12 Machine learning techniques for the management of digitised collections

Mathias Coeckelbergs and Seth Van Hooland

Introduction

If we believe media such as the New York Times, artificial intelligence (AI) and machine learning techniques have the potential to automate a wide range of societal challenges, from the detection of cancer to self-driving cars (Lohr 2016). Given enough content to analyse and practice on as a training set, algorithms can develop statistical models to replace decision-making ordinarily perceived as requiring human intelligence, such as driving a car in traffic or interpreting an X-ray scan. There is no lack of content to analyse. There is too much of it to be handled manually.

In this context, automation is no longer seen as a feature that is “nice to have”, but is a necessity. However, there are a lot of misunderstandings and false hope circulating in the archives and records management community. We use the terms information management and archives and records management interchangeably throughout this paper. A complex debate could be had on the exact boundaries and definitions of these two fields, but automation has a role to play in both. However, we decided not to focus on the exact definitions of each field. This chapter therefore seeks to give professional users such as researchers and students a better understanding both of the possibilities and the limits of automation.

The first half of the chapter develops a typology of the different methods and tools which have been in use for decades to automate particular aspects within the life cycle of information in archives. We believe that machine learning methods can greatly contribute to the life cycle of information, as they can perform manipulations on the raw data, unmediated by previous selection mechanisms. This overview also includes a review of the relevant current literature. The latter half of the chapter then focusses on a more detailed description of one specific technique which can be applied by virtually any records manager or archival collection holder: topic modelling (TM). In the context of this chapter, we consider TM as a “gateway drug” for the information management community. Once you get “hooked” on easily implemented, unsupervised analyses, it opens the door to more enhanced machine learning techniques. It is easy to get

quick results and it may trigger insights into the implementation of more complex and resource-intensive machine learning methods and tools. To make the introduction to TM as pragmatic as possible, we present the technique with the help of a real-life, large-scale case study. In the study, TM was applied on an archival corpus of the European Commission (EC) using a statistically significant sample. Having presented the results of the case study, the chapter concludes with some remarks on how the information management community can develop a more pragmatic view on automation.

The remainder of this chapter will first provide an overview of the state of the art of both automation for information management and how machine learning was introduced in the text mining field in general and in records management in particular. Then, we discuss the usefulness of one specific topic modelling algorithm, Latent Dirichlet Allocation (LDA) and how it can help us give an initial assessment of the applications of machine learning in an archival setting. This will allow us to develop a case study in which we use LDA to link documents to entries of the EuroVoc thesaurus. The chapter ends with a discussion and conclusion of the above.

A short history of automation in information management

Electronic records management may seem to be a tautology, since records management is traditionally associated with archives that keep track of their records without the use of information technology. The term electronic records management is used more and more often as the field is shifting almost exclusively to digital methods. A complex debate could be had on the exact boundaries and definitions of the two fields, but automation has a role to play in both. However, we decided not to focus on the exact definition of the boundaries of each field. Archival documents can also include e-mails, for example. But, as Bailey pointed out almost a decade ago, the status or role of automation within the professional community remains a “somewhat ill-defined and rather arbitrary one” (Bailey 2009, 91). Take, for example, the principles underpinning virtually every electronic document and records management system on the market today: records must exist within a container of some kind (as if they were a physical item). Here we understand records as the documents we want to archive and containers as the smallest unit in which they are collected. They can only exist in one place within the system (again, as if they were a physical item) and are allocated this location within a hierarchical classification scheme that organises content as if physically arranging it. We add additional metadata to records to enable multiple search routes in the same way as subject-based index cards (cards sorted according to a pre-defined list of subjects) did for our paper-holdings in the days before databases; and we define retention and destruction periods according to when the file was opened or closed or the record was created, in a way that

would have been immediately familiar to a registry officer from the 1960s. In short, we try to manage electronic records as we have always managed paper records: and this is why we are failing. For in doing so we have missed the crux of the issue, which is that the real challenge is not the fact that records are now electronic: it is the sheer volume with which they are now being produced. That requires an entirely different approach to their management.

It seems that a new wave of enthusiasm emerges every decade. With hindsight, three different stages can be identified, each of which can be roughly mapped to a particular decade. On each occasion, the archivist community puts its hope in the arrival of new hardware and software paradigms:

- In the 90s began the ubiquitous usage of desktops and client-server architecture. The popularity of monolithic EDRMS (electronic document and records management systems) applications, such as Documentum reached a maximum.
- At the end of the 90s and the beginning of the century there developed the global usage of the Web and the introduction of Web-based software.
- From 2010 onwards, cloud computing started to take off commercially at a large scale. The period is marked by major advances in natural-language processing (NLP) and increased interest in Linked Data for the creation and sharing of virtual authority files. For example, the Social Networks and Archival Context (SNAC) Project demonstrates how existing descriptions in the Encoded Archival Context – Corporate bodies, Persons, Families (EAC-CPF) format can be leveraged as Linked Data. During the same period, interest in the use of Machine Learning to extract meaning from non-structured text also rose.

While it is relatively straightforward to draw up a typology of methods, as we have just described, it can be more of a challenge to describe the various aspects of the life cycle of documents and files to be automated. In many cases, it is more accurate to talk about semi-automation, since human intervention is often still required in the life cycle of documents. Elements of this semi-automated cycle include the following:

- appraisal and record declaration, based on worth: in a paper-based environment, records managers and archivists were responsible for identifying the business value of documents. In an electronic environment, this task has been shifted onto staff members. A large body of literature documents the inadequacy of such an approach, as described by (Vellino 2016);
- classification of a record into a file;
- development of a controlled vocabulary and its application;
- identification of personally identifiable information.

The process of appraisal establishes the value of information, qualifying its value and determining the retention period (Duranti 1994, 329). A recent thorough study on the possibilities and limits of automatically appraising e-mail was conducted by Vellino (2016). He describes the need to formalise the organisational context, which requires considerable effort.

In the first step, a qualitative study of the e-mail appraisal behaviour of records management experts is performed based on semi-structured interviews and cognitive inquiries, followed by data analysis. Based on this input, an abstract classification model was built, consisting of two top-level categories: “e-mails of business value” and “e-mails of no business value”, which were further divided into 13 sub-categories. The insights of eight records management experts were used as training data for a support vector machine (SVM), one of the most widely used machine learning methods for classification. The trained model was then tested on a collection of 846 e-mails, taken from two of these experts’ mailboxes. The main result of this experiment was that, although it is very complex to establish a consensus within an organisation on what is “business value”, machine classification models nevertheless achieve a high degree of accuracy. The most important criteria appeared to be the keywords in the subject field and the textual features from the e-mail body. This result shows that a smart textual representation, based on machine learning, together with insight into the indexing of these emails by their subject field, gives promising results for automatic methods of classifying e-mails for long-term preservation as an on-demand service.

At this point we are confronted with one of the central problems of metadata: they are infinitely extendable. Boydens has pointed out this danger by referring to Friedrich Nietzsche’s “Vom Nutzen und Nachtheil der Historie fuer das Leben” (“About the benefits and disadvantages of history for life”, in which Nietzsche aims to demonstrate that we need to “*serve history only insofar as it serves living*” [Boydens 1999]). The different layers of metadata that are added, one on top of another, result in highly complex documentation practices that are difficult to maintain. The ability to comment on a user review of an Amazon product provides an illustrative example. The user comment, which can be considered a form of metadata on the book that a customer bought, is turned into data itself by attaching metadata to it in the form of a review of the comment.

Metadata evolve through time, but so do the objects and realities that they describe. Boydens and Hooland (2011) explained that the notion of the *lead department* is one of the conceptual building blocks of a records management policy. Throughout the entire life cycle of a record, it should be clear who has responsibility for a file. However, as Desrochers says, “the properties of organizational structures have shifted from a hierarchical command and control structure – a Weberian state – to one where an organization exists and interacts in a network with other institutions” (Desrochers et al. 2011). Phenomena such as mergers, acquisitions, take-overs and bankruptcies also

challenge the notion of having an established entity responsible for a file. In the aftermath of the Fortis and Dexia bankruptcies, questions arise regarding their management, based on the information decisions that were taken. If no clear records management policy was implemented at the time of these crises, it will be impossible to present legally sound information in about who was aware of what information and when.

The National Archives and Records Administration (NARA) has identified five different approaches to automation. Based on their typology, three distinct methods can be differentiated:

- **Rule-based automation:** Based on prior established rules, documents can be assigned records, filed and given retention rules. These rules can vary widely in their complexity. An example of a relatively simple rule is the approach promoted in the “capstone” method of managing e-mail. This records management method requires organisations to designate specific inboxes for accounts that are likely to receive crucial e-mails. In this case, an organisation might, for example, declare the rule that, starting from a certain level in the hierarchy of an organisation, all e-mail accounts are ingested as records unless the user specifies the contrary. Rules might also be triggered by certain types of metadata that indicate a certain type of document (e.g., a contract). More complex rules can be composed by combining formal and contextual elements of a document with actual content by making use of regular expressions. For example, all documents which contain the string “final” can be grouped, as they probably indicate all finalised documents in a collaboration workflow.
- **Business processes and automated workflows:** Most organisations use process-driven software to manage specific workflows such as procurement. This approach consists of assigning specific metadata within these applications in an automated manner. For example, in an application that allows citizens to apply for permits, each document can be assigned a record when the application file is finalised and the file is transferred to the digital archive. Currently, most workflow packages can be connected to other applications, making this scenario relatively easy to implement for structured documentation. If an organisation only uses this model, it should be aware that the unstructured documentation must also be managed and brought into context with the automated records from workflow applications.
- **Auto-classification:** This is the most complex and advanced way to automate workflows, using the content of documents to convert them into records and assign them the necessary metadata. These processes often use machine learning software, where a training corpus is used to configure the software iteratively. The more formal and repetitive the content of the documents, the better the results of this approach. It can also be combined with rule-based methods and workflow automation based on those rules. The software and methodology that are

available in the field of auto-classification are relatively new but offer many advantages over traditional classification, based on keywords or functions, rather than using generic rules that do not take into account the actual content of the documents. The raw data of these documents contain valuable information that is left untouched by these methods. They usually require making highly specific and granular classification schemes more general and evolving towards a “big bucket” method of classification. In this case, the rule-based system becomes too detailed to still be efficient, given the complexity of the system. The difficulty of this method lies in the fact that there are currently no satisfactory benchmark studies available that objectively compare the performance of different methods and tools.

The premise for auto-classification is having existing categories based on a classification scheme that is already in place, which can be used to manually label a training set. The data, together with this set of labels, can then be used to train an algorithm. Here we can also refer to the essential caveat in the NARA report (Automated Electronic Records Management Report, Plan of the NARA). It points out in the introduction that “All the automated approaches described in this report and plan depend on having a solid records management policy in place. Automation is a tool, not a replacement for a professional records and information management program” (NARA 2014, 4).

The starting point of this section was to know what the worst-case scenario is: a large collection of non-structured documents with little to no manually pre-assigned metadata offers a significant challenge for the implementation of machine learning methods. In the remainder of this chapter, we will discuss a workflow for tackling this issue together with some preliminary results.

Rise of machine learning and its introduction in records management

The history of Machine Learning can be traced back to the 1950s, with the inception of Turing’s Learning Machine and stochastic neural analog reinforcement calculator (SNARC), the first neural network algorithm, built by Marvin Minsky and Dean Edmonds. These early, statistics-based algorithms laid the groundwork for the subsequent development of important machine learning tools such as the perceptron by Frank Rosenblatt in 1958 (Rosenblatt 1958), the nearest neighbour algorithm in 1967 (Mack and Rosenblatt 1979), backpropagation in 1986 (Rumelhart 1986) and the random forest algorithm in 1995 (Tin-Kam 1995). All these discoveries form an important part of the current machine learning landscape, even though at the time of their discovery and immediately thereafter they were not as widely used as they are today.

The computational underpinnings of the discovery of these algorithms were still in an embryonic stage, meaning that the available data were to a certain extent structured, allowing the rules to be deduced on which to build a system. Of course, these rules were not able to deal with all conceivable situations, but the exceptions could be measured. Since encoding explicit rules was easier both computationally as well as conceptually than using machine learning, rule-based systems have been very popular until fairly recently. It is only with the advent of big data methods and the computational power associated with them that investigations into machine learning methods could begin.

Even though the recent popularity of neural networks, as the most successful branch of machine learning methods, moves the goalposts for state-of-the-art research, they are still considered to be a black box. Although their results are changing the way we think about the treatment of data, it is still impossible to grasp what exactly is going on inside. Together with these algorithmic advances during the last two decades, we have seen a rise in not only the amount of data and documents available but also in the variety of data types, the complexity of resources and the unstructured nature of information. This shift in the landscape has made the rule-based methods that thrived in the 20th century outdated at best and often even obsolete in the context of the surge of big data.

In summary, we may claim that we are witnessing a shift from knowledge-driven methods (rule-based) to data-driven methods. This means that traditional rules are generally left behind, so that statistics-based machine learning systems can provide structure in the wealth of information available today. The danger of the rule-based approach is that, if the rules miss a scenario, noise is generated as output, which requires ever more rules to describe every possible scenario. In the machine learning approach, on the other hand, the user has to feed the system good examples of input data with the desired labels, which allows the system to learn the relationships between input and output in this training phase. The system then learns to classify data according to the thresholds that it acquires during training. Of course, this method is also not without its problems, since noise can arise from bad training data or overfitting, the learning deficit where the algorithm does not abstract well from the data. Nevertheless, the main asset of this approach is that it is based on the data themselves and not on knowledge input from the user. Chiticariu (2013) has noted that, although machine learning seems to be a more advanced method of dealing with data, this is a phenomenon mainly demonstrated in academia, whereas industry still heavily relies on rule-based applications. Industrial users are among the many voices declaring that the distinction between the two does not have to be that strong; and several hybrid methods have been proposed, for example, by Villena-Román (2011). Many, however, are often more reluctant to admit their use of rules, disguising them with euphemisms such as “dependency restrictions” (Schmitz Mausam et al. 2012) or “entity type constraints” (Yao 2011).

Topic modelling

The field of unsupervised machine learning, where an algorithm is designed to find clusters in the data without predetermined labels provided by humans, has proven to be important in providing a first analysis step in big data pipelines. Supervised machine learning, where labelled data is readily provided, can occur further down the pipeline, or else as a stand-alone technique that is focussed more strongly on specific NLP tasks such as text classification using pre-defined categories. Since our starting point of this chapter is raw textual data, we shall proceed with a primary focus on unsupervised machine learning. In this field, TM has gained momentum within the Humanities over the last few years in analysing topics represented by large volumes of full text. Within the digital humanities (DH) community, TM has attracted a fair amount of interest and is increasingly being used to access and explore large corpora of full-text documents (Chang 2009; Goldstone et al. 2012; Klein et al. 2015).

In this specific implementation of TM we use the famous LDA algorithm (Blei 2003), which forms the basis of several derived TM algorithms. Applied to large textual datasets, it shows great promise in successfully clustering similar texts, i.e., determining features based on patterns found in the raw data. This approach, along with other text-mining algorithms, has gained momentum in document classification based on features learned by a machine learning system, as pointed out by Suominen et al. (2016) in the specific field of bibliometrics or by Newman (2010) in a library context. Similarly, Roe (2016) uses LDA to draw a map of all human knowledge contained in the French *Encyclopédie* of *d'Alembert and Diderot*. This application of LDA indicates how this taxonomy of human knowledge, contained in their encyclopaedia, can be approximated automatically.

TM can be considered as a “plain vanilla” approach to text mining, since it is developed in the context of information retrieval. The main goal of this field is to retrieve the most relevant documents corresponding to a query. This can be compared to word embeddings, for example, which are equally famous but have been developed within the field of computational linguistics, where the main goal is not the retrieval of relevant documents but the correct modelling of the semantics. TM can be applied in any context where there is a large volume of non-structured documents (e-mail, social media, reports, etc.) and the outcomes can be used for different goals.

To investigate the value of TM for the archival community, this chapter seeks to critically assess its potential for the “distant reading” of archival data. Developed by Moretti (2007), distant reading practices make use of statistics and computational linguistics to extract specific features automatically from large corpora, providing a means to spot trends and shifts over time. This is in contrast to close reading, which uses only human understanding and classification to identify the relatedness of texts and the relevance of features in order to achieve a classification. Traditionally, historians explore

archives based on an inventory, which contains metadata at the fonds, series or file level. Only very rarely do historians have access to metadata on a document level. However, within the current context of mass digitisation of archival holdings, institutions often end up with millions of OCRred text files, having only minimal “tombstone” metadata at either the fonds, series, file or document level. In the absence of traditional access paths, based on metadata that are often not available or limited, innovative distant reading methods such as TM can provide alternative ways to explore large archival holdings and immediately drill down to the content at a document level.

Case study

When and how did environmental considerations start to influence agricultural policy development by the EC? What are the key documents to analyse the debate on nuclear energy production from the 1960s onwards? These are two examples of research questions that historians might frame. In this context, the mass digitisation of the EC’s archives offers exciting new ways to query and analyse the archival corpus in an automated fashion. However, there is a large gap between the promises made by “big data” advocates, who rely on statistics to discover patterns and trends in large volumes of non-structured data, and how historians can derive value from automatically generated metadata to explore archives and find answers to their research questions.

Currently, millions of scanned and OCRred files are available that hold the potential to significantly change the way historians of the construction and evolution of the European Union can carry out their research. However, due to a lack of resources, only minimal metadata are available at the file and document levels, severely undermining the accessibility of this archival collection.

This chapter explores the capabilities and limits of TM empirically in order to automatically extract key concepts from a large body of documents spanning multiple decades. By mapping the topics to headings from the EUROVOC thesaurus, this proof of concept offers the future opportunity to represent the topics thus identified with the help of a hierarchical search interface for end users.

Data collection and pre-processing

Based on a subset of the EC archives, this section presents how TM can be implemented and discusses the results. After signing a non-disclosure agreement (NDA), the MaSTIC research group³ of the Université libre de Bruxelles (The French-speaking Free University of Brussels) obtained a 138.3 GB corpus of 24,787 documents from the European Commission Archives. The dataset was created following the issuance of the Council Regulation (EEC, Euratom) No. 354/83 of 1 February 1983 concerning

opening the historical archives of the European Economic Community and the European Atomic Energy Community to the public. Classified documents within the files were declassified in conformity with Article 5 of the regulation. These files can be consulted by European citizens but are not currently made electronically available by the historical archives service of the EC, as little to no metadata are attached to the files.

The dataset is multilingual, spanning a period ranging from 1958 to 1982: it contains documents in French, Dutch, German, Italian, Danish, English and Greek, as those were the official languages of the European Economic Community of the time. As noted above, the dataset provides almost no metadata: apart from an XML file corresponding to each PDF, which contains basic information such as a unique identifier, a creation date, the number of a reference volume and the language and title of the document, little additional information is given. Although the metadata provides limited cues for classifying the documents, for example, the year of publication, none of these is related to the contents of the documents. There is no insight into what the documents encompass in terms of topics and themes, which makes the dataset nearly unusable for end users. This observation is reinforced by the fact that the original files contain several versions of the same document in different languages, making the automatic indexing of its content very difficult and effectively creating a lot of noise for the purposes of a classical, full-text information retrieval system.

To work with as many as 24,787 PDF files, a few preprocessing steps were needed. These steps, described below, include the creation of *.txt (raw text) files and the language detection of their content. Whilst PDF files are often the standard for storing historical documents and archives, the format does not allow for easy use within other, readily available applications. Using a small script in the programming language Python, a *.txt file was created for each of the existing PDF documents, making the dataset easily readable by other software. The script kept the existing folder structure as well as the filenames, ensuring that only the format of the data was changed. To create a file for each existing language version of a document, a Python script based on `langid.py` (Lui et al. 2012) was used, a language detection Python library that achieves 98.7% and 99.2% accuracy on the EuroGov (Sigurbjörnsson et al. 2005) EuroParl (Koehn 2005) corpora, respectively (which are two multilingual, parallel corpora that deal with EU-related matters). This process brought the total number of text files to 205,370, or 7.4 GB, an estimated 835,717,292 words or 1,671,434 pages.

Methodology

It should be noted that the number of topics for TM must be determined in advance. Following the work of Suominen et al. (2016), we used several configurations of the algorithm before choosing a total of 100 topics. LDA produces a list of keywords for each of the topics present in a textual dataset.

These keywords are supposedly the most representative tokens for that topic. When these are combined, a human operator must be able to deduce the underlying theme: for example, keywords such as *countries*, *cooperation*, *developing*, *trade*, *development*, *community*, *international*, *states*, *associated* and *aid* most probably refer to the topic of *international cooperation*. While this might often be considered straightforward, research shows that it is often more of an art than a science (Chang 2009), even though the automation of topic deduction (Lau 2014) seems possible. Nevertheless, deducing a list of top keywords selected by an algorithm remains a difficult task. This problem should thus not be taken lightly, especially in the case of large archival fonds whose content is not completely known. With that in mind, we resorted to matching the most prominent entries with EuroVoc terms, allowing us to base our work on a solid, well-documented foundation on the one hand and to harness the power of a hierarchical, multilingual thesaurus on the other. EuroVoc (<http://eurovoc.europa.eu/>) is the EU's multilingual thesaurus, which we use here as the gold standard for indexing documents. Even though LDA provides a *distribution* of topics for each document, i.e., a collection of topics, where a probability is assigned to each document indicating to what extent each topic occurs in that document, we resorted to hard partitioning in this case study, i.e., assigning a single topic to every document. This choice was made because hard partitioning is a proof-of-concept approach to a semi-automatic classification of historical archives (Suominen et al. 2016); the alternative soft-partitioning approach, in which several topics are assigned to a document, would have proven too time-consuming while achieving only a small refinement of the results.

Evaluation metrics for TM are an important consideration, focussing mostly on a quantitative assessment of topics based on coherence measures. An implementation of these measures can be found in the Palmetto Toolbox, a detailed discussion of which is beyond the scope of this chapter. In general, these measures are based on metrics that score the relatedness of the words for each topic. Although these measures are important for the interpretability of topics, and the assurance that the clusters found in the data are semantically realistic, they are not that important for our current purposes. They are discussed in more detail in our previous work (Coeckelbergs et al. 2020). Rather than verifying whether the clusters themselves reveal the semantics of the underlying document collection, which can be achieved using the Palmetto Toolbox, we seek to evaluate whether top terms from topics can be matched to strongly related concepts from the EuroVoc thesaurus. To allow for a more precise and qualitative evaluation, the results of which can be found in Table 12.1, we selected a subset of the whole archival fonds: for this proof-of-concept we used three subsets of the dataset, consisting of the three EU Commissions for which English texts were available: the Ortoli presidency (73–77), the Jenkins presidency (77–81) and the Thorn presidency (81–85, but our data stops at 82).

Table 12.1 Examples of the results: URIs, labels and tokens for topics

URI	Label	Tokens				
http://eurovoc.europa.eu/2965	agricultural aid	agricultural	areas	aid	measures	
http://eurovoc.europa.eu/852	ECSC aid	Coal	steel	ECSC	aid	
http://eurovoc.europa.eu/1418	textile industry	Fabrics	textile	woven	knitted	

Instead of manually querying the EuroVoc website, we used the built-in IMPORTXML (<https://support.google.com/docs/answer/3093342>) and IMPORTHTML (<https://support.google.com/docs/answer/3093339>) functions of Google Sheets. These functions allow us to query EuroVoc automatically and easily from within Google Sheets, where we had previously stored the output of TM; and, once the correct term has been selected, to keep its URI and preferred term (PT).

Results

Three examples of the results are illustrated in Table 12.1, each line corresponds to a topic, where the first and second column show the URI and the label of the topic, respectively, and the following columns give some of the tokens deemed representative of the topic by the algorithm. For clarity's sake, only the first four out of ten tokens are displayed.

However, it is important to emphasise that we were unable to attach a label to around 30% of the clusters, either due to the very general nature of the tokens (e.g., *agreement, community, parties, negotiations*) or to the fact that we did not manage to find a semantic link between them (e.g., *lights, bmw, brazil, eec, coffee*). For some topics, OCR noise (e.g., *cf, ii, ir*) was the main cause. While the OCR errors cannot be corrected automatically, the other unmatchable output could be reduced by using a smaller number of topics in the LDA configuration. In practice, we discovered that OCR errors tend to cluster into a single topic, which can then be left out of consideration.

The evaluation of the annotated LDA output was carried out by three different people, and the work of each annotator was verified by another. During this evaluation, there was no discrepancy between annotators based on Cohen's kappa coefficient, a well-known method of quantifying inter-annotator agreement. This indicates that the matching between LDA output and EuroVoc terms is consistent between people. We plan to submit our findings for expert evaluation to a domain expert, i.e., an archivist of the EC. Our approach differs from the one described by Newman (2010), which relied on semi-automatic evaluation of results using word pairs which serve as a base line for strongly correlated words (from Wikipedia, among other sources). Since the aim of this work is to evaluate how LDA can be used to

help annotate corpora with an existing controlled vocabulary and not to evaluate the human interpretation of LDA itself, our approach thus prevents an additional step which might introduce noise. Relying on an expert review helps in this process.

While there is a need for agreement between annotators for the controlled vocabulary matching, there is no indication of where LDA has correctly assigned topics to documents, only that it is possible to match LDA output to an existing thesaurus. With that in mind, we resorted to three topics selected randomly from each presidency and manually checking all documents for which this topic is the primary subject. Other means of evaluating LDA output exist, including the *topic intrusion* task introduced by Chang (2009): humans are given a document and four lists of words, each list amounting to a topic, and have to decide which one list out of the four is incorrect for that document. Given the clear and comprehensive results of a close manual inspection, such a method was not used. In the course of this close inspection of several hundred text files, no discrepancy between an actual document and its assigned topic could be found, even though it is clear that some documents are more relevant than others: this is the only logical and expected outcome, as LDA results in soft partitioning (a document is about several topics) and we considered only the primary topic of our documents.

From this dual approach to the manual evaluation of our results, it is clear that LDA offers a relatively fast and undeniably cheap alternative to manual metadata creation. Clear examples of success include the documents specified by LDA as part of the ECSC aid topic: the algorithm returned documents whose respective titles are “Memorandum on the financial aid awarding [sic] by the Member States to the coal industry in 1976”, “Introduction of a Community aid system for intra-community trade in power-station coal”, etc. After an extensive search of the results, the authors have failed to detect any document that was not directly or indirectly related to the topic deduced. Nonetheless, as noted earlier, around 30% of the clusters were not successfully matched with a label, mainly because of bad OCR.

Relevance of the research and future perspectives

In this chapter, we have discussed the results of applying TM on a large archival corpus to assess the potential of this statistical approach to the exploration of large collections of full text. The analysis yielded results scoring high in precision, but for which recall is unavailable.

Based on these results, what are the wider implications of this research for the overall archival community?

The tools and methodology described in this chapter can be of interest to any archival holder interested in adding a subject-based access to digitised full-text holdings. As document-level subject indexing is often out of scope, the automated and low-cost approach of TM can offer an interesting extra

access path for end-users. Typical use-cases would include digitised newspapers, journals or meeting reports, etc. The fact that the methodology is language independent offers also an important additional advantage, as it can be applied to archives in a multitude of languages.

What our methodology currently lacks is the ability to determine the depth of the term from a thesaurus to which an extracted term should be mapped, i.e., how precise a thesaurus term should be. Work in that direction is planned, enabling practitioners and end users alike to better visualise the documents collaboratively within the wider context of the whole thesaurus. By doing so, this research will help historians and archivists to develop a better understanding of how large volumes of full-text documents can be made more accessible.

Bibliography

- Bailey, David J. 2009. *The Political Economy of European Social Democracy: A Critical Realist Approach*. London: Taylor Francis Ltd.
- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993–1022.
- Boydens, Isabelle 1999. *Informatique, Normes et Temps*. Bruxelles: Bruylant (Emile).
- Boydens, Isabelle, and Seth Van Hooland. 2011. "Hermeneutics Applied to the Quality of Empirical Databases." *Journal of Documentation* 67 (2): 279–289.
- Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. 2009. "Reading Tea Leaves: How Humans Interpret Topic Models." In *Advances in Neural Information Processing Systems*, edited by Yoshua Bengio, D. Schuurmans, J. D. Lafferty, Christopher Kenneth Ingle Williams, Aron Culotta, 288–296. New York, NY: Curran Associates.
- Chiticariu, Laura, Yunyao Li, and Frederick R. Reiss. 2013. "Rule-based Information Extraction is Dead! Long Live Rule-based Information Extraction Systems!" In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, edited by David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard, 827–832. Seattle, WA: Association for Computational Linguistics.
- Coeckelbergs, Mathias, and Seth Van Hooland. 2020. "Concepts in Topics. Using Word Embeddings to Leverage the Outcomes of Topic Modeling for the Exploration of Digitized Archival Collections." In *Proceedings of the first EAI Data and Information in Online Environments (DIONE) Conference*, edited by Rogério Mugnaini, 41–52. Cham: Springer.
- Desrochers, Donna M., and Jane V. Wellman. 2011. *Trends in College Spending 1999–2009*. Washington, DC: Delta Cost Project. https://deltacostproject.org/sites/default/files/products/Trends2011_Final_090711.pdf
- Duranti, Luciana, 1994. "The Concept of Appraisal and Archival Theory." *American Archivist* 57, 328–344.
- Goldstone, Andrew, and Ted Underwood. 2012. "What Can Topic Models of Pmla Teach Us about the History of Literary Scholarship." *Journal of Digital Humanities* 2 (1): 40–49.
- Hofmann, Thomas. 1999. "Probabilistic Latent Semantic Indexing." In *Proceedings of the 22nd Annual International SIGIR Conference on Research and Development*

- in *Information Retrieval*, edited by Fredric Gey, Marti Hearst, and Richard Tong, 50–57. New York, NY: Association for Computing Machinery.
- Kelemen, Jozef. 2007. “From Artificial Neural Networks to Emotion Machines with Marvin Minsky.” *Acta Polytechnica Hungarica* 4 (4): 5–16.
- Kerr, Sara J. 2016. “Jane Austen in Vector Space: Applying Vector Space models to 19th century literature.” In *Proceedings of the Japanese Association for Digital Humanities 2016 Conference*, 19–22. Tokyo: The Historiographical Institute.
- Klein, Lauren F., Jacob Eisenstein, and Iris Sun. 2015. “Exploratory Thematic Analysis for Digitized Archival Collections.” *Digital Scholarship in the Humanities* 30 (suppl 1): i130–i141.
- Koehn, Philipp. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. MT Summit, 79–86. <https://homepages.inf.ed.ac.uk/pkoehn/publications/europarl-mtsummit05.pdf>
- Lau, Jey Han, David Newman, and Timothy Baldwin. 2014. “Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality.” In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 530–539. Stroudsburg: The Association for Computational Linguistics.
- Lohr, Steve. 2016. “IBM Is Counting on Its Bet on Watson, and Paying Big Money for It” New York Times, October 17, 2016. <https://www.nytimes.com/2016/10/17/technology/ibm-is-counting-on-its-bet-on-watson-and-paying-big-money-for-it.html>
- Lui, Marco and Timothy Baldwin. 2012. “langid.py: An Off-the-shelf Language Identification Tool.” In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012), Demo session*, edited by Min Zhang, 25–30. Stroudsburg: The Association for Computational Linguistics.
- Mack, Y.P., and Murray Rosenblatt. 1979. “Multivariate k-Nearest Neighbor Density Estimates.” *Journal of Multivariate Analysis* 9 (1): 1–15.
- Moretti, Franco. 2007. *Graphs, Maps, Trees: Abstract Models for a Literary History*. London: Verso.
- National Archives and Records Administration. 2014. “Automated Electronic Records Management Report/Plan.” September 2014. <https://www.archives.gov/files/records-mgmt/prmd/A31report-9-19-14.pdf>
- National Archives and Records Administration. 2015. “White Paper on the Capstone Approach and Capstone GRS.” April 2015. <https://www.archives.gov/files/records-mgmt/email-management/final-capstone-white-paper.pdf>
- Newman, David, Youn Noh, Edmund M. Talley, Sarvnaz Karimi, and Timothy Baldwin. 2010. “Evaluating Topic Models for Digital Libraries.” In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, 215–224. New York, NY: Association for Computing Machinery.
- Roe, Glenn, Clovis Gladstone, and Robert Morrissey. 2016. “Discourses and Discipline in the Enlightenment: Topic Modeling the French Encyclopédie.” *Frontiers in Digital Humanities* 2: 2297–2668.
- Rosenblatt, Frank 1958. “The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain.” *Psychological Review* 65 (6): 386–408.
- Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. 1986. “Learning Representations by Back-Propagating Errors.” *Nature* 323: 533–536.
- Schmitz Mausam, Michael, Robert Bart, Stephen Soderland, and Oren Etzioni. 2012. “Open Language Learning for Information Extraction.” In *Proceedings of*

- the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL), edited by Jun'ichi Tsujii, James Henderson, and Marius Paşca, 523–534. Stroudsburg: The Association for Computational Linguistics.
- Sigurbjörnsson, Börkur Jaap Kamps, and Maarten de Rijke. 2005. “EuroGOV: Engineering a Multilingual Web Corpus.” In *Accessing Multilingual Information Repositories, Workshop of Cross-Language Evaluation Forum for European Languages (CLEF)*, Lecture Notes in Computer Science, vol 4022, edited by Carol Peters et al., 825–836. Berlin, Heidelberg: Springer. https://doi.org/10.1007/11878773_90
- Suominen, Arho, and Hannes Toivanen. 2016. “Map of Science with Topic Modeling: Comparison of Unsupervised Learning and Human-Assigned Subject Classification.” *Journal of the Association for Information Science and Technology (JASIST)* 67 (10): 2464–2476.
- Tin-Kam, Ho. 1995. “Random Decision Forests.” In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 278–282. Montreal, QC: Institute of Electrical and Electronics Engineers.
- Van Hooland, Seth, Max De Wilde, Ruben Verborgh, Thomas Steiner, and Rik Van de Walle. 2015. “Exploring Entity Recognition and Disambiguation for Cultural Heritage Collections.” *Digital Scholarship in the Humanities* 30 (2): 262–279.
- Vellino, André, and Inge Alberts. 2016. “Assisting the Appraisal of e-mail Records with Automatic Classification.” *Records Management Journal* 26: 293–313.
- Villena-Román, Julio, Sonia Collada-Pérez, Sara Lana-Serrano, and José C. González Cristóbal. 2011. “Hybrid Approach Combining Machine Learning and a Rule-Based Expert System for Text Categorization.” In *Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, 323–328. Palm Beach, FL: Association for the Advancement of Artificial Intelligence.
- Yao, Limin, Aria Haghighi, Sebastian Riedel, and Andrew McCallum. 2011. “Structured Relation Discovery using Generative Models.” In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1456–1466. Edinburgh: Association for Computational Linguistics.