Authorship attribution. Found. Trends Inf. Retrieval **1**(3), 233–334 (2006)
24. Kirschenbaum, M.G.: The remaking of reading: data mining and the digital humanities. In: Proceedings of the National Science Foundation Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation, Baltimore, MD (2007)
25. Rothman, J.: An Attempt to Discover the Laws of Literature. The New Yorker (2014)
26. Moretti, F.: Graphs, Maps, Trees: Abstract Models for a Literary History. Verso Books, London (2005)
27. Liu, A.: The state of the digital humanities: a report and a critique. Arts Human. High. Educ. **11**(1–2), 8–41 (2012)
28. Merriman, B.: A Science of Literature. Boston Review (2015)
29. Jockers, M.L.: Macroanalysis: Digital Methods and Literary History. University of Illinois Press, Urbana (2013)
30. Graham, S., Milligan, I., Weingart, S.: The Historian's Macroscope: Big Digital History. Imperial College Press, London (2016)
31. Fish, S.: Mind your P's, B's: The digital humanities and interpretation. New York Times 23, no. 1 (2012)
32. Blevins, C.: The Perpetual Sunrise of Methodology (2015). http://www.cameronblevins.org/posts/perpetual-sunrise-methodology/
33. Owens, T.: Discovery, justification are different: Notes on science-ing the humanities (2012)
34. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
35. Meeks, E., Weingart, S.: The digital humanities contribution to topic modeling. J. Digit. Humanit. **2**(1) (2012)
36. Yang, T.I., Torget, A.J., Mihalcea, R.: Topic modeling on historical newspapers. In: Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pp. 96–104. Association for Computational Linguistics (2011)
37. Hall, D., Jurafsky, D., Manning, C.D.: Studying the history of ideas using topic models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (2008)
38. Weingart, S.: Topic Modeling and Network Analysis. The Scottbot Irregular (2011)
39. Rhody, L.: Topic modeling and figurative language. J. Digit. Humanit. **2**(1) (2012)
40. Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L., Blei, D.M.: Reading tea leaves: how humans interpret topic models. In: Advances in Neural Information Processing Systems, pp. 288–296 (2009)

41. Nanni, F., Fabo, P.R.: Entities as topic labels: improving topic interpretability and evaluability combining entity linking and labeled LDA. arXiv preprint arXiv:1604.07809 (2016)
42. Maas, A.L., Ng, A.Y.: A probabilistic model for semantic word vectors. In: NIPS Workshop on Deep Learning and Unsupervised Feature Learning (2010)
43. Wolfreys, J.: Readings: Acts of Close Reading in Literary Theory. Edinburgh University Press, Edinburgh (2000)
44. Domingos, P.: A few useful things to know about machine learning. Commun. ACM **55**(10), 78–87 (2012)
45. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 137–142. Springer, Heidelberg (1998). doi:10.1007/BFb0026683
46. Diederich, J., Kindermann, J., Leopold, E., Paass, G.: Authorship attribution with support vector machines. Appl. Intell. **19**, 109–123 (2003)
47. Sculley, D., Pasanek, B.M.: Meaning and mining: the impact of implicit assumptions in data mining for the humanities. Literary Linguist. Comput. **23**(4), 409–424 (2008)
48. Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **5**, 1798–1828 (2013)
49. Christenson, H.: HathiTrust. Libr. Res. Techn. Serv. (2011)
50. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Proceedings of Workshop at ICLR (2013)
51. Mikolov, T., Yih, W.T., Zweig, G.: Linguistic regularities in continuous space word representations. In: HLT-NAACL (2013)
52. Bjerva, J., Praet, R.: Word embeddings pointing the way for late antiquity. In: LaTeCH (2015)
53. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (2012)
54. Socher, R., Perelygin, A., Wu, J.Y., Chuang, J., Manning, C.D., Ng, A.Y., Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (2013)
55. Huang, E.H., Socher, R., Manning, C.D., Ng, A.Y.: Improving word representations via global context and multiple word prototypes. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (2012)
56. Trask, A., Michalak, P., Liu, J.: sense2vec-a fast and accurate method for word sense disambiguation. In: Neural Word Embeddings (2015)
57. Nanni, F., Kuemper, H., Ponzetto, S.P.: Semi-supervised textual analysis, historical research helping each other: some thoughts and observations. Int. J. Humanit. Arts Comput. (2016)
58. Mimno, D.: Computational historiography: data mining in a century of classics journals. J. Comput. Cult. Heritage **5**, 1–19 (2012)
59. Siemens, L.: It's a team if you use 'reply all': an exploration of research teams in digital humanities environments. Literary Linguist. Comput. **24**, 225–233 (2009)
60. Crymble, A.: Historians are becoming computer science customerspostscript (2015). http://ihrdighist.blogs.sas.ac.uk/2015/06/24/historians-are-becoming-computer-science-customers-postscript/
61. Thaller, M.: Controversies around the Digital Humanities: An Agenda. Historical Social Research/Historische Sozialforschung (2012)

62. Cohen, D.J., Frisch, M., Gallagher, P., Mintz, S., Sword, K., Taylor, A.M., Thomas, W.G., Turkel, W.J.: Interchange: the promise of digital history. J. Am. Hist. (2008)
63. Autor, D.H.: Why are there still so many jobs? The history and future of workplace automation. J. Econ. Perspect. **29**, 3–30 (2015)
64. Top, N.M.: Counterterrorism's new tool: 'Metanetwork' analysis (2009)