Exploring archives with probabilistic models: Topic Modelling for the valorisation of digitised archives of the European Commission

Simon Hengchen, Mathias Coeckelbergs, Seth van Hooland
Université libre de Bruxelles
Brussels, Belgium
shengche;mcoeckel;svhoolan@ulb.ac.be

Ruben Verborgh
Ghent University – iMinds
Ghent, Belgium
ruben.verborgh@ugent.be

Thomas Steiner
Google Germany
Germany
tomac@google.com

***Abstract*—Topic Modelling (TM) has gained momentum over the last few years within the humanities to analyze topics represented in large volumes of full text. This paper proposes an experiment with the usage of TM based on a large subset of digitized archival holdings of the European Commission (EC). Currently, millions of scanned and OCRed files are available and hold the potential to significantly change the way historians of the construction and evolution of the European Union can perform their research. However, due to a lack of resources, only minimal metadata are available on a file and document level, seriously undermining the accessibility of this archival collection. The article explores in an empirical manner the possibilities and limits of TM to automatically extract key concepts from a large body of documents spanning multiple decades. By mapping the topics to headings of the EUROVOC thesaurus, the proof of concept described in this paper offers the future possibility to represent the identified topics with the help of a hierarchical search interface for end-users.**

***Keywords*-LDA; topic modelling; archives; topic modeling**

I. Introduction

When and how did environmental considerations start to influence the agricultural policy development from the European Commission (EC)? What are the key documents to analyse the debate on nuclear energy production from the 1960s onwards? These are two examples of typical research questions historians might have in mind. The mass digitisation of the EC's archives offer in this context new exciting possibilities to query and analyse in an automated manner the archival corpus. 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 actually derive value from automatically generated metadata to explore archives and find answers to their research questions.

Within the Digital Humanities (DH) community, Topic Modelling (TM) has attracted a fair amount of interest and is increasingly being used to access and explore large corpora of full-text documents (Klein et al., 2015; Chang et al., 2009; Goldstone and Underwood, 2012). Topic modelling – in our case, latent Dirichlet allocation, LDA (Blei et al., 2003) – is an unsupervised machine learning technique. Applied to large textual datasets, as introduced by Hofmann (1999), it shows great promise at successfully clustering similar texts. This approach, along with other text-mining routines, has gained momentum for document classification, as pointed out by Suominen and Toivanen (2015) in the specific field of bibliometrics or by Newman et al. (2010) in a library context. Similarily, Roe et al. (2016) uses LDA in order to draw a map of all human knowledge – as seen by d'Alembert and Diderot – contained in the French *Encyclopédie*. In archival science, computational methods are not new (Hedstrom, 1993) and different solutions for text categorisation have been tested, as pointed out by Díaz et al. (2004) and Frank and Paynter (2004), who used support vector machines (SVM).

By using a real-life case study of archives having undergone optical character recognition (OCR), this paper wants to critically assess the potential of TM for the archival community to experiment with "distant reading". Developed by Moretti (2005), distant reading practices make use of statistics and computational linguistics to automatically extract specific features from large corpora, allowing to spot trends and shifts over time. Traditionally, historians explore archives based on an inventory, which contains metadata on a fonds, series or file level. Only very rarely 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 OCRed text files, having only minimal "tombstone" metadata on either a fonds, series, file or document level. In the absence of traditional access paths, 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. Other approaches, such as Named-Entity Recognition (NER) (van Hooland et al., 2015) or Word2vec (Kerr, 2016) also offer opportunities to automatically extract specific features from full-text. Future experiments will be rolled out on the same corpus in order to compare the possibilities and limits of TM, NER and Word2vec.

978-1-4673-9005-7/16/$31.00 ©2016 IEEE

II. Methodology

Based on a statistically significant subset from the EC archives, this paper presents how TM can be applied and discusses the results. After the signature of a Non-Disclosure Agreement (NDA), the MaSTIC research group[1] of the Université libre de Bruxelles obtained a 138.3-GB, 24,787-document corpus from the European Commission Archives. The dataset has been created following the Council Regulation (EEC, Euratom) No 354/83 of 1 February 1983 concerning the opening to the public of the historical archives of the European Economic Community and the European Atomic Energy Community.[2] Classified documents in the files have been declassified in conformity with Article 5 of the aforementioned regulation. The files can be consulted by citizens, but are currently not made electronically available by the historical archives of the EC, as little to no metadata are attached to the files.

The dataset, spanning a period ranging from 1958 to 1982, is multilingual: it contains documents in French, Dutch, German, Italian, Danish, English and Greek, as those were the then official languages of the now-called European Union. As was already mentioned, the dataset presents close to no metadata: apart from an XML file corresponding to each PDF and containing basic information such as a unique identifier, a creation date, the number of a reference volume and the language and title of the document, few additional information is given. There is no insight as to what the documents encompass in terms of topics and themes, which makes the dataset nearly unusable for end-users. This conclusion is reinforced by the fact that the original files contain several linguistic versions of the same document, making the indexing of its content very hard and effectively inducing a lot of noise in the case of a classical, full-text information retrieval system.

In order to work with the 24,787 PDF files, a few preprocessing steps were needed. These steps, described below, include the creation of `*.txt` 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 really allow for an easy use within other, readily available applications. Using a small python script, a `*.txt` file for each of the existing PDF documents was made, 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 changed. If `*.txt` files are easier to work with, files consisting of the same content in different languages are close to useless in most existing applications. In order to create a file for each existing linguistic version of a document, a python script was used, based on langid.py (Lui and Baldwin, 2012), a language-detection Python library that achieves 98.7% and 99.2% accuracy on EuroGov and EuroParl – two multilingual, parallel corpora which deal with EU-related matters –, respectively. This process brings the total number of text files to 205,370, in 7.4 GB – an estimated 835,717,292 words or 1,671,434 pages.

LDA produces, for each of the topics[3] present in a textual dataset, a list of keywords. These keywords are supposedly the most representative tokens of that topic – combined together, a human operator must deduce the underlying theme: for example, it can be inferred from keywords `countries cooperation developing trade development community international states associated aid` that this collection of documents is about the topic of *international cooperation*. Whilst this might be considered straightforward, research shows that it is often more of an art than a science (Chang et al., 2009), even though the automation of topic deducing (Lau et al., 2014) seems possible. This task 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 tokens with EuroVoc[4] 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. Despite the fact that LDA provides a *distribution* of topics for each document, we resorted, in this case study, to hard-partitioning (Suominen and Toivanen, 2015): as this is a proof-of-concept approach to a semi-automatic classification of historical archives, soft-partitioning – associating several topics to a document – would have proven too time-consuming for too low a gain. Indeed, it can be argued that if topic models do detect the main topic of a document, as it is the case in our study, subsequent less-important topics can be *assumed* to be correct.

In addition, in order to allow for a more precise and qualitative evaluation of our method, 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; our data stops at 82). The total number of text files used for the manual evaluation is 11,868.

Instead of manually querying the EuroVoc website,

[1] http://mastic.ulb.ac.be

[2] The legal text and all its amendments are available at http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1475395564392\&uri=CELEX\:31983R0354

[3] It should be noted that the number of topics must be determined beforehand. Following the work of Suominen and Toivanen (2015) and others, we have used several configurations of the algorithm before choosing for a total of 100 topics.

[4] EuroVoc, http://eurovoc.europa.eu/ is the EU's multilingual thesaurus.

the built-in `IMPORTXML` and `IMPORTHTML`[5] functions of Google Sheets were used. These functions allow us to automatically and easily query EuroVoc from within Google Sheets – where we had previously stored the output of topic modelling – and, once the correct term has been selected, to keep its URI and preferred term (PT).

III. Results & Discussion

Three examples of the results are illustrated in Table I: each line corresponds to a topic, where the first and second columns depict respectively the URI and the label of the topic, and the following columns some of the token deemed representative of the topic by the algorithm.

However, it is important to underline that we were unable to attach a label to around 30% of the clusters, due to either the very general nature of the tokens (`agreement community parties negotiations`) or the fact that we did not manage to find a semantic link between them (`lights bmw brazil eec coffee`). For some topics, OCR noise (`cf ii ir`) was the main cause. Whilst 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.

The evaluation of the annotated LDA output has been carried out by three different people, and the work of each annotator has been verified by another. During this evaluation, there was no discrepancy between annotators, indicating that the matching between LDA output and EuroVoc terms is consistent across people. Submitting our findings to a domain expert – i.e., an archivist of the EC – for expert evaluation is planned. Our approach differs from the one described in Newman et al. (2010), which relied on a semi-automatic evaluation of results using word-pairs (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 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.

Whilst an agreement between annotators for the controlled-vocabulary matching is needed, it does not indicate whereas LDA 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 select three topics out of each presidency, and to manually check all documents that have this topic as primary subject matter. Other means of evaluating LDA output exist, including the *topic intrusion* task introduced by Chang et al. (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 thorough results of a close manual inspection, such a method has not been used. During 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 only logical and could be expected, as LDA produces soft-partitioning (a document is about several topics) and we only considered the primary topic for our documents.

From this double-sided, 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 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 extensive searching into the results, the authors have failed to detect a document that was not directly or indirectly related to the deduced topic. Nonetheless, as indicated above, it should be reported that around 30% of the clusters were not successfully matched with a label: reasons include bad OCR and an incorrect number of topics specified when running LDA.

IV. Perspectives

In this paper, we have discussed the results of applying TM on a large archival corpus in order to assess the potential of this statistical approach towards the exploration of large collections of full text – an analysis that yielded results scoring high in precision, but for which recall is unavailable. The approach is language-independent and can thus be applied on archives in a multitude of languages. What our methodology currently lacks is the ability to determine the depth of the extracted term – how precise should a term of the thesaurus be used –, but work in that direction is planned, enabling practitioners and end-users alike to better visualise the documents between each other and in the bigger context of the whole thesaurus. Also, as was mentioned in the introduction, other methods such as NER and Word2vec will be applied on the same corpus, in order to analyse the possibilities and limits of TM compared to other approaches of automatically creating access paths across a large archival corpus. By doing so, this research can help historians and archivists to develop a better understanding of how large volumes of full-text documents can be made more accessible.

[5]Documentation for these functions is available at https://support.google.com/docs/answer/3093342 and https://support.google.com/docs/answer/3093339, respectively.

References

S. van Hooland, M. De Wilde, R. Verborgh, T. Steiner, and R. Van de Walle, "Exploring entity recognition and disambiguation for cultural heritage collections," *Digital Scholarship in the Humanities*, vol. 30, no. 2, pp. 262–279, 2015.

F. Moretti, *Graphs, maps, trees: abstract models for a literary history*. Verso, 2005.

A. Goldstone and T. Underwood, "What can topic models of pmla teach us about the history of literary scholarship," *Journal of Digital Humanities*, vol. 2, no. 1, pp. 40–49, 2012.

L. F. Klein, J. Eisenstein, and I. Sun, "Exploratory thematic analysis for digitized archival collections," *Digital Scholarship in the Humanities*, p. fqv052, 2015.

M. Lui and T. Baldwin, "langid. py: An off-the-shelf language identification tool," in *Proceedings of the ACL 2012 system demonstrations*. Association for Computational Linguistics, 2012, pp. 25–30.

A. Suominen and H. Toivanen, "Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification," *Journal of the Association for Information Science and Technology*, 2015.

D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1999, pp. 50–57.

D. Newman, Y. Noh, E. Talley, S. Karimi, and T. Baldwin, "Evaluating topic models for digital libraries," in *Proceedings of the 10th annual joint conference on Digital libraries*. ACM, 2010, pp. 215–224.

J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei, "Reading tea leaves: How humans interpret topic models," in *Advances in neural information processing systems*, 2009, pp. 288–296.

J. H. Lau, D. Newman, and T. Baldwin, "Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality." in *EACL*, 2014, pp. 530–539.

G. Roe, C. Gladstone, and R. Morrissey, "Discourses and disciplines in the enlightenment: Topic modeling the french encyclopédie," *Frontiers in Digital Humanities*, vol. 2, p. 8, 2016.

D. Diderot *et al.*, *Encyclopédie: ou Dictionnaire raisonné des sciences, des arts et des métiers*, 1751.

S. Kerr, "Jane Austen in vector space: Applying vector space models to 19th century literature," in *Proceedings of the JADH 2016 Conference*, 2016, pp. 19–22.

E. Frank and G. W. Paynter, "Predicting Library of Congress classifications from Library of Congress subject headings," *Journal of the American Society for Information Science and Technology*, vol. 55, no. 3, pp. 214–227, 2004.

I. Díaz, J. Ranilla, E. Montañes, J. Fernández, and E. F. Combarro, "Improving performance of text categorization by combining filtering and support vector machines," *Journal of the American society for information science and technology*, vol. 55, no. 7, pp. 579–592, 2004.

M. Hedstrom, "Teaching archivists about electronic records and automated techniques: A needs assessment," *The American Archivist*, vol. 56, no. 3, pp. 424–433, 1993.

Table I
MATCHING OF EUROVOC TERMS WITH LDA OUTPUT

URI	label	tokens			
http://eurovoc.europa.eu/2965	agricultural aid	agricultural	areas	aid	measures
		premium	directive	number	eec
		farms	production		
http://eurovoc.europa.eu/852	ECSC aid	coal	steel	ecsc	aid
		industry	production	iron	decision
		measures	community		
http://eurovoc.europa.eu/1418	textile industry	fabrics	textile	woven	knitted
		crocheted	fibres	community	agreement
		products	yarn		