Combining Topic Models for Corpus Exploration

Applying LDA for Complex Corpus Research Tasks in a Digital Humanities Project

Carsten Schnober[†‡]
[‡]Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for International Educational Research
schnober@ukp.informatik.tu-darmstadt.de

Iryna Gurevych[†‡]
[†]Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science
Technische Universität Darmstadt
gurevych@ukp.informatik.tu-darmstadt.de
https://www.ukp.tu-darmstadt.de/

ABSTRACT

We investigate new ways of applying LDA topic models: rather than optimizing a single model for a specific use case, we train multiple models based on different parameters and vocabularies which are combined on-the-fly to comply with varying information retrieval tasks. We also show a semi-automatic method which helps users to identify relevant topics across multiple models.

Our methods are demonstrated and evaluated on a real-world use case: a large-scale corpus-based digital humanities project called *Welt der Kinder* ("Children and their World"). We illustrate our approach in that context and show that it can be generalized to other scenarios.

We evaluate this work using empirical methods from information retrieval, but also show visualizations and use cases as actually applied in the project.

Keywords

Digital Humanities; Topic Modeling; LDA; Information Retrieval

1. INTRODUCTION

LDA topic models [2] have served as a helpful tool for corpus exploration in many practical use cases. They are nowadays an established tool in the field of digital humanities where they assist researchers in exploring large amounts of text data [8, 20, 12, 9]. In most of these cases, the data has been of manageable size or has been filtered in advance to contain only relevant portions, and, more importantly, domain and use cases have been defined relatively strictly.

However, there is a practically infinite number of permutations with pre-processing options and LDA parameters to produce suitable topic models. While these parameters are often successfully fine-tuned based on statistical measures and on experience to optimize a model for a specific research task, the same model might not be very useful for a different task.

This work is motivated by a use case in which corpus researchers have widely varying requirements that often involve various angles and different levels of complexity. Researchers are interested in different aspects in terms of content and lexis, for instance in opinionated expressions or in certain countries, regions, or persons. In settings like that, single topic models that work well in a fixed setup with well-defined research tasks, often fail to provide suitable topics even with a very large number of topics. We aim to combine human and machine workflows to overcome that limitation.

In order to create a user-friendly interface, we have implemented a project-specific web application which is backed by an Apache Solr[1] server. It incorporates the present approaches to topic models, as well as term search, facet search based on document metadata, and different views on the data (Figure 1).

We present a new approach in which we generate multiple, mutually independent topic models to facilitate the exploration of a large, noisy, and heterogeneous corpus. This helps to overcome the said limitations of single topic models, that are often unable to generate topics suitable for all information requests expressed by humanities researchers. Thanks to cheap storage and increased computational power, keeping pre-computed topic mixtures for a large corpus on disk or even in memory and to aggregate them upon request has become feasible.

In our context, corpus exploration means corpus-based text research in which history scientists investigate a large corpus. They require access to statistics and to single documents with regard to their relevance for very specific and complex research tasks. In such tasks, approaches based on simple term search often are not sufficient to handle problems like word sense ambiguity, OCR errors, and authors using implicit terminology. Our approach aims to tackle such challenges that frequently occur in research projects in the field of digital humanities.

In order to allow researchers to take different angles on a corpus, we generate rather general as well as highly specific

TM'15, October 19, 2015, Melbourne, Australia.

ACM 978-1-4503-3784-7/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2809936.2809939 .

[1]Apache Solr: `http://lucene.apache.org/solr/`

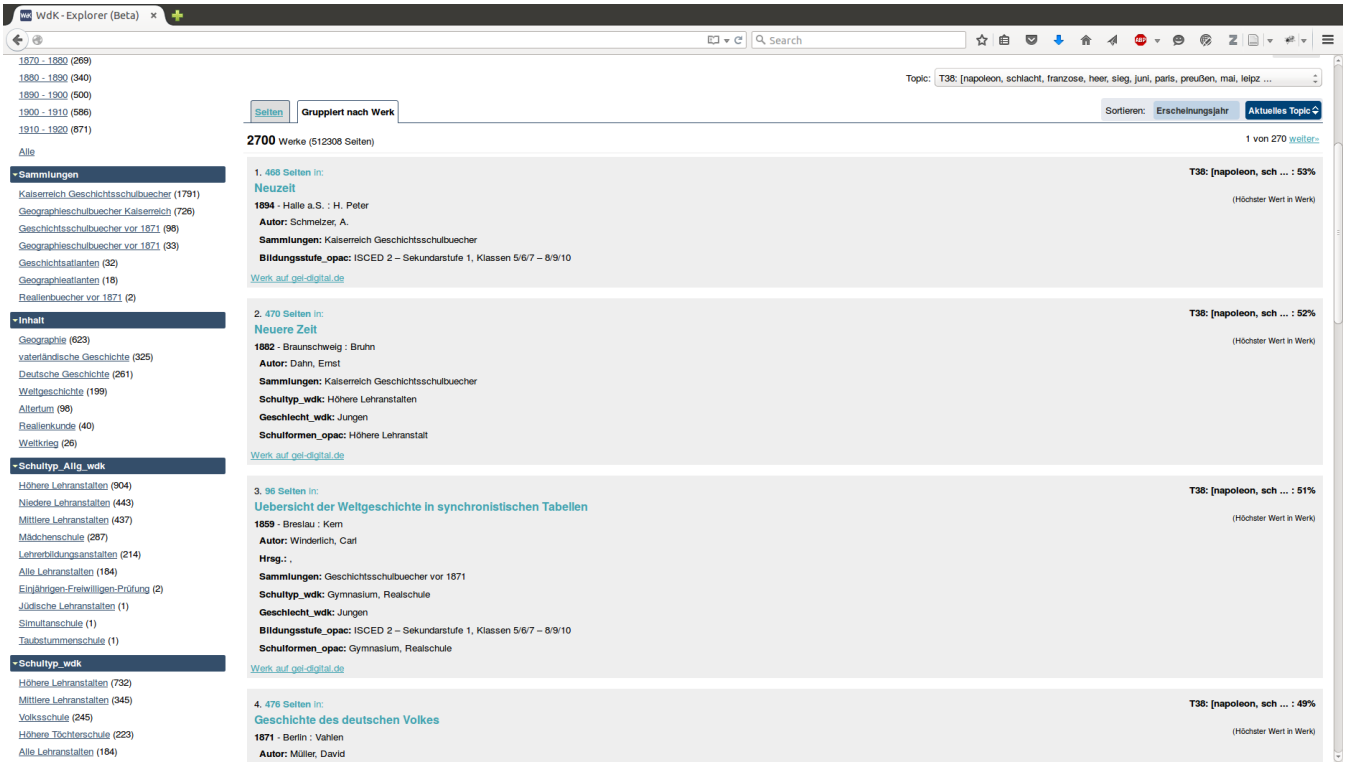

Figure 1: A project-specific web interface gives corpus researchers access to the retrieval functionality presented in this work.

topic models based on disjoint and on overlapping sub-sets of the total vocabulary and on different LDA parameters.

In our concrete use case, the vocabulary sub-sets on which the various topics models are based are pre-filtered by part-of-speech tags, named entities, and a sentiment lexicon respectively. Other use cases may benefit from different filtering choices.

From an information retrieval point of view, we show how to combine topics that have been generated by any of these models in order to fulfil the requirements of complex corpus research tasks. We thus apply standard information retrieval evaluation metrics in order to measure the performance of our approach and compare it to baselines based on tf-idf [18] and on a single topic model.

Evaluation is difficult, especially with respect to recall because the corpus is large and it is often hard to decide, even for humans, whether a document is relevant for a specific query. We have thus manually assembled sub-corpora with relevance annotations. These sample corpora consist of a few hundred documents represent sub-tasks of more general, complex research tasks and are used to measure recall.

1.1 LDA optimization: a hard problem

Improving the results of a topic model estimation process can be approached from two main sides: on the one hand, there are statistical settings including the Dirichlet parameters, and on the other hand there are the linguistic assumptions regarding text pre-processing.

The essential LDA hyper-parameters are the number of topics k, and the Dirichlet priors α (concentration of per-document topic distributions) and β (concentration of per-topic word distributions). Many measures have been proposed to find or approximate optimal values for these parameters both automatically and by human evaluation [22, 3]. Nevertheless, finding optimal parameter settings remains a hard problem, depending on the data and the context.

On the pre-processing side, there are many further decisions to take, for instance: should one lemmatize the text; even if the data is noisy and/or in a language in which lemmatization is error-prone? Should one remove and/or substitute stopwords, numbers, punctuations, frequent words, rare words? And if so, which ones, and what are decent thresholds?

Most practitioners rely on common best practices to answer such questions based on experience and intuition, but empirical evidence is hard to provide. Additionally, it has been shown that models that have been optimized on statistical measures such as perplexity and held-out likelihood do not necessarily correspond to what is perceived as best by human users [3].

The conclusion on which this work is based is that the effect of optimizing a single model for a certain task hits limits sooner or later. Instead, we generate multiple models where each one seeks to perform well on a specific job – and together, they can handle more complex tasks.

2. MOTIVATION AND BACKGROUND

The presented approach has been motivated by a real-world digital humanities project, called *Welt der Kinder* ("Children and their World")[2], in which several humanities researchers (historians, in our case) work on very different research tasks, but on the same corpus.

Our text corpus currently comprises approximately 3,500 historical German textbooks[3] from the 19th century, covering a wide range of domains. It includes approximately 600,000 pages and 59M tokens.

Digitization of the corpus has been performed within a dedicated, previous project *GEI digital*[4] [19]. Due to the relatively low accuracy of the OCR process (approximately 90 percent on character-level) on that data – scans of over 100 years old paper, variations of German Gothic print, non-standardised orthography – supervised techniques yield bad results even on simple tasks such as sentence boundary detection much less complex tasks such as part-of-speech tagging and lemmatization. Figure 2 shows the content of an example page in German.

A central research goal in this project is to find empirical evidence for hypotheses about the quantitative intensity of certain discursive matters: which topics are discussed (or are not) in certain sub-collections? For instance, do textbooks made for boys' schools discuss war-related matters more frequently than the ones for girls' schools? Are the Napoleonic wars more present during certain time periods than during others for propagandistic reasons?

We use the topics derived from different models and their respective distributions over the corpus to quantify and to visualize such questions. Researchers are able to freely define the topics they are interested in and to immediately get a graphical illustration of the topic distributions. Figure 3 shows an example in which the proportion of a topic is visualised across time.

Sub-collections based on time as well as on other categories as mentioned above are generated on-the-fly, using faceted search or keyword search. Figure 4 shows an extract of categories that are available as facets in our corpus and that can be used and combined to create sub-collections on-the-fly. Our approach allows to extract topic proportions from different models among such instantly created sub-corpora.

The overall task can be re-phrased as: a) "how many documents (pages) are there in a sub-collection that discuss a certain topic?" or "what is the average proportion of a topic within a sub-collection?" The challenge we tackle in this work is that such topics – in the more abstract linguistic sense rather than in the formal sense of topic modelling – often involve various angles and aspects and are therefore too complex to be represented by a single topic or model. Examples are given in sections 4 and 5. Figure 2 shows a hit for a topical search for documents regarding the Napoleonic wars.

In a pilot study we conducted within the project, we have identified sample research questions that are relevant for historians in a real-world scenario. For this work, we use

[2] *Children and their World* project homepage: `http://welt-der-kinder.gei.de/`

[3] All terms are immediately translated into English in this paper; the original terms in German are provided in footnotes where necessary for clarification.

[4] *GEI digital*: `http://gei-digital.gei.de`

Figure 4: Some of the metadata categories of our corpus that can be used to create sub-collections on-the-fly and to extract and visualize topic proportions as shown in Figure 3.

Figure 2: An OCR'd document in German with high relevance for the topic related to Napoleonic wars. Topically relevant terms are bold-faced. This view provides additional information to the users such as automatically extracted named entities. Relevant topics from other available models are also displayed. The number in the top-right corner displays the relative relevance of the selected topic for the document.

these sample questions in order to assemble realistic research tasks. An important result of that pilot study has been that the questions vary significantly on many levels. They differ with respect to the actual content, but also in terms of specificity. Some questions circle around very broad issues such as "transport" or "colonization", others are very specific, for instance about a certain war. Additionally, more or less subtly expressed opinions about certain entities (countries, peoples, etc.) are of particular interest within this project.

A secondary requirement comes close to a classic information retrieval task: for corpus-based research in digital humanities the concept of "distant reading" [13] has been shown to play an important role. Corpus researchers need to be able to virtually zoom in from a distant scope reflecting statistical tendencies over the whole corpus or subcollections, down to the level of single document instances. Again, such requests take place in the context of specific information requests of varying complexity.

Topic modelling has been chosen for this project to facilitate research about concepts, or "latent topics" [2], rather than for exact terms. For instance, the result set for a concept such as *Asia* should include documents that do not necessarily contain the term "Asia" but possibly more or less directly related terms such as "China", "India", "Mongolia", or "Tibet".

In our particular use case, this also concerns words that were misspelled or misrecognized during the digitization process, for instance "Asla" instead of "Asia". Assuming that orthographic variations and recognition errors of a word occur in similar contexts, erroneously recognized terms are assigned with largest weights in the same topics as their correctly spelled and recognized counterparts.

In this work, however, we focus on the problem of various degrees of specificity and different aspects and angles that are implied in the research questions posed by historians. A single model cannot possibly generate topics that serve all the information needs for such a variety of requirements involving named entities, opinionated expressions etc. Our approach is thus to generate multiple mutually independent topic models using standard LDA techniques.

3. RELATED WORK

Since their early days, LDA topic models [2] have been attractive for exploring large corpora because of their intuitive, human-readable output and because they can be generated in a completely unsupervised way on an unannotated text corpus. Taking co-occurrences into account, it estimates a fixed number of topics that are represented by a weighted list of terms. The most highly-ranked terms are the most representatives for such a topic.

Topic models have been improved by incorporating additional knowledge or metadata where available. An extension that is often used for analyses of corpora that span across time are Dynamic Topic Models [1] where topics and word weights are adapted over time. Structural Topic Models [17] take general metadata into account when estimating a topic model.

Joint Sentiment/Topic Models [10] explicitly incorporate sentiment analysis into LDA topic models, based on a sentiment lexicon. In Multi-grain LDA [21] global and local topics are distinguished, again in a sentiment analysis setting.

The goals set in these approaches tackle problems that are similar to the ones posed by the corpus researchers in our project. However, such methods are optimized for a single task such as sentiment analysis; in our case, each one of those tasks represents only one amongst many others.

LDA topic models have been shown to work well in practical research tasks. [8] and [9] discover topics within different corpora of scientific publications. [20] applies LDA on his-

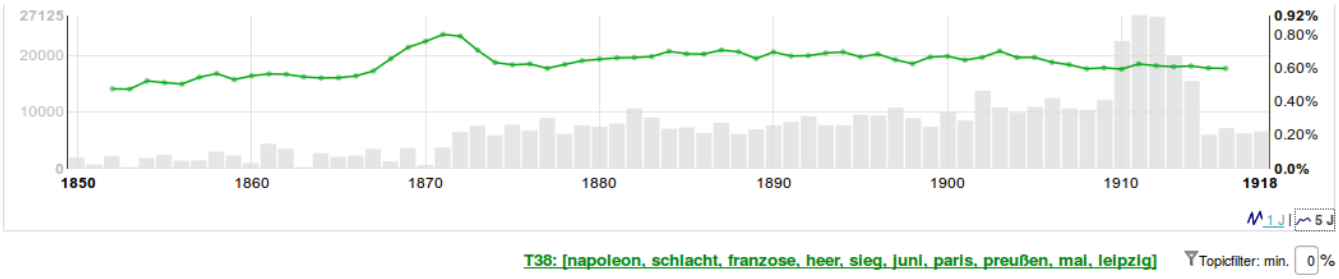

Figure 3: LDA topic proportions quantify the intensity of a topic over time and/or in sub-collections (5-years sliding window mean). The bars shows the absolute number of pages in which the selected topic is the most relevant one, referring to the left-hand side y-axis scale, the green line shows the relative topic proportion, referring to the right-hand side y-axis scale. This example shows a peak in a topic related to *Napoleon* shortly after the Franco-Prussian war in 1870/71. The peak in the absolute numbers between 1910 and 1915 is due to a non-evenly distributed collection that is skewed towards the later years of the selected time-span.

torical newspapers. [4] applies LDA for sociological studies regarding discussions about government assistance for the arts in the U.S.

These approaches apply single models to perform predefined tasks. In contrast, we generalise so that we can provide a research tool that is flexible enough to work on heterogeneous data on various research tasks.

4. COMBINING INDEPENDENT TOPIC MODELS ON-THE-FLY

The following steps are described in this section:

1. (off-line) Estimate topic models and infer topic mixtures for all documents.
2. Formulate research task(s) and concepts.
3. Identify relevant topics in generated topic models (manually or semi-automatically).
4. Retrieve documents according to accumulated topic scores.

In step 1, multiple topic models are generated. This needs to be done only once and can take between hours and days, depending on corpus size, computing capacities, and number of topics. It is therefore performed *off-line*, i.e. as a preparatory step to build the system that can then be used on-line by the users.

In step 2, a task is formulated by a corpus researcher and broken down into concepts that can be used in an information retrieval-like setting. Subsequently, these concepts are mapped to topics that were generated by any of the models. In step 4, documents that are relevant for the defined task are retrieved from the corpus.

4.1 Generation of Topic Models

We have generated three topic models to illustrate and test our approach. The models differ mostly with regard to their topic numbers (k) and the vocabulary sub-sets they are based on:

M1 Nouns only, $k = 500$

M2 Named entities only, $k = 200$

M3 Sentiment words only, $k = 25$

The topic numbers have been chosen by manually observing the quality of resulting topics in a small pilot study in cooperation with human experts (historians). Quality judgements were based on the methods presented by [3] and took consistency, specificity, and redundancy into account.

Model **M1** was generated to address rather general research questions. Its vocabulary therefore comprises all nouns. As nouns are capitalized in German orthography, model **M1** is simply based on all the capitalized words in the corpus[5]. The problem of capitalized non-nouns at the beginning of a sentence, to a large portion determiners, is significantly mitigated with a simple stopword filter. This method helps us to work around the afore-mentioned, non-satisfying accuracy of state-of-the-art part-of-speech taggers on our noisy data.

Model **M2** is motivated by the corpus researchers' explicitly expressed interest in countries and persons. The model is thus generated only on words that have been tagged as named entities, using the Stanford Named Entity Recognizer [7]. Results on our data are far less accurate than for contemporary, clean data because of OCR errors and deprecated orthography. We have therefore extended the NER detection by adding words from a manually compiled list of historic names and spellings that occur in our corpus.

Model **M3** generates topics based on opinionated expressions. We use a static word list extracted from the German sentiment lexicon SentiWS [16] while ignoring the polarity provided by the lexicon. For this model, only words that are listed in the sentiment lexicon remain in the vocabulary, all others are removed.

These model types have been chosen to exemplify our approach and to comply with the real requirements expressed by the involved humanities researchers. Other models can be added in the same fashion when required for other tasks and/or data.

The vocabulary sizes after appliying the described filters lie between 30,000 (model **M3**) and 200,000 (model **M1**).

[5]Original German capitalizations are not reflected in the English translations.

4.2 Identifying Relevant Topics

This section describes how topics from different models are selected, manually or semi-automatically, to be used for a specific corpus research task. The selected topics serve as a basis for document scoring in the subsequent step (section 4.3).

A research task is phrased as an information request such as 'find documents that refer to hostilities taking place in Asia.' This can be broken down into two relevant concepts, or latent topics, provisionally labelled as "Asia" and "hostility". The next step is this to identify topics that address these concepts.

4.2.1 Manual Topic Selection

The first method for selecting relevant topics is to manually go through the top words of each topic in each model. For the afore-mentioned example, human researchers look for topics that refer to a) *Asia* and b) *hostility* in two separate steps. For each one of these two concepts, human experts were asked to compile a list of intuitively most representative terms. For *Asia*, "Asia" itself and "China", "Japan", and "India" were chosen, for *hostility* the terms are "enemy", "hostility", and "hostile". From the list of topics output by any of the three models, we then search those that have any of the respective terms among their top 10 ranked terms.

In the models **M1** to **M3**, the following candidate topics[6] are found by our experts, based on manual investigation of the top n terms of each topic, for *Asia*:

M1 (a) highlands, mountains, Indus, lowlands, Tibet, Iran, sea, Himalayas, part, China, [...]

(b) empire, king, Syria, province, Cyrus, Asia, Asia Minor[7], Persian, rule, [...]

(c) China, Japan, Peking[8], railway, city, trade, Siberia, Korea, harbour, [...]

(d) Europe, Asia, Africa, America, Australia, continent, island, structure, [...]

(e) (7 more)

M2 (a) China, India, Japan, Chinese, Asia, Siberia, Persia, Russian, Tibet, [...]

(b) Alexander, Greek, Olympia, Greece, India, Apollo, Athens, God, Eurystheus, Asia, [...]

(c) (6 more)

M3 None

From these topics, the human experts manually select those that intuitively correspond best to their research task concerning Asia and choose topics **M1**c and **M2**a. The other topics are discarded as they appear to refer to other issues (e.g. topic **M2**b clearly refers to Alexander the Great and Ancient Greece), their scope is too general (topic **M1**d lists all continents), or refer to different aspects like geographic entities (topic **M1**a). The latter insight demonstrates how difficult automatic topic selection is, as a lot of background knowledge is required to recognize topic **M1**a as being related to geography and thus not useful for our task.

We repeat the topic selection process for the second concept, *hostility*, by searching topics which have any of the words "enemy", "hostile", or "hostility" among their top 10 terms and find the following candidate topics for the context *hostility*:

M1 (a) emperor, king, Henry, plan, enemy, Frederick, Germany, pope, death, [...]

(b) enemy, battle, army, victory, fight, attack, camp, Romans, day, [...]

(c) enemy, horse, side, step, horseman, attack, bridge, army, path, escape, [...]

(d) (13 more)

M2 None

M3 (a) enemy, fleet, attack, escape, hostile, strong, save, destroy, right, solid, [...]

(b) death, enemy, friend, judge, [...]

(c) fight, war, heavy, defeat, bloody, new, enemy, [...]

For model **M1**, the candidate list shows that some of the topics are quite similar to each other. The human experts choose topics **M1**b and **M3**a as most specifically related to the concept of *hostility*.

In the manual selection process, human experts can make use of their linguistic and domain knowledge to select topics that match the concept in question. However, as shown in this example, the selection process can be cumbersome if there are many topics and, depending on the model, the decision between similar topics if often hard and the list of top n terms might be misleading.

4.2.2 Semi-automatic Topic Selection

Next, we describe a method to semi-automatically identify relevant topics from different models based on a few manually compiled seed terms. For that purpose, we re-use the terms that were defined for the two concepts *Asia* and *hostility* in section 4.2.1. A relevance score is computed based on the harmonic mean over the probabilities of $term_1 ... term_n$ of n seed terms for each available topic in all the models. Hence, the relevance score of a topic is:

$$relevance_{topic} = \frac{n}{\sum_{i=1}^{n} \frac{1}{p(w|topic_i)}} \quad (1)$$

In order to simplify the computations and to smoothen the results, we assign zero weights to all terms that are not ranked among the top 100 of a topic. Because term weights approach a Zipfian distribution, the exact weights of such lower-ranked terms typically lie only marginally above 0, anyway.

If any of the seed terms has zero weight (after smoothing), we define the total topic relevance score as 0; the harmonic mean is not defined in this case. Consequently, only topics are considered in which all the seed terms occur among their top 100 words.

The intuition behind choosing the harmonic mean for scoring topic relevance (rather than, for instance, the average) is that topics with a relatively large score for all of the seed

[6] We represent topics here and throughout this work by a list of their top-weighted terms.

[7] German: *Kleinasien*

[8] former romanization of Beijing

terms should be preferred; the harmonic mean tends towards the least of the input elements.

Eventually, the topic with maximum score from each of the models is selected to represent a concept, if it exceeds an experimentally defined threshold of 0.001. For the terms "Asia", "China", "Japan", "India", the remaining candidate topics are (with relevance scores in brackets):

M1 None

M2 (a) (0.0064) China, India, Japan, Chinese, Asia, Siberia, Persia, Russian, [...]

(b) (0.0050) Europe, Asia, America, Africa, Australia, European, China, India, Asian, [...]

(c) (0.0013) English, German, England, China, European, Germany, America, India, French, British, [...]

M3 None

Models **M1** and **M3** do not yield any candidate topics with a relevance score above the threshold, that is why the only automatically selected topic is **M2**a in this example.

The other concept in question, *hostility*, is represented by the seed terms "enemy", "hostile", "hostility". These are the relevant topics (again, with relevance scores):

M1 None

M2 None

M3 (a) (1.9275) enemy, fleet, attack, escape, hostile, [...]

(b) (0.2511) fight, war, heavy, defeat, bloody, new, enemy, [...]

Again, only one model (**M3**) returns candidate topics at all and the top scoring topic, **M3**a, is selected. This is the same topic as one of those also selected manually in the previous section.

Note that models **M1** and **M2** could not be expected to yield a topic for the given seed terms because they have been generated exclusively on nouns or named entities respectively (cf. section 4.1). The adjective "hostile" is thus not included in their vocabularies.

However, the presented method still works as long as there is at least one topic from any of the models that can be used. In fact, this demonstrates a strength of our combined models approach: if necessary, we can select topics from one or few models while others may fail to provide any topics that represents a concept appropriately.

4.3 Scoring Documents

Having selected one or multiple relevant topics $1...n$ for a specific task, the relevant topic scores are aggregated for each document.

The aggregated score over $topic_1...topic_n$ for a document is computed through the harmonic mean again over the probabilities for the document and the selected topics. $p(topic_i|doc)$ is inferred from the model that has generated $topic_i$ using Gibbs sampling [8] (cf. section 4.4).

$$relevance_{doc} = \frac{n}{\sum_{i=1}^{n} \frac{1}{p(topic_i|doc)}} \quad (2)$$

The intuition behind using the harmonic mean to combine topic weights rather than the average, is similar as for the automatic topic selection (cf. section 4.2.1): documents with relatively large weights for all of the topics should be preferred over ones that score high for one topic, but low for others.

4.4 Implementation

All the topic models used in this approach are estimated using standard LDA techniques as described by [2]. Our implementations for model estimation and for inference are based on the Mallet toolkit [11] and make use of parallel LDA [14] to speed up the computationally expensive model generation process. We incorporated the pre-processing steps and the model estimation into a UIMA pipeline [6] using uimaFIT [15] and analysis engines from the DKPro repository [5].

The topic mixture of a document is inferred using Gibbs sampling as described in [8]. Because the inference process is too time-consuming to be performed on-the-fly on the full document collection, it is done *off-line* once the models have been generated.

For each document, the topic mixtures are inferred and stored for each of the models separately. With each document, a total of $|w| = \sum_{m=1}^{n} k_m$ topic weights (or proportions) are stored where n is the number of models and k_m is the number of topics for model m.

5. EVALUATION

For the purpose of evaluation, we have modelled our approach as a pure information retrieval task and evaluate performance by precision among the top-50-ranked documents (*Prec@50*), recall, and aggregated F1 score of the two. As mentioned previously, the size of our corpus does not allow to annotate and count all documents. That is why we have manually assembled smaller sample corpora of a few hundred documents for the two exemplary tasks in this section. These manually annotated sub-corpora serve as gold standards in the recall measures.

For our combined topic models approach, we report two results, denoted as *TMs (manual)* and *TMs (semi-auto)*. The descriptions refer to relevant topics being selected manually (section 4.2.1) or semi-automatically (section 4.2.2) respectively.

Simple term search based on manually defined terms for each concept serves as one baseline, with tf-idf as a relevance measure [18]. These are the same terms that are also used as seed terms for topic selection in our approach. To retrieve documents for multiple concepts in the tf-idf baseline, we take the intersection of the results for each one.

For the term search we use an Apache Solr engine and apply the built-in stemming feature so that a search for "China", for instance, includes matches for "Chinese" in order to improve recall.

The other baseline is derived using a single topic model for retrieval, reported as *Single TM*. For that approach, we manually pick a single topic from the general-purpose topic model (**M1**) that comes closest to all concepts in question. Based on the chosen topic, we retrieve all documents in which that topic is ranked among the top 10 and rank them by the topic weight.

5.1 Sample Task 1: *Hostilities in Asia*

This research question of historians has already been described in section 4.1; it circles around the concepts *Asia* and *hostility* and is expressed as 'find text passages that discuss Asia in the context of hostilities'. In order to represent the two concepts, we re-use the seed terms "Asia", "China", "India", and "Japan" for *Asia*, and "enemy", "hostile", and "hostility" for *hostility*, as shown in section 4.2. These terms are also used for the tf-idf baseline.

For the single topic model baseline, the topic *China, Japan, Peking, ...*, referred as **M1**c in section 4.2.1, is used.

The sample sub-corpus we have assembled to measure the recall comprises 327 documents that specifically discuss hostilities around the former German colony of Qingdao[9] in Eastern China, representing a relevant sub-topic for the research task.

Table 1 reports the precision among the 50 most highly ranked documents for each method and the recall for the *Tsingtau* sub-collection.

	Prec@50	Recall	F1
TMs (manual)	0.88	0.486	**0.626**
TMs (semi-auto)	0.88	0.495	**0.634**
tf-idf	0.92	0.3	0.452
Single TM	0.1	0.706	0.175

Table 1: Results for *Asia* and *hostility*; recall for the *Tsingtau* sub-corpus (327 documents).

5.2 Sample Task 2: *Uprisings in Ancient Rome*

The second example research task is expressed by history researchers as 'Uprisings in Ancient Rome'. It can be broken down into the concepts *Ancient Rome* and *uprisings*. As seed terms, the human researchers define the terms "Rome" and "uprising"[10] respectively. For the tf-idf baseline, the former also matches word forms such as "Roman" due to the stemming algorithm applied.

A interesting challenge comes into play with this task: the term "Rome" is slightly ambiguous as it can refer specifically to Ancient Rome, but can also mean the city of Rome in general.

The following are the candidate topics for *Ancient Rome* for both manual and semi-automatic topic selection (including relevance scores):

M1 (a) (0.0046) pope, emperor, Gregory, Henry, Rome, VII, king, ban, [...]

(b) (0.0031) Rome, war, Romans, city, king, peace, Gaul, Carthage, year, [...]

(c) (0.0030) Rome, city, war, Tarquinius, king, Roman, Latin, Brutus, son, [...]

(d) (0.0020) Pompey, Caesar, Rome, Csar, year, Spain, Crassus, war, army, [...]

M2 (a) (0.0185) Caesar, Rome, Gaul, Spain, Cicero, Italy, Julius, Pompey, [...]

(b) (0.0178) Rome, Honorius, Italy, Ravenna, Greek, Robert, Apulia, [...]

(c) (0.0174) Hannibal, Rome, Carthage, Italy, Spain, Africa, Carthaginian, [...]

(d) (0.0136) Rome, Italy, Alba, Florence, Naples, Parma, Latium, Modena, Bologna, Tuscany, [...]

M3 None

According to the topic relevance scores, topics **M1**a and **M2**a are selected by the semi-automatic topic selection. However, the human experts ruled out **M1**a as clearly referring to papal matters rather than to Ancient Rome, and selected topic **M1**b instead. This decision has a big impact on the results as shown in table 2.

For the second concept, only one candidate topic was found using the term "uprising":

M1 None

M2 None

M3 (a) (1.0656) uprising, new, violence, revolution, great, murder, war, cruel, top, manner, [...]

Consequently, topic **M3**a is selected both by the manual and by the automatic topic selection process.

For the *Single TM* baseline, the following topic from model **M1** has been chosen manually as most representative for both "Ancient Rome" and "uprising":

- emperor, cruelty, people, Rome, conspiracy, government, life, Christian, year, insurrection, [...]

The sub-corpus on which we measure the recall comprises 302 documents that refer to the *Catiline Conspiracy* which resulted in a civil war and is thus clearly relevant for the posed research task.

Table 2 reports the retrieval results for this task.

	Prec@50	Recall	F1
TMs (manual)	1	0.662	**0.797**
TMs (semi-auto)	0.3	0.01	0.019
tf-idf	0.88	0.108	0.192
Single TM	0.78	0.417	0.543

Table 2: Results for *uprisings* and *Ancient Rome*; recall for the *Catiline Conspiracy* sub-corpus.

5.3 Analysis

The first sample task (section 5.1) demonstrates the strength of our combined topic models approach: the recall compared to the search-based method shows that it retrieves many documents that cannot be found using term search. Analysis of the results shows that false negatives which are not retrieved with tf-idf are relevant for Asia, as expected, but do not explicitly mention any of the search terms "Asia", "China", "India", or "Japan".

The comparison with the single model approach shows that a single topic is not sufficiently specific to cover both

[9]German: *Tsingtau*
[10]German: *Aufstand*

aspects of the research task (*Asia* and *hostilities*). The recall is very high though, due to the fact that most relevant documents are assigned a significant weight for the chosen topic. However, this comes at the expense of a low precision: there are many false positives retrieved that may be relevant for either of the two aspects in the research task, but not for both of them.

For sample task 2 (section 5.2), the results for the manually selected combined topics are even more successful than for the first task: both precision and recall outperform the tf-idf-based term search. This turns out to be due to the ambiguity of the search term "Rome": among the false positives for the tf-idf baseline, there are many documents that deal with Rome in contexts other than Ancient Rome.

However, the semi-automatic topic selection approach fails in this task for the same reason, the ambiguity of the term "Rome": the topic that achieves that highest relevance score for that term does not actually refer to Ancient Rome but to papal matters and Rome in the medieval. As a consequence, most retrieved documents do not refer to Ancient Rome for the *TMs (Combined)* method. Regarding practical applications, our conclusion is that the automatic topic selection can only be used to suggest relevant topics to human users for confirmation. However, fine-tuning the automatic selection process goes beyond the scope of this work.

The single topic model achieves a higher precision for the second research task than for the first one which can be explained by the fact that the manually selected topic unambiguously refers to Ancient Rome.

However, the combination of two specifically selected topics in the combined approach still significantly outperforms the single topic approach.

6. CONCLUSION AND FUTURE WORK

In this work, we demonstrate a method to effectively make use of multiple LDA topic models based on varying parameters and specialized vocabularies. They are combined in an efficient way so that document scores can be computed on-the-fly even for large collections.

In contrast to previous works that aimed at incorporating special aspects such as sentiment analysis [10, 21] into topic modelling, we let history scientists or other researchers making use of large corpora define aspects and foci flexibly, depending on a current task.

In future work, we will focus on increasing the number, quality, and specificity of the single topic models and on more robust (semi-)automatic topic selection. This would allow corpus-based research to define even more specific and more complex retrieval tasks and to identify suitable topics even when the total number of topics is too large for manual selection.

Acknowledgements

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant I/82806, by the German Institute for International Educational Research[11] (*Deutsches Institut für Internationale Pädagogische Forschung*, DIPF), and by the Leibniz Association[12] (*Leibniz-Gemeinschaft*).

[11] `http://www.dipf.de`

[12] `http://www.leibniz-gemeinschaft.de/`

7. REFERENCES

[1] D. M. Blei and J. D. Lafferty. Dynamic Topic Models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 113–120, New York, NY, USA, 2006. ACM.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, Mar. 2003.

[3] 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 22*, pages 288–296, Vancouver, British Columbia, Canada, 2009.

[4] P. DiMaggio, M. Nag, and D. Blei. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics*, 41(6):570–606, Dec. 2013.

[5] R. Eckart de Castilho and I. Gurevych. A broad-coverage collection of portable NLP components for building shareable analysis pipelines. *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*, pages 1–11, 2014.

[6] D. Ferrucci and A. Lally. UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. *Nat. Lang. Eng.*, 10(3-4):327–348, Sept. 2004.

[7] J. R. Finkel, T. Grenager, and C. Manning. Incorporating Non-Local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370, Ann Arbor, Michigan, USA, June 2005. Association for Computational Linguistics.

[8] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.

[9] D. Hall, D. Jurafsky, and C. D. Manning. Studying the history of ideas using topic models. In *Proceedings of the conference on empirical methods in natural language processing*, pages 363–371. Association for Computational Linguistics, 2008.

[10] C. Lin and Y. He. Joint Sentiment/Topic Model for Sentiment Analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pages 375–384, New York, NY, USA, 2009. ACM.

[11] A. K. McCallum. MALLET: A Machine Learning for Language Toolkit, 2002.

[12] D. Mimno and A. McCallum. Organizing the OCA: Learning Faceted Subjects from a Library of Digital Books. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '07, pages 376–385, New York, NY, USA, 2007. ACM.

[13] F. Moretti. *Distant Reading*. Verso, London; New York, 2013.

[14] D. Newman, A. Asuncion, P. Smyth, and M. Welling. Distributed algorithms for topic models. *The Journal of Machine Learning Research*, 10:1801–1828, 2009.

[15] P. Ogren and S. Bethard. Building test suites for UIMA components. In *Proceedings of the Workshop on Software Engineering, Testing, and Quality Assurance for Natural Language Processing (SETQA-NLP 2009)*, pages 1–4, Boulder, Colorado, June 2009. Association for Computational Linguistics.

[16] R. Remus, U. Quasthoff, and G. Heyer. SentiWS - A Publicly Available German-language Resource for Sentiment Analysis. In N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 2010. European Language Resources Association (ELRA).

[17] M. E. Roberts, B. M. Stewart, and E. M. Airoldi. Structural Topic Models. Technical report, Working Paper., 2014.

[18] K. Spärck Jones. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, 28(1):11–21, Jan. 1972.

[19] R. Strötgen. New information infrastructures for textbook research at the Georg Eckert Institute. *History of Education & Children's Literature*, 9(1):149–162, Jan. 2014.

[20] C. Templeton, T. Brown, S. Battacharyya, and J. Boyd-Graber. Mining the Dispatch under Supervision: Using Casualty Counts to Guide Topics from the Richmond Daily Dispatch Corpus. In *Chicago Colloquium on Digital Humanities and Computer Science*, Chicago, Illinois, USA, 2011.

[21] I. Titov and R. McDonald. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web*, pages 111–120, Beijing, China, 2008. ACM.

[22] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1105–1112, Montréal, Québec, Canada, June 2009. ACM.