HistorEx: Exploring Historical Text Corpora Using Word and Document Embeddings

Sven Müller[1a], Michael Brunzel[1b], Daniela Kaun[1c], Russa Biswas[1,2], Maria Koutraki[3], Tabea Tietz[1,2], and Harald Sack[1,2]

[1] Karlsruhe Institute of Technology, Institute AIFB, Germany
[a]uodlt@student.kit.edu , [b]ubebi@student.kit.edu , [c]ugebn@student.kit.edu
[2] FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany
firstname.lastname@fiz-karlsruhe.de
[3] Leibniz University Hannover, L3S Research Center, Germany
koutraki@l3s.de

Abstract. Written text can be understood as a means to acquire insights into the nature of past and present cultures and societies. Numerous projects have been devoted to digitizing and publishing historical textual documents in digital libraries which scientists can utilize as valuable resources for research. However, the extent of textual data available exceeds humans' abilities to explore the data efficiently. In this paper, a framework is presented which combines unsupervised machine learning techniques and natural language processing on the example of historical text documents on the $19th$ century of the USA. Named entities are extracted from semi-structured text, which is enriched with complementary information from Wikidata. Word embeddings are leveraged to enable further analysis of the text corpus, which is visualized in a web-based application.

Keywords: Word Embeddings · Document Vectors · Wikidata · Cultural Heritage · Visualization · Recommender System

1 Introduction

Humanities as well as social sciences, historical sciences and political sciences are essentially text-based sciences, i.e. philologies, and hence strongly depend on the diligent analysis of text corpora. Cultural heritage text data grants researchers from these text-based sciences precious insights into past and present cultures and social structures. However, the extent of information available in the form of unstructured text provides a difficult challenge for researchers to fully grasp the content of the documents in a cultural heritage data collection, the interrelations between the documents as well as the places and figures they involve in a reasonable amount of time and work effort. As a result, the meaning of these data for the societies they describe often remains uncharted. This problem calls for interdisciplinary work between technical and non-technical researchers with the goal to process these documents in a way that they can be explored, understood, and placed into the context of the time and culture they were authored in

[1]. To achieve this, the data has to be analyzed and enriched with machine understandable information before it is presented to the user in a comprehensible interface for exploration [2].

The presented work also addresses this challenge and contributes a framework that combines unsupervised Machine Learning techniques with Natural Language Processing (NLP) with the goal to uncover semantic connections among collections of textual documents on the example of a historical text corpus on the $19th$ century USA[4]. The data was originally published by the Humboldt Chair for Digital Humanities of the University of Leipzig and made available by the Coding da V1nc1 hackathon[5]. In total the collection comprises 334 documents with various types of literary works including local reports on the American Civil War to biographies and novels. As part of the presented framework, named entities are extracted and partially enriched with additional information from the Wikidata knowledge graph. Furthermore semantic representations of the documents are obtained using word vectors and document vectors. The framework also includes an interface that exploits the results from the proposed approach and enables users on the web to explore the text collection.

2 Description of HistorEx

The proposed workflow, as illustrated in Fig. 1, follows two separate pipelines: ① Entity extraction is performed using Beautiful Soup[6]. The persons and locations mentioned in the text are directly imported into Dash[7]. The extracted authors are enriched with additional information from the Wikidata knowledge base. ② After preprocessing, semantic representations of each document are obtained by a neural net approach. In the final step, all data from both the pipelines is integrated into an interactive web application for data exploration.

2.1 Vector Representation

The semantic representations of the documents, i.e. the word and document vectors are generated using the Distributed Memory Model of Paragraph Vectors (PV-DM) [3]. The word and the document vectors are initialized randomly. The document vector is unique to a document, whereas the word vectors are shared among all the documents. Therefore, the word vectors learned from all the documents in the collection. Noise-contrastive estimation (NCE) is used for optimization of the neural network. The quality of the trained representations is assessed during the training by manually monitoring the most-similar words to a corresponding set of selected tokens which leads to a training for eight epochs with a batch size of 512 and a context window-size of six (three on each side of the target-token).

[4] `http://www.perseus.tufts.edu/hopper/collection?collection=Perseus:collection:cwar`
[5] `https://codingdavinci.de/about/`, last retrieved: March 05, 2019
[6] `https://www.crummy.com/software/BeautifulSoup/bs4/doc/`
[7] `https://plot.ly/products/dash/`

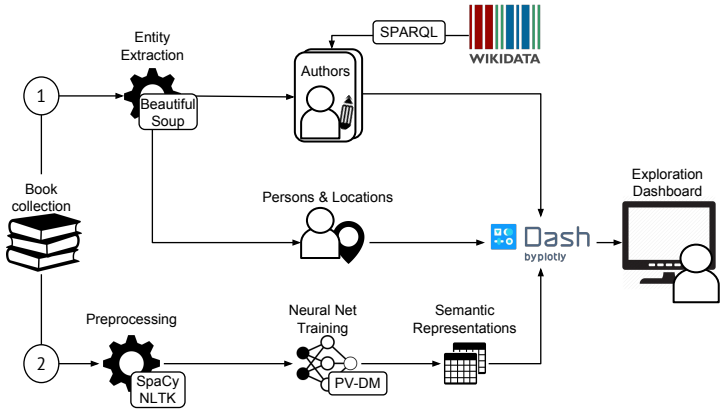


Fig. 1: Structure of the proposed double tracked pipeline.

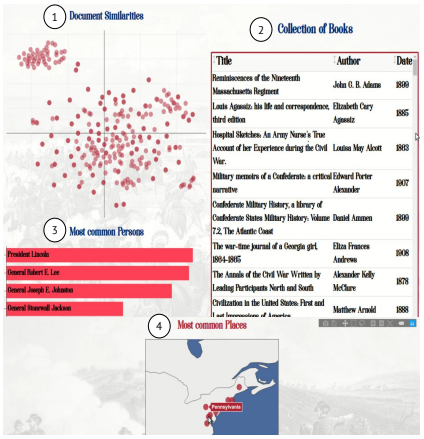
2.2 Data Exploration Dashboard

The results of the proposed approach are presented in an interactive dashboard, depicted in Fig. 2a and 2b. It provides the user with several entry points to explore the text collection visually. It was implemented with Dash. The exploration dashboard and its user interaction possibilities are described in the Demo (section 3) of this paper.

3 Demo

The HistorEx dashboard provides the user with several entry points to the content of the collection. Fig. 2a presents an initial overview of the entire collection. A T-SNE plot ① lets the user explore the similarity between the documents in the corpus through visualized clusters. In the presented plot, each dot represents one book. Here, the user can also zoom in to the graph to get a closer look on the data and hovering each dot reveals its title. A table ② shows the collection of books overall and a bar chart let's the user explore the most common persons mentioned in the collection ③. Hovering individual bars reveals their quantity of occurrence. The places mentioned in the texts are visualized in a map view ④ which are plotted using GeoPy. Next to the general overview, the user can systematically explore the content by searching for specific entities or documents using the search bar ⑤ implemented with an autosuggestion feature. Fig. 2b shows the result of a book search. Next to its title, the author information ⑥, automatically retrieved from Wikidata, is given along with a link to the original Wikidata page. The user can furthermore explore persons occurring in the book ⑦ and receive document recommendations based on similarity ⑧. A map view (not pictured), similarly to ④ presented in Fig. 2a reveals the places named in the explored document. The HistorEx demo is publicly available[8].

[8] `https://ise-fizkarlsruhe.github.io/CourseProjects2019#historex`



(a) Dashboard: Data collection overview with ① T-SNE plot of document similarities, an overview of ② books in the collection, ③ mentioned persons, ④ mentioned places. Due to space efficiency, the screenshots only show part of the information given in the dashboard.

(b) Book exploration with ⑤ a search bar ⑥ book author information, ⑦ common persons bar chart, ⑧ similar documents.

Fig. 2: HistorEx Demo Interface

4 Experiments and Results

In this section, the experiments conducted on 308 historical books which amounts to roughly 30 million tokens are described. First, the word vectors generated by PV-DM method are evaluated against SimLex-999[9]. Secondly, the document vectors generated by PV-DM are compared against the document vectors generated by averaging the Google pre-trained word vectors. Thirdly, most similar top-k documents are retrieved based on cosine similarity for both types of document vectors. Also, both types of document vectors are projected to a low dimensional space to gain better insight of the similar documents using t-Stochastic Neighbor Embedding (t-SNE) [10].

It has been observed that these two different types of document vectors exhibit different orientation in the low dimensional space and also in top-k recommendation of most similar documents. This is due to the usage of different set of vocabularies in Wikipedia and historical text. Moreover, Wikipedia is an open encyclopedia consisting of information from various domains whereas historical data is restricted to a specific geographical location in a certain era. A few results from the experiments and the t-SNE plot have been provided at Github[11]

[9] https://fh295.github.io/simlex.html
[10] https://lvdmaaten.github.io/tsne/
[11] https://github.com/ISE-FIZKarlsruhe/HistorEx

5 Conclusion

In this paper, HistorEx, a framework to analyze and explore semi-structured historical text collections on the example of a historical text corpus on the $19th$ century of the USA was presented. HistorEx includes the extraction of named entities from the text corpus and enrichment of document authors with knowledge from Wikidata. Furthermore, semantic representations of the documents are obtained using word vectors and document vectors. The results are integrated into an interactive dashboard to facilitate document exploration.

Experiments to assess the quality of the word and document vectors show that the application of the PV-DM model provides promising results on the example text collection. Furthermore, evidence is provided that the approach seems especially useful for the analysis of unstructured historical English text collections, as long as the amount of text is sufficiently large in order to train meaningful semantic representations with the neural net. Future work will focus on enriching not only document authors with additional information from the Wikidata knowledge graph but also persons and locations mentioned in the documents to improve the exploration environment of the framework and enrich the collection with further context.

Acknowledgement. This paper is motivated by the the first German open cultural data hackathon, Coding da V1nc1. It supports interdisciplinary work on cultural heritage data by bringing together GLAM institutions, programmers and designers to develop ideas and prototypes for the cultural sector and for the public.

References

1. Gold, M. K. *Debates in the Digital Humanities*. University of Minnesota Press, 2012.
2. Jänicke, S., Franzini, G., Cheema, M. F., and Scheuermann, G. Visual text analysis in digital humanities. In *Computer Graphics Forum* (2017), vol. 36, Wiley Online Library, pp. 226–250.
3. Le, Q. V., and Mikolov, T. Distributed representations of sentences and documents. *CoRR abs/1405.4053* (2014).