IJDC | *Conference Paper*

An Approach for Curating Collections of Historical Documents with the Use of Topic Detection Technologies

Medina Andresel
AIT Austrian Institute of Technology, Vienna, Austria

Sergiu Gordea
AIT Austrian Institute of Technology, Vienna, Austria

Srdjan Stevanetic
AIT Austrian Institute of Technology, Vienna, Austria

Mina Schütz
AIT Austrian Institute of Technology, Vienna, Austria

Abstract

Digital curation of materials available in large online repositories is required to enable the reuse of Cultural Heritage resources in specific activities like education or scientific research. The digitization of such valuable objects is an important task for making them accessible through digital platforms such as Europeana, therefore ensuring the success of transcription campaigns via the Transcribathon platform is highly important for this goal. Based on impact assessment results, people are more engaged in the transcription process if the content is oriented to specific themes, such as the First World War. Currently, efforts to group related documents into thematic collections are in general hand-crafted and due to the large ingestion of new material they are difficult to maintain and update. The current solutions based on text retrieval are not able to support the discovery of related content since the existing collections are multi-lingual and contain heterogeneous items like postcards, letters, journals, photographs etc. Technological advances in natural language understanding and in data management have led to the automation of document categorization via automatic topic detection. To use existing topic detection technologies on Europeana collections there are several challenges to be addressed: (1) ensure representative and qualitative training data, (2) ensure the quality of the learned topics, and (3) efficient and scalable solutions for searching related content based on the automatically detected topics, and for suggesting the most relevant topics on new items. This paper describes such challenges and the proposed solutions in more detail, thus offering a novel perspective on how digital curation practices can be enhanced with the help of machine learning technologies.

Submitted 13 June 2022 ~ *Accepted* 13 June 2022

Correspondence should be addressed to Sergiu Gordea Email: sergiu.gordea@ait.ac.at

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

The digitization of cultural heritage (CH) objects, and their aggregation in large publicly accessible repositories, stimulate their reuse in other sectors including education or scientific research. Europeana, Europe's digital platform for cultural heritage, aggregates 50+ million CH objects from all European countries. [1] The descriptions of Europeana records are freely accessible and a large part of the digitised content is open for reuse. However, effective access to the information available in historical documents is not yet achieved through the simple scanning of old manuscripts (i.e. due to illegible hand writing, changed alphabets or scripts, damaged documents, etc). The extraction of textual information available of scanned documents into a plain text representation (i.e. which is easily understandable both by humans and machines) is a key activity for enabling its reuse by the communities of practice. Such activities are supported by online transcription tools like Transcribathon.[2] This tool implements support for transcribing manuscripts available in the Europeana repository. Through the engagement of the user community in crowdsourcing campaigns, many historical documents were worked on and barrier free access to historical information is provided on the Europeana site.

However, stimulating user participation in crowdsourcing campaigns is still a challenging task. The impact assessment report[3] on Transcribathon campaigns indicates that the best engagement is achieved by organising competitions on specific themes or topics which may raise the interest of individual users or local communities. Past crowdsourcing campaigns were focusing on transcribing manuscripts related to historical events such as the First World War or the Revolution from 1989, while the current efforts go towards unlocking the information related to the urban and societal development from "the long 19th Century", starting with the French Revolution until the First World War.

The informational content from the selected documents plays an important role for the success of Transcribathon runs. However, the manual curation of very specific collections from Europeana remains a challenging and expensive task given the size of the repository, the multilingualism, and the heterogeneity of the data records.

In this paper we present the current efforts for providing automated support for content curation activities in the Transcribathon platform by employing Natural Language Processing (NLP) and Machine Learning (ML) technologies. The NLP technologies are employed within the pre-processing pipeline to aggregate appropriate input for learning a topic detection model. For the first experiments we used a non-supervised solution for topic detection based on Latent Dirichlet Allocation (LDA). Within this process, we select the most relevant terms for each topic model, which are further used for searching new materials that can be associated to each individual topic.

As presented in the conclusions of the paper, topic detection solutions can be successfully employed for curating collections of historical documents when the following challenges are addressed: (1) ensure representative and qualitative training data, (2) identify the most suited topic detection models and (3) implement efficient and scalable solutions for searching related content based on the automatically detected topics. The technical solutions used to address these individual challenges are presented in the following sections.

Even if the topic detection technology is language agnostic, the learned models depend on the input data, which in the case of the Europeana repository includes descriptions available in one of the official European languages, but also Hebrew, Russian and Norwegian. By employing automatic language detection and machine translation, the proposed approach is able to

[1] https://www.europeana.eu
[2] https://europeana.transcribathon.eu/
[3] Impact Assessment Report: https://pro.europeana.eu/post/impact-assessment-report-enrich-europeana-transcribathon

automatically group Europeana records in clusters of similar content, independent from the language of the original object descriptions.

The following are the main research questions behind the experimental evaluation and the conclusions presented within this paper:

1. Which are the most appropriate models used for automatic clustering of Transcribathon documents in topics?

2. Can the topic-based search be effectively and efficiently used to support curating collections of historical documents?

The rest of the paper is organized as follows: firstly, we give an overview of the related work and state-of-the-art approaches for topic modeling. Secondly, we present the methodology of our approach - content curation using topic-based information retrieval – and its pre-processing pipeline, learning a topic model and the topic-based information retrieval technique. Next the experimental evaluation is presented, followed by the results discussions, and the conclusions and future work.

Related Work

Topic Modelling, more commonly known as topic detection, is a natural language processing task, which is used in many domains, such as social media analysis and information retrieval. The goal of this task is to find semantically important patterns and new information about the underlying text data (Jelodar, 2019). The retrieved topics give an overview over the main themes in a document based on the words in each text (George, 2018). However, topic modelling approaches are usually sensitive to noise and therefore not stable and they must be optimized and iterated until the best model is achieved (Vayansky, 2020). To identify the best model, there are several metrics proposed in the literature, however, in general, human input might still be required to assess the quality of the obtained topics.

There are various approaches to find relevant topics in a corpus of documents. *Latent Semantic Analysis* (LSA) (Dumais, 2005) is a technique where words co-occurrences are used to derive concepts. Those concepts inherit terms that have a similar semantic meaning, based on the occurrences in the documents. Among the extended versions of LSA is the *Probabilistic Latent Semantic Analysis* method (PLSA) (Hofmann, 1999) which also considers the context of the words. This solution is useful when the documents contain ambiguous terms.

Probably the most popular approach for topic modelling is the *Latent Dirichlet Allocation* (LDA) (Blei, 2003). The standard assumption is that each document can consist of multiple topics, and the topics are generated as a probability distribution over all the words in the corpus. There are various adjusted and enhanced versions of LDA. For instance, *Hierarchical Latent Dirichlet Allocation* (hLDA) treats the topics as a hierarchy and generates subtopics (George, 2018). Other variants of LDA include Latent Dirichlet Mixture Model (LDMM) and matrix factorization through LDA (fLDA) (Vayansky, 2020). Another extension to LDA is Correlated Topic Modelling (CTM) (Blei, 2005), which is a statistical model that captures correlations and relationships in the extracted topics through a logistic normal distribution which is especially useful for exploration, predictions, and filtering. An alternative to LDA-based approaches is proposed in (Moody, 2016) that tries to learn the parameters for the Dirichlet distribution of topics to documents and words to topics. In this model, word embeddings are used to consider the semantic relations between words co-occurring in the corpus, in various contexts and the approach builds on the existing *word2vec* model (Mikolov, 2013) which is typically used for unravelling syntactic and semantic similarities in language from large and unsupervised sets of documents, like Wikipedia.

To measure the quality of the obtained topic model, and to identify the optimal number of topics discovered, metrics such as perplexity, coherence and various clustering coefficients are

being used. In this work we consider two metrics, described below, that complement each other: coherence (Syed, 2017) and inter-topic distance (Sievert, 2014).

A set of statements or facts is said to be coherent if they support each other. Topic coherence measures a single topic by measuring the degree of semantic similarity between high scoring words in the topic. These measurements help to distinguish between topics that are semantically interpretable topics and topics that are artifacts of statistical inference. There are several coherence measures, thus we focus is on the following: *Cv* measure, which is based on a sliding window, one-set segmentation of the top words and an indirect confirmation measure that uses normalized pointwise mutual information (NPMI) and the cosine similarity. We rely on the existing implementation from Gensim library.[4]

Inter-topic distance is computed based on the Jensen-Shannon distance which computes the difference between two probability distributions. Based on this distance, the principal components analysis is applied to represent each topic as a circle in a 2D vector space. Next, the distance between every two topics is computed as the distance between their circles (i.e. by subtracting the radiuses of the two topics from the distance between their centre points). The overall inter-topic distance is computed as the average distance between all topics in the model. The PyLDAVis library [5] was used in the current paper to compute the inter-topic distance and to visually represent the learned topic models.

Content Curation using Topic-based Information Retrieval

The main goal of the current work is to organize materials available in the Transcribathon tool in several groups of closely related documents. To achieve this purpose, we rely on the LDA technology to cluster documents around finer grade topics based on their description.

Apache Spark MLlib [6] library offers a robust implementation for LDA, which can easily scale to a large set of documents. Still, providing the appropriate data for model learning requires the implementation of a pre-processing pipeline, which is presented in the following subsection.

While variating the learning parameters, we aim at identifying the most appropriate model for the given dataset. The learned LDA model is computing for each Transcribathon the probability of belonging to every individual topics. Consequently, it offers the required functionality for curating the existing historical documents in more homogeneous data collections.

The second goal of our work is to offer an effective and efficient solution for curating new materials from Europeana into the topic driven collections. The most relevant words associated to each topic are used to find candidate documents using the Europeana search API[7]. The scoring of the search results is not correlated with the LDA probabilities for document-topic relationships. Therefore, the recommendations of new materials for individual topics follows a searching and reranking approach, where the recommendations are ordered by the LDA probabilities.

In general, the quality of the learned topics is a matter of subjective evaluation. However, metrics based on statistical data or probability distributions can be used as objective means for comparing different LDA Models. Therefore, we follow the strategy of identifying the LDA model that offers the better document clustering, meaning that the overlap between the topics is minimal and the cohesion within the topics is maximized (i.e. the average similarity between the documents assigned to an individual topic is maximized). Consequently, the decision for

[4] https://radimrehurek.com/gensim/models/coherencemodel.html
[5] https://pyldavis.readthedocs.io/en/latest/readme.html
[6] https://spark.apache.org/docs/latest/ml-guide.html
[7] https://pro.europeana.eu/page/search

selecting the best LDA model is based on a combination of the inter-topic distance and topic coherence metrics.

Further level details for the implementation of the proposed approach are provided in the following subsections, while its effectiveness is measured within the experimental evaluation part of the paper.

Pre-processing pipeline

Given the heterogeneity and the different languages used for describing the documents available in Transcribathon (i.e. named as stories on the website), a data processing pipeline is required to generate the input for learning meaningful topics. Therefore, each document description is run through the processing pipeline illustrated in Figure 1.

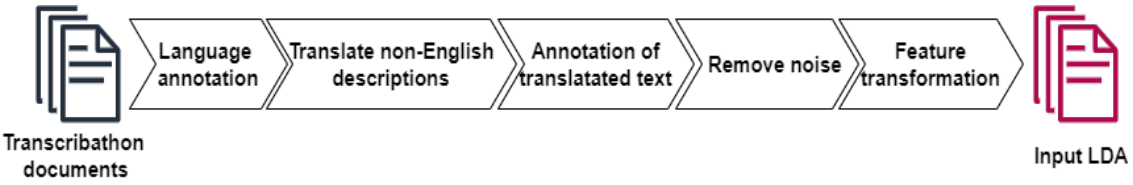

Figure 1. Pre-processing pipeline

The first step in the pipeline is used to convert the original document descriptions into plain text by removing the available HTML markups, followed by the automatic detection of the language used within the original text. All non-English descriptions are automatically translated using the Google Translation API.[8] While the translation service still has some difficulties in translating mixed languages (i.e. which are sometimes still present in the document description), the language detection is once again applied on the translated text, and the noisy translations are removed from the dataset.

After ensuring that all documents are assigned with valid descriptions in the English language, the resulting text is transformed into a document term count vector from which a predefined list of stop words is also removed (i.e. a list of common words with very low semantic information). The document term count vector represents the input for LDA model learning.

Learning Topic Models

LDA is an unsupervised machine learning approach which uses two configuration parameters: ***K*** – the number of topics to be learned and ***I*** – the number of iterations run for building the topic model. In each iteration, the probability distribution of terms to topics and of documents to topics are adjusted to improve the document to topic assignments. To find the most appropriate values for the LDA configuration, we run several experiments with different values for ***K*** and ***I***.

The entire process used for identifying the best topic model is illustrated in Figure 2. This process generates multiple LDA models (i.e. named *LDA_K_I*) for each combination of the ***K*** and ***I*** values, from which we have the goal to select the most appropriate one for the given dataset. This assessment is made by using the two complementary metrics. While the coherence - *coh(LDA_K_I)* - indicates how well the documents assigned to the same topic relate to each other, the inter-topic distance – *itdist(LDA_K_I)* - is an indicator on how well the document clusters are separated from each other. The better models are maximizing the values for both these metrics, however, the metrics are not indicating the same optimal values for ***K*** and ***I***. As the inter-topic distance is a surjective function with respect to ***K***, we use the derivate of the

[8] https://cloud.google.com/translate

function with respect to ***K*** to assess the quality of the models (see also Experimental Evaluation section).

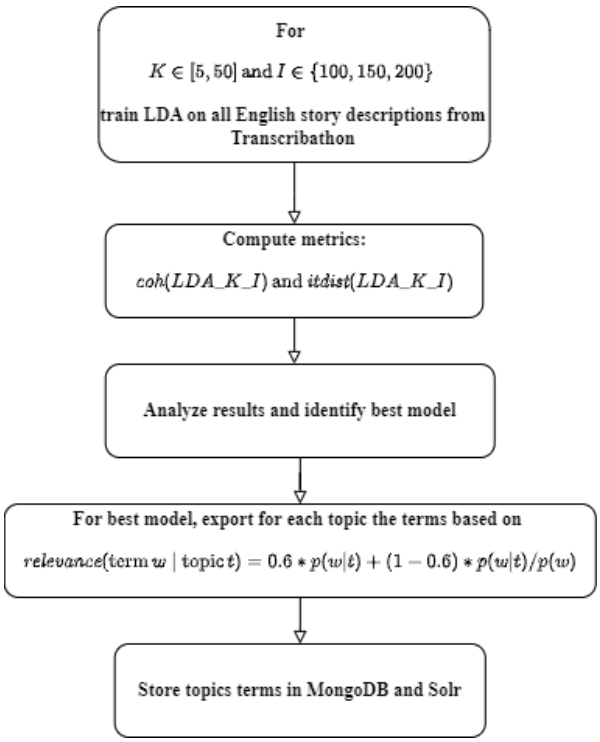

Figure 2. Procedure for finding the best topic models for the Transcribathon dataset.

The representation of the topics learned within the selected model are meant to be used for assigning new incoming documents to the existing topics and for searching new candidate documents in external repositories. Therefore, we implemented a Topic Management API, which is storing the topics representation into a Mongo database [9] and indexes them on a Solr [10] server. While the number of terms in a topic is increasing with the size of the dataset, we integrate an additional processing step to select the most relevant terms for each stored topic.

The selection of topic terms is based on the relevance function. The LDA model learning computes the probability distribution for the words in the dataset – $p(w)$, and for the words belonging to a topic – p(w|t). We select only a limited subset of topic terms containing the most significant words for each topic. The relevance is described in (Sievert, 2014), and we set the value for $=0.6$. The top 500 most relevant terms for each topic and their relevancy score are stored in the database and indexed for efficient searching.

Topic-based Information Retrieval Approach

This subsection describes the functionality for storing topics in an inverted index which are efficiently used as a search engine (based on Solr technology) and for identifying the most relevant topics for new documents.

For searching topic-related content in Europeana, efficient computation of answers is crucial given the very large number of ingested documents, thus, to apply the learned topic model, a trade-off between accuracy and efficiency must be implemented. That is why our approach is to keeps only the top 500 most relevant terms for each topic. The vocabulary of the Transcribathon

[9] https://www.mongodb.com/
[10] https://solr.apache.org/

dataset contains approximatively 28,000 words. While LDA computes the probability for each word to belong to a given topic, the probability vector contains a long tail of very small probabilities. Consequently, such words have almost no influence in deciding if a document belongs or not to a given topic.

The topic management API supports the retrieval of Europeana documents belonging to a given topic. The main challenge in realizing such service is the computational complexity for such a large repository. Currently, the Europeana platform aggregates more than 50 million CH objects, and new records are ingested periodically. The pure LDA approach requires the application of the pre-processing pipeline and the topic prediction for each of these documents.

To address the processing complexity, we rely on the Solr search capabilities for selecting relevant documents for LDA processing, making this service efficient in practice. A Solr query is generated based on the topic terms, where the relevance score is used for boosting individual terms. Note that Solr computes the similarity between query terms and documents based on cosine similarity, which is not well corelated with the LDA model which uses conditional probabilities (which is currently not realizable using the standard Solr search). Consequently, recommending new documents from Europeana for a given topic follows a two-step approach. In the first instance, relevant documents are pre-selected by searching the repository and the LDA probabilities are computed for re-ranking the results. The documents with low LDA probability are removed from the recommendation list.

The underlying LDA assumption indicates that a document may belong to one or more topics. Still a threshold needs to be applied for deciding which documents should be dropped from the recommendation list. Within the experimental evaluation we aim at measuring the recommendation precision in two scenarios: 1) the precision at Top 10 considering only the main topics of the document and 2) the precision at Top 10 considering the main and the second topics for the document. Consequently, for the first scenario, we consider only the documents with a probability higher than 0.5 as being relevant recommendations, while for the second scenario, a threshold of 0.3 is applied. In this way we can achieve a scalable solution for recommending new documents from Europeana for each individual topic.

Experimental Evaluation

The experimental evaluation presented in the following subsections aims at answering the following questions:

- Which are the ***K*** and ***I*** parameters for learning a good LDA Model on the complete Transcribathon dataset?
- When recommending documents for the learned topics by using the proposed approach, which are the configurations required to obtain a good *Precision* at Top 10?

At the time of instrumenting this evaluation, there were 31,957 documents (i.e. stories) available in the Transcribathon platform. The great majority of these documents are related to the crowdsourced collections on the historical events such as the First World War, the 1989 Revolution in Eastern Europe, the Industrial Revolution and other materials showcasing the cultural and urban development from the 19th Century. Their free text description is available in one of different European languages, only 273 of them having the original description in English. After running the pre-processing pipeline (i.e. which includes the translation to English), a vocabulary of 28,000 words is retained for learning the LDA models.

Evaluation Method

For answering the first question proposed for the evaluation, based on previous experience, we are variating the values for ***K*** in the range of [5,50] with step of 5, while for ***I*** we consider values

in the range of [100, 300] with a step of 50. The coherence and inter-topic distance metrics for all investigated setups are computed and represented graphically for expert evaluation which selected the most appropriate LDA model.

For the second question proposed in the evaluation, we use the *Precision@10* metrics to evaluate the quality of the recommendations computed with the proposed approach. When searching the candidate documents for a given topic, we take in consideration a number ***N*** of search results, by variating ***N*** in the range of with a step of 50. The LDA model is used to re-rank the recommendation list and the relevant recommendations are computed based on the LDA probabilities, by using two thresholds. By using a lower threshold of 0.3, we consider to be relevant recommendations also the ones for which the current topic is not the main topic of the document. While using a threshold of 0.5, the recommendations are relevant only for document's main topic. For each topic in the model, *Precision@10* (i.e. fraction of relevant documents in the top 10) is computed based on the recommendation list by using the before mentioned decision criteria.

Selection of the Most Optimal LDA Model

Figure 3 and Figure 4 present the metrics for evaluating the learned LDA models with different configurations for ***K*** and ***I*** parameters. We aim at identifying the model which performs well in terms of clustering from perspective of topic cohesion (i.e. measured by *Coherence*) and the separation of the document clusters (i.e. measured by the normalized *Inter-topic Distance*).

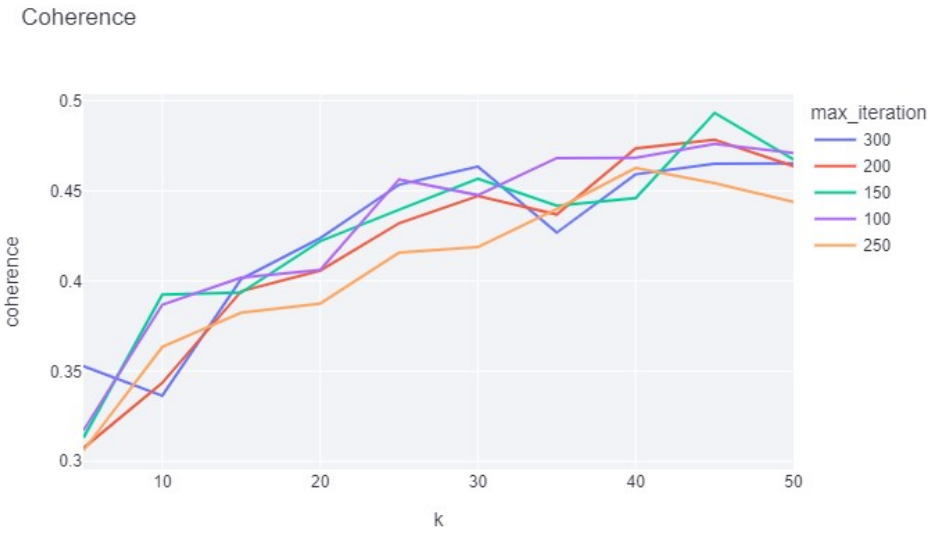

Figure 3. Coherence based on ***K*** (number of topics) and ***I***(number of iterations)

Given that coherence and inter-topic distance measure different aspects of the clustering performance, it is not surprising that they do not agree on the best LDA model. The normalized inter-topic distance indicates a maximum for ***K***=10, which is slightly better than for ***K*** =15. The model learned with 100 iterations underperforms, while the other models have a comparable performance from this perspective. Therefore, one may conclude that the model training riches a saturation when ***I*** > 150. The coherence metric indicates a higher variation of the results and starting with the ***K*** = 30 the performance of the models with different number of iterations starts to be unstable, which is an indicator of model overfitting. While the topic cohesion increases for all models with ***K*** < 25, the performance for most of the models converges for ***K***=15 (except for ***I*** = 250), meaning that at this point, the models have comparable performance, independent from the number of iterations. The variation between ***K***=15, ***K***=20 and ***K***=25 is still placed in a small range. Based on these observations, the most appropriate model for this dataset is

considered ***LDA_15_150*** (meaning the topic model obtained by training LDA with 15 number of topics and 150 as maximal number of iterations).

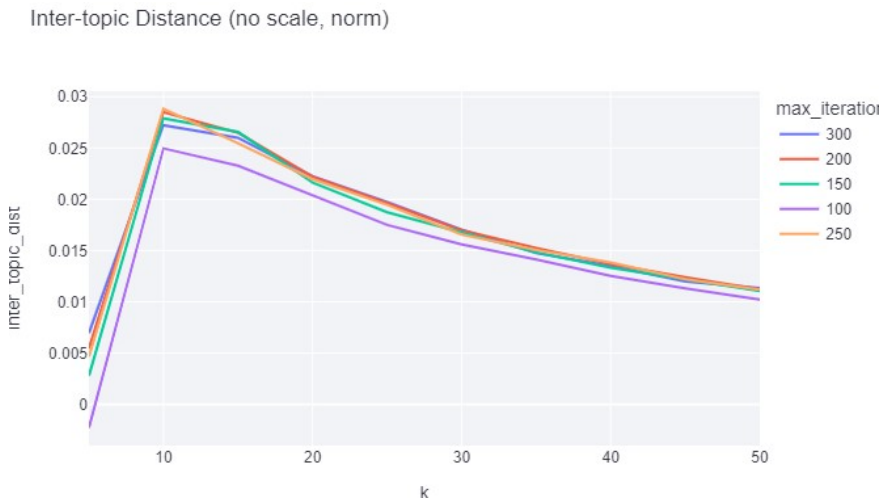

Figure 4. Normalized Inter-topic Distance based on ***K*** (number of topics) and ***I*** (number of iterations)

The topics learned with the selected model are presented in a visual form for cross validation by domain experts using the pyLDAVis as shown in Figure 5. The visualization of the inter-topic distance map, presents a 2D visualization of the topics, indicating a good clustering of documents over the 15 topics. The topics have low overlap and a balanced distribution of documents in topics (e.g. in terms of topic size). The most relevant terms for each topic can be visualized as a histogram, in which the term frequency is indicate for the given topic and for the whole vocabulary. By looking into the terms of Topic 4, one may understand that this topic is grouping postcards, photos, illustrations sent home from the front by the soldiers during the First Word War.

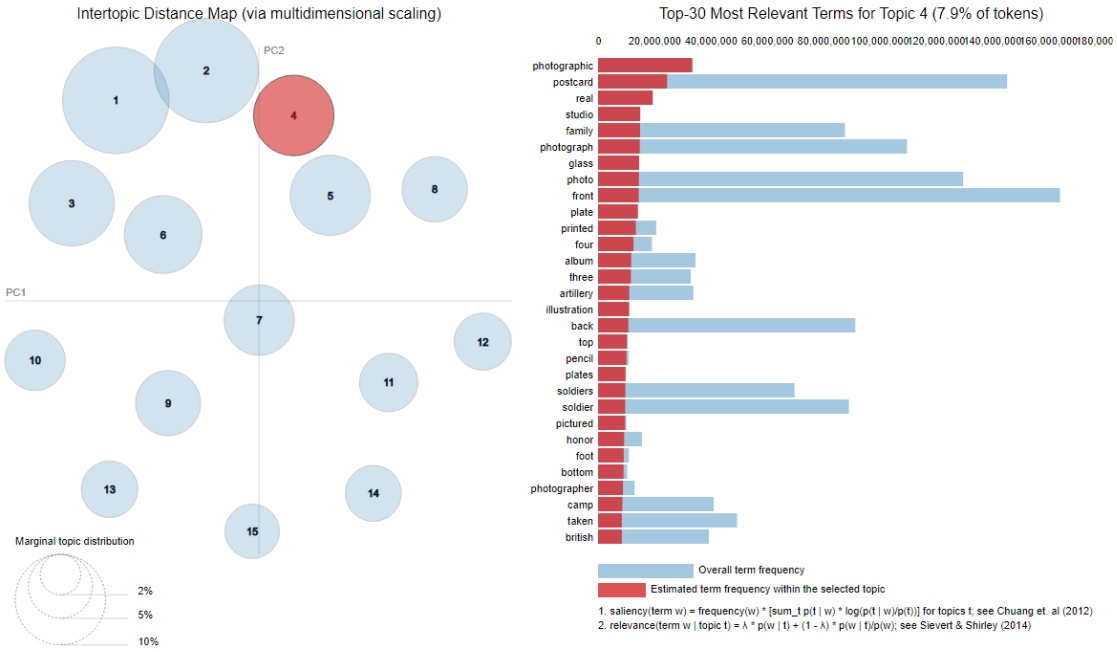

Figure 5. Visualization of LDA_15_150 (15 topics and 150 iterations).

Topic-based Search Evaluation

In this subsection the results of the evaluation for the topic-based search are presented. The results for *Precision@10* for different number of candidate documents - N and the relevant document thresholds of 0.3 and 0.5 are presented in Error: Reference source not found6. The precision of recommendations increases with N, but also the processing complexity. When using a threshold of 0.5, the proposed approach reaches a precision greater than 50% is obtained for N≥550 (for N=600, Precision@10 = 56%). This represents the case when the documents are considered relevant only for their main topic. When relaxing this constraint to consider the documents relevant for their second topic by setting the value for the threshold to 0.3, an average precision above 80% is obtained when re-ranking 600 documents retrieved through Solr search.

By analysing the precision for individual topics, one can conclude that most of the topic representations are able to generate good top 10 recommendations, for N=600. However, there are several smaller and more specific topics for which the number of preselected documents will need to be increased to compute good recommendations, such as topics 8 and 9. The threshold of 0.3, was set based on heuristic interpretation of probabilities. However, this threshold could be also computed based on the highest probability of the third topic assigned to all documents.

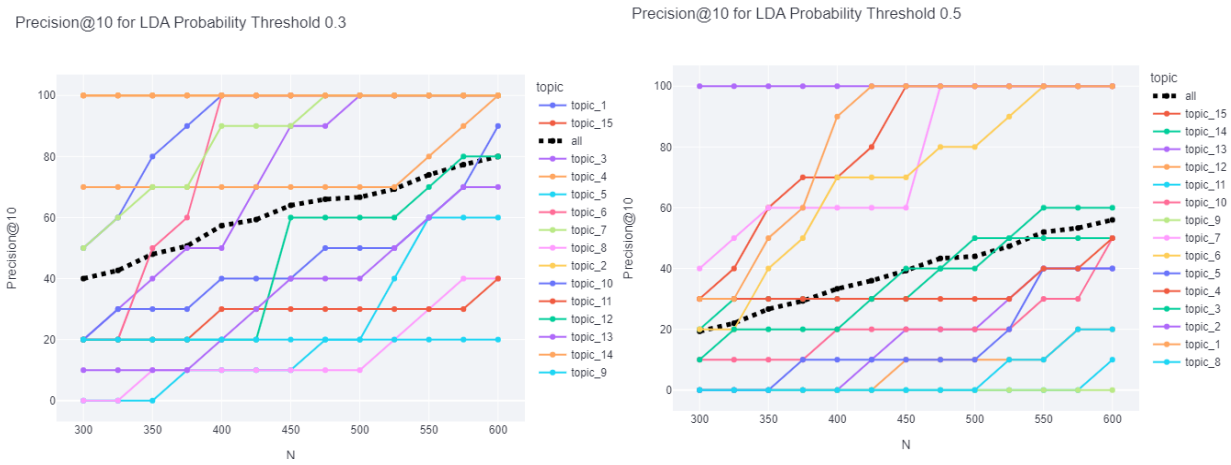

Figure 6. Precision@10 for each topic based on probability threshold of 0.3, respectively 0.5 and N ∈ [300,600]

Results Discussion

The first observation is that for each ***I*** in [100, 300], the coherence and inter-topic distance metrics vary the least when the number of topics is 15. This means that according to both metrics, the overall optimum is achieved at this point. For all other ***K*** values, the variation is much larger, meaning that the model performance is not stable with respect to ***I***. This motives our selection of LDA_15_150 as the best model for the Transcribathon dataset.

The second empirical observation indicates that the representation of the topics by their 500 most relevant terms using different values for parameter doesn't have a relevant impact on the precision, therefore not included in the experimental evaluation. The experimental results presented in the paper are those for

The third observation is that the pure search-based approach is not able to approximate the LDA topic assignment for documents. That is due to the fact that LDA uses the conditional probabilities when assigning documents to the topics. Even if this might be implemented in internal repository of the Transcribathon platform, it is not feasible to assume that such probabilities would be included in general purpose repositories like Europeana. However, as

shown in the experimental evaluation, the proposed approach can still identify relevant documents, without the need to compute the LDA document-topic assignments for the documents available in external repositories. We can achieve a good recommendation precision in the top 10 by retrieving between 500 to 600 documents which contain topic terms.

The presented experimental evaluation was focusing on recommending documents for topics learned with the LDA model. The other scenario of assigning new ingested documents to the existing topics is also relevant for digital curation activities. There is an open question if a good approximation of the of the LDA similarity function can be implemented based on Solr by including the topic-term probabilities in the Solr search. Note that after building the LDA model, the topic-term probability matrix is fixed with a size of $\boldsymbol{K}$* **T**, where $\boldsymbol{T}$ represent the number of terms stored for the representation of each topic.

Conclusions and Future Work

This paper presents a scalable approach for clustering large corpora of historical documents in finer grade collections. We proposed a well-defined protocol for learning and choosing the best topic model to support the curation of new materials for Transcribathon campaigns. We combine the common search functionality from large CH repositories like Europeana to reduce the computation efforts required by LDA based document clustering.

This work offers a novel perspective in how platforms like Transcribathon or Europeana can use topic-based searching and recommendations for curating online collections. Such functionality is enabled through the advances made in the past years in the AI domain, including development of performant machine translation and natural language processing technologies.

For future work, we would like to evaluate the Solr-based topic detection from the user perspective and based on this evaluation investigate if it makes sense to create a customized similarity metric based on the LDA conditional probability. We would also like to investigate other topic modelling techniques such as BERTopic which supports the use of multilingual sentence embeddings which seems promising and a good competitor for LDA.

References

Blei, D.M., & Lafferty, J.D. (2005). Correlated Topic Models. NIPS'05: Proceedings of the 18th International Conference on Neural Information Processing Systems, 147-154.

Blei, D.M., & Lafferty J.D. (2006). Dynamic Topic Models. ICML 2006 - Proceedings of the 23rd International Conference on Machine Learning., 113-120.

Blei D. M., Ng A. Y. & Jordan M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning Research*, 3(Jan), 993-1022.

Dumais S. T. (2005). Latent Semantic Analysis. *Annual Review of Information Science and Technology* 38, 188–230.

George, L. E. (2018). A study of topic modeling methods. In 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS). (pp. 109-113). IEEE.

Hofmann, T. (1999). Probabilistic Latent Semantic Analysis.

Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI'99).

Jelodar, H. W. (2019). Latent dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools Appl.* 78, 15169–15211.

Mikolov, K, Sutskever, I., Chen, K., Corrado, G.S., & Dean, J. (2013). Distributed Representations of Words and Phrases. NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, 3111-3119.

Moody, C. E. (2016). Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec. arXiv preprint arXiv:1605.02019.

Sievert, C. & Shirley, K (2014). LDAvis: A method for visualizing and interpreting topics. In Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, 63-70.

Syed, S., & Spruit, M. (2017). Full-Text or Abstract?Examining Topic Coherence Scores Using Latent. 2017 International Conference on Data Science and Advanced Analytics. In 2017 IEEE International conference on data science and advanced analytics (DSAA) (pp. 165-174). IEEE.

Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. Inf. Syst. 94.