# My Approach = Your Apparatus?

## Entropy-Based Topic Modeling on Multiple Domain-Specific Text Collections

Julian Risch
Hasso Plattner Institute, University of Potsdam
Prof.-Dr.-Helmert-Str. 2–3
14482 Potsdam, Germany
julian.risch@hpi.de

Ralf Krestel
Hasso Plattner Institute, University of Potsdam
Prof.-Dr.-Helmert-Str. 2–3
14482 Potsdam, Germany
ralf.krestel@hpi.de

## ABSTRACT

Comparative text mining extends from genre analysis and political bias detection to the revelation of cultural and geographic differences, through to the search for prior art across patents and scientific papers. These applications use cross-collection topic modeling for the exploration, clustering, and comparison of large sets of documents, such as digital libraries. However, topic modeling on documents from different collections is challenging because of domain-specific vocabulary.

We present a cross-collection topic model combined with automatic domain term extraction and phrase segmentation. This model distinguishes collection-specific and collection-independent words based on information entropy and reveals commonalities and differences of multiple text collections. We evaluate our model on patents, scientific papers, newspaper articles, forum posts, and Wikipedia articles. In comparison to state-of-the-art cross-collection topic modeling, our model achieves up to 13% higher topic coherence, up to 4% lower perplexity, and up to 31% higher document classification accuracy. More importantly, our approach is the first topic model that ensures disjunct general and specific word distributions, resulting in clear-cut topic representations.

## CCS CONCEPTS

•**Information systems** →**Document topic models;** *Document collection models;* •**Applied computing** →*Document searching;*

## KEYWORDS

Topic modeling; Automatic domain term extraction; Entropy

**ACM Reference format:**
Julian Risch and Ralf Krestel. 2018. My Approach = Your Apparatus?. In *Proceedings of The 18th ACM/IEEE Joint Conference on Digital Libraries, Fort Worth, TX, USA, June 3–7, 2018 (JCDL '18),* 10 pages.
DOI: 10.1145/3197026.3197038


978-1-4503-5178-2/18/06…$15.00
DOI: 10.1145/3197026.3197038

## 1 CROSS-COLLECTION TOPIC MODELS

A variety of information retrieval and data mining applications deals with unstructured text data. With unsupervised machine learning, topic models iteratively estimate probabilistic representations of topics and documents. Based on word co-occurrence frequencies, topic models cluster text documents by their latent topics and thus structure large document collections. Despite the overwhelming, unstructured amount of data, topic models thereby enable users to search for documents by topic, to explore document collections, and to extract meaningful information.

As digital libraries grow, getting an overview and keeping track of large document collections becomes even more important. However, combining the knowledge from multiple collections is challenging. Linguistic contrasts, such as domain-specific vocabulary, complicate topic modeling. Cross-collection topic models extend previous single-collection models to multiple collections. They aim to model document-topic representations despite linguistic contrasts and to reveal per-topic similarities and differences of collections.

An exemplary application of cross-collection topic modeling is the search for prior art during the patent examination process [24]. For this task, related work from any publicly available text collection, such as granted patents or scientific papers, needs to be retrieved. Patent-specific words, such as "device" or "apparatus" and paper-specific words, such as "algorithm" or "approach", hamper the effective usage of keyword-based approaches and word co-occurrence statistics.

Another field of application is bias detection in newspapers. For this task, we consider each newspaper's articles as an individual collection. We distinguish collection-independent and collection-specific words based on their frequency distribution: Collection-independent words have similar frequency across all collections, whereas collection-specific words have significantly different frequency per collection. Still, collection-specific words might occur in all collections, but they occur much more frequently in one particular collection compared to other collections. In newspapers, collection-specific words serve as a bias indicator, because these words occur more frequently in one newspaper. Further bias analysis then focuses on interpretation of these differences in language use. Similarly, cross-collection topic models can identify linguistic differences in document collections with different cultural and regional background. For example, topic modeling on traveler forum posts about different countries aims to reveal country-specific words per topic and regional and cultural differences in forum posts.

As opposed to previous work, we focus on domain-specific language and consider collections as domains with specific vocabulary. We propose a novel topic model that combines state-of-the-art cross-collection topic modeling [23] with the concept of termhood from the field of automatic domain term extraction [4]. With the help of an entropy-based termhood measure, our model ranks words according to their collection-specificity. To split the vocabulary in a set of collection-specific words and collection-independent words, the model sets an entropy threshold and estimates the proportion of collection-specific words for each dataset individually and automatically. As a consequence, the topic model guarantees a clear-cut separation of words with collection-specific and collection-independent frequency distribution. A mixture of collection-specific and collection-independent word distributions represents each latent topic. The precise distinction of collection-specific and collection-independent words is a novelty in cross-collection topic modeling. Furthermore, in order to resolve semantic ambiguity of single words, our topic model considers also multi-word phrases. We evaluate our model and a state-of-the-art approach with regard to three quality measures: topic coherence, language model perplexity, and document classification accuracy. The evaluation is based on four datasets with either two or three collections.

Section 2 presents related work in the fields of probabilistic topic models and automatic domain term extraction. We present our topic model together with the entropy-based termhood estimation, the entropy threshold definition, and the estimation procedure of the proportion of collection-specific words in Section 3. Section 4 describes the experiment setup and compares the proposed model with state-of-the-art cross-collection topic modeling. We conclude with a summarization of our contributions and paths for future work in Section 5.

## 2 RELATED WORK

*Latent Dirichlet Allocation.* Being the standard topic model for single text collections, latent Dirichlet allocation (LDA) models each document as a probability distribution over topics and each topic as a probability distribution over words [3]. Gibbs sampling [11] is used for the estimation of these latent distributions, as it is less complex to implement and achieves up to two orders of magnitude faster runtime compared to variational Bayes and expectation propagation [30]. While LDA is based on the assumption that documents with similar topic distribution exhibit similar word distribution, documents from different collections use collection-specific words and thus have different word distributions. As a consequence, LDA is not suited for modeling topics on multiple collections with domain-specific vocabulary.

*Cross-Collection Latent Dirichlet Allocation.* Comparative text mining extends text mining techniques to more than one document collection for the purpose of revealing collection's similarities and differences. The cross-collection mixture model (ccMix) discovers common topics across collections of news articles from different publishers and of product reviews from different companies [31]. To extend topic modeling to multiple collections, ccMix draws each word either from a topic's collection-independent or a topic's respective collection-specific word distribution. However, ccMix uses only a single, user-defined parameter as the probability that words are collection-independent or collection-specific. Therefore, ccMix cannot distinguish collection-specific and collection-independent words precisely. Instead of a single, user-defined parameter, cross-collection LDA (ccLDA) learns a probability distribution of collection-independent and collection-specific words per topic and per collection [23]. Applied to traveler forum posts about the UK, Singapore, and India, ccLDA identifies topics about "food" or "weather" alongside with collection-specific (region-specific) words per topic, such as "masala" and "seafood" or "monsoon" and "snow". An alternative to collection-specific topic distributions is proposed by Eisenstein et al.: Their approach models a background word distribution and the differences in log-frequencies from this distribution for collection-specific words [6].

More recently, supervised cross-collection topic models have been proposed for cross-domain learning, extending LDA and ccLDA [2, 10]. Interestingly, no topic model so far utilizes a word's frequency of occurrence across all collections in order to determine whether the word is collection-specific or not. They all lack the ability to precisely identify a word as collection-independent if it occurs with similar frequency of occurrence across all collections. Counterintuitively, in ccMix, ccLDA, and subsequent models, a word can be simultaneously part of a collection-independent and a collection-specific word distribution. In contrast, our approach guarantees a clear-cut distinction of collection-specific and collection-independent words.

*Multi-Lingual and Other Cross-Collection Topic Models.* Multi-lingual topic modeling deals with stronger linguistic differences between document collections. Zhang et al. model topics across different languages by incorporating a bilingual dictionary into a probabilistic topic model [32]. While their approach requires a dictionary, our experiments on multi-lingual Wikipedia articles demonstrate that we are able to model multi-lingual topics without a dictionary. LDA can be extended to multiple collections by running several LDA instances in parallel, unified by manually identifying common topics [34] or with a hierarchical model of per-topic word distributions [5]. On a newspaper dataset with different regions as collections, the hierarchical model learns a master topic across all regions and region-specific descendant topics. However, the model does not distinguish collection-specific from collection-independent words and descendant topics across different regions are almost identical for generic topics. In contrast, our approach guarantees for each topic that the collection-independent topic representation has no word in common with each collection-specific topic representation. Another contribution of Zhang et al. focuses on asymmetric and weakly-related collections, extending ccLDA with collection-specific topics [33]. In contrast, we focus on collection-specific words in common topics.

Fang et al. identify verbs, adjectives, and adverbs as differing opinion words in three news collections per topic [8]. They model only one word distribution per topic and assume that different collections share many words per topic and that only opinion words differ. This assumption does not hold for collections with strong linguistic contrasts and especially not for multi-lingual datasets. Furthermore, their model does not distinguish collection-specific and collection-independent opinion words. To the best of our knowledge, no previous work models topics on patents and scientific papers simultaneously. However, there is related work that models topics of these collections separately, such as a topic-model-based recommender system for prior art in patents [18] and a topic model on abstracts of scientific papers [11].

*Automatic Term Extraction.* The field of automatic (domain) term extraction (ATE) deals with the extraction of words or compounds of multiple words that are considered domain-specific terms in text documents. To this end, Kageura and Umino define termhood as "the degree that a linguistic unit is related to domain-specific concepts" [15]. Several papers address the extraction of domain-specific terms by identifying term candidates based on part-of-speech patterns and ranking them by their termhood afterwards. It is a common assumption that: "The information that a term candidate carries is also an important indicator of its termhood." [17]. Therefore, term candidates are ranked according to their termhood, which is measured with variants of term frequency-inverse document frequency (TF-IDF).

In contrast to TF-IDF, Inter-Domain Entropy (IDE) considers the distribution of a word's relative term frequency across all domains [4]. The closer this distribution is to a uniform distribution, the lower is a word's termhood. While IDE provides a ranking of words according to their termhood, the ranked list needs to be split into domain-specific and domain-independent word sets based on a threshold, which varies for different datasets. However, most previous work arbitrarily considers the top-10% or top-100 words to be specific terms or sets thresholds empirically [4, 9]. Instead of statistical information about word frequencies, Li et al. incorporate semantic information from learned latent topics into a novel termhood measure [20]. On single collections, Wilson and Chew extend the standard LDA model with different term weighting schemes [29]. Thereby, the dominance and scattering of high-frequency words can be controlled and stop word removal becomes obsolete. In our work, we combine topic modeling and domain term extraction to model multiple document collections and their linguistic characteristics.

*Topic Models on Phrases.* In order to reduce semantic ambiguity, text mining applications can take multi-word phrases into account (instead of single words only). For example, the unigrams "support", "vector", and "machine" have a different meaning if they are processed as a phrase and therefore topic models based on both unigrams and bigrams outperform unigram topic models at information retrieval tasks. [28]. Topical phrase mining has been applied to model medical terms [12], combined with topical change over time [14], and scaled to large datasets [7]. With a supervised approach, Kawamae et al. investigate the relationship between training labels and corresponding phrases [16]. While phrases can be modeled as a hierarchy of Pitman-Yor processes [21], Lau et al. show that also n-gram tokenization as a pre-processing step improves topic quality [19]. Recently, a data-driven approach automates phrase segmentation with robust performance at different domains [22, 26]. Hence, topic modeling and phrase segmentation can be applied in independent, automatic steps. We include automatic phrase segmentation in our proposed topic model during the pre-processing step of tokenization.

# 3 ENTROPY-BASED TOPIC MODELING

Our approach combines ccLDA as a cross-collection topic model with an entropy-based measure of termhood. As a result, we propose a novel topic model that splits the vocabulary in collection-specific and collection-independent words and provides more meaningful topic mixture representations of documents. Collection-specific and collection-independent words form each topic's representation and reveal commonalities and differences of the collections.

## 3.1 Basic Cross-Collection Model

Our model and ccLDA have in common that they both contain per-topic collection-independent word distributions $\varphi$ and per-topic and per-collection collection-specific word distributions $\sigma$. Furthermore, there are per-document topic distributions $\theta$. These three distributions have Dirichlet priors $\alpha$, $\beta$, and $\delta$. Which word $w$ is sampled in a document depends on the document's collection $c$, the topic $z$ of word $w$, and the binary decision variable $x$. $x$ determines whether $w$ is sampled from a collection-independent or a collection-specific word distribution. The original ccLDA samples $x$ from a Bernoulli distribution for each occurrence of a word. As a consequence, out of two different occurrences of the same word, one occurrence might be considered collection-specific and the other collection-independent.

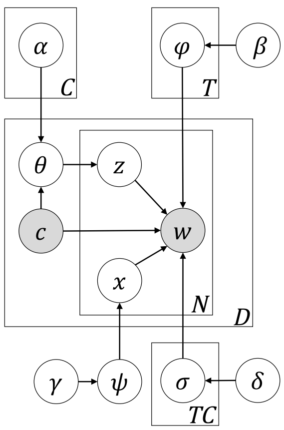
## 3.2 Extended Cross-Collection Model

The main difference compared to ccLDA is the following: ccLDA determines separately for each occurrence of a word whether this occurrence in particular is collection-specific, whereas our model determines globally whether a word is collection-specific for all its occurrences across all documents. If our model assumes that a word is collection-specific, this assumption holds for every document in which the word occurs. In contrast, if ccLDA assumes that a word is collection-specific, another occurrence of the same word in the same sentence could be considered collection-independent. Furthermore, ccLDA learns a per-topic and per-collection probability distribution from which the model randomly samples whether a particular occurrence of a word is collection-specific. In contrast, our model does not sample this property randomly but uses a precise differentiation of termhood on the vocabulary level.

According to an entropy-based termhood measure, the vocabulary is split into two sets: collection-specific and collection-independent words. With regard to this measure, $x$ is sampled from a Bernoulli distribution $\psi$ with the parameter $\gamma$. $\gamma$ corresponds to the proportion of occurrences of collection-specific words in the entire dataset and is described together with the entropy-based termhood measure in Section 3.3. Via the Bernoulli distribution $\psi$ and the proportion of collection-specific words $\gamma$, the entropy-based termhood measure determines the mixture of words from $\varphi$ and $\sigma$ across all documents. Section 4 gives examples of how this conceptual difference causes significant change in resulting topics. Figure 1 depicts the graphical model of our approach and the corresponding generative process is as follows:

(1) Draw a collection-independent multinomial word distribution $\varphi_z$ from Dirichlet($\beta$) for each topic $z$.

(2) Draw a collection-specific multinomial word distribution $\sigma_{z,c}$ from Dirichlet($\delta$) for each topic $z$ and each collection $c$.
(3) For each document $d$, choose a collection $c$ and draw a topic distribution $\theta_d$ from Dirichlet($\alpha$).
  (a) For each word $w$ in $d$, draw a topic $z$ from $\theta_d$.
  (b) Draw $x$ from Bernoulli distribution $\psi$ according to proportion of collection-specific words $\gamma$.
  (c) If $x = 0$, draw $w$ from $\varphi_z$; else from $\sigma_{z,c}$[1].




| | |
|---|---|
| $c$ | collection |
| $w$ | word |
| $x$ | determines whether $w$ depends on $\sigma$ or $\varphi$ |
| $z$ | topic |
| $C$ | number of collections |
| $T$ | number of topics |
| $D$ | number of documents |
| $N$ | number of words/doc. |
| $\alpha, \beta, \delta$ | Dirichlet priors |
| $\gamma$ | proportion of collection-specific words |
| $\theta$ | topic distribution |
| $\sigma$ | collection-specific word distribution |
| $\varphi$ | collection-independent word distribution |
| $\psi$ | Bernoulli distribution of variable $x$ |

**Figure 1: Graphical representation of the proposed entropy-based topic model.**

### 3.3 Cross-Collection Word Entropy

In this paper, we distinguish collection-specific and collection-independent words based on entropy. For each word, the entropy of its frequency distribution across all collections measures the word's termhood. Intuitively, the termhood shall be low for words that are evenly distributed across documents from all collections. In contrast, the termhood shall be high for words that occur more frequently in documents of a specific collection. Together with an entropy threshold, the termhood determines whether a word is collection-specific. In our case, a random variable $X$ represents the collection $c$ of a given word $w$, where $P(X = c) = P(c|w)$ holds. In the following formulas, we consider a word $w$ as an entry in the vocabulary and not as a particular occurrence of a word in a document. Thus, we define the normalized entropy $H(w)$ as:

$$H(w) = \frac{1}{log_2 C} \sum_{c=1}^{C} -P(c|w) \cdot log_2 P(c|w).$$

If $w$ is uniformly distributed across all collections, $P(c|w)$ is equal for all $c \in [1, C]$ and $\sum_{c=1}^{C} -P(c|w) \cdot log_2 P(c|w)$ reaches its maximum, $log_2 C$. To obtain an entropy value in the interval of [0,1] that reaches its maximum for collection-specific words, we normalize entropy with the factor $log_2 C$. We consider $P(c|w)$ as the probability that the document's collection is $c$, given the occurrence of word $w$. Because the probability $P(c|w)$ is unknown a priori, we estimate this frequency on training documents before the topic sampling process.

### 3.4 Estimation of Word Probabilities

In order to estimate posterior $P(c|w)$, we estimate evidence $P(w)$, prior $P(c)$, and likelihood $P(w|c)$ on training documents:

- $P(w)$ is the probability that word $w$ is randomly chosen from any document in the entire dataset.
- $P(c)$ is the probability that a randomly chosen word is from a document of collection $c$.
- $P(w|c)$ is the probability that word $w$ occurs in a randomly chosen document of collection $c$.

Due to the inherent sparsity of limited training data, some words occur only once in the entire corpus. With only a single observation of such word, called hapax legomena, there is only limited knowledge about this word's true frequency distribution across collections. Although hapax legomena have been never observed in any other collection, they might have a non-zero probability of occurrence. For this reason, we use Laplace smoothing in order to create pseudocounts for unobserved words.

Based on entropy and the corresponding termhood[2], we can sort words according to their estimated collection-independence. To incorporate this sorted vocabulary into the topic model, we define an entropy threshold that splits the vocabulary in a set of collection-independent words and a set of collection-specific words. The next section describes how to estimate this threshold as a hyperparameter for arbitrary datasets automatically.

### 3.5 Estimation of the Entropy Threshold

We rank every word according to its entropy-based termhood. On the one hand, there are certainly collection-specific words, which occur many times in each document of one collection and never in any document of the other collection. On the other hand, there are certainly collection-independent words, which occur often and with equal frequency in each document, independently of the collection. In between, the uncertainty is highest for words with the fewest observations: hapax legomena.

We set the entropy threshold such that hapax legomena are closest to it. This threshold corresponds to the highest uncertainty whether a word is collection-specific or collection-independent. Even if a corpus contains no hapax legomenon at all, this threshold can be calculated. Imagine, there is a hapax legomenon $w_{hl}$ that occurs exactly once in collection $c_1$ and never in collection $c_2$ in a dataset consisting of two collections. After Laplace smoothing (adding one pseudocount occurrence to each collection), $w_{hl}$ has two occurrences in $c_1$ and one occurrence in $c_2$. Thus, we can estimate $P(c_1|w_{hl}) = \frac{2}{3}$ and $P(c_2|w_{hl}) = \frac{1}{3}$. The entropy $H_{w_{hl}}$ follows as:

$$H_{w_{hl}} = \sum_{i \in \{1,2\}} -P(c_i|w_{hl}) \cdot log_2 P(c_i|w_{hl})$$

For an arbitrary number of collections $n$, collections $c_1, c_2, \ldots, c_n$, $P(c_i|w_{hl}) = \frac{2}{n+1}$ for $i = 1$, and $P(c_i|w_{hl}) = \frac{1}{n+1}$ for all integers $i$,

[1]The conceptual difference in distinction from ccLDA is to draw $w$ from *disjoint* $\varphi_z$ and $\sigma_{z,c}$ according to *entropy-based* $\gamma$.

[2]Termhood of word $w$ corresponds to $1 - H(w)$

$1 < i \leq n$, this formula generalizes to:

$$H_{w_{hl}} = \sum_{i \in \{1,2,\ldots,n\}} -P(c_i|w_{hl}) \cdot log_2 P(c_i|w_{hl})$$

According to Zipf's law for natural language corpora, a word's frequency is inversely proportional to its rank in frequency. Thus, most words in a corpus occur exactly once and by definition are hapax legomena, having entropy value $H_{w_{hl}}$. As a consequence, we can expect this value to be the most frequent entropy value in a dataset. We show that this holds in our four evaluation datasets in Section 4.3.

If the entropy threshold is set below this most frequent value, all hapax legomena are collection-independent. If the threshold is set above this value, all hapax legomena are collection-specific. As a consequence, a small change of the entropy threshold results in a large shift of the proportion of collection-specific words. Notice that Laplace smoothing has been applied already. Otherwise, a word that occurs exactly once would be collection-specific with certainty 100%. By applying Laplace smoothing, certainty for being collection-specific ranges from 0 (inclusive) to 1 (exclusive).

Because hapax legomena have been observed once in one collection and never in any other collection, we consider them as collection-specific words. Topic representations list only the most frequent words per topic and therefore, hapax legomena typically do not appear in these representations.

$H_{w_{hl}}$ splits the vocabulary into two sets according to the entropy ranking, thus defining a set of collection-specific and a set of collection-independent words. From the sizes of these sets and word frequency observations in the training dataset, we estimate the hyperparameter $\gamma$, which corresponds to the proportion of collection-specific words in the dataset. $\gamma$ is estimated by the sum of occurrences of all collection-specific words divided by the total number of occurrences of words in the dataset.

For the estimation of word distributions and topic distributions, we use Gibbs sampling based on the equations described in the original ccLDA paper [23]. Our implementation of the entropy-based cross-collection model is open-sourced online[3].

### 3.6 Further Considerations

*Arbitrary Number of Collections.* For two collections, the definition of collection-specific words is intuitive: these words occur most likely in one collection and unlikely in the other. For more than two collections, for example three, this definition is less intuitive: If a word occurs once in each of two collections but not in the third collection, is this word collection-specific? We compute the entropy of a hapax legomenon as the entropy threshold, which separates collection-specific and collection-independent words. Based on this threshold and our definition of collection-specificity, we consider words that occur once in two collections but not in the third collection as collection-independent, because their entropy is higher than the entropy of a hapax legomenon. The broader definition of collection-specificity is justified, because it considers a word as collection-specific if its frequency distribution across collections differs from a uniform distribution with larger extent than a hapax legomenon's distribution. We evaluate our model on datasets with two and with three collections in Section 4.6 and show examples for collection-specific words.

*Multi-Lingual Corpora.* Our entropy-based topic model is suited for multi-lingual corpora as well. If a corpus contains collections of two different languages, the linguistic differences are especially strong. In this case, collection-independent words are words that occur in documents of either language, such as named entities or loanwords. Collection-specific words are language-specific words that have no correspondence with the same spelling in the other language. Our topic model provides topic-wise language-specific and language-independent words, which makes the results of our work interesting for linguists and machine translation. In Section 4.6, we evaluate our model's multi-lingual capabilities prototypically on English, French, and German Wikipedia articles.

*Multi-Word Phrases.* To reduce semantic ambiguity of single words, we tokenize multi-word phrases during text pre-processing. For an automatic phrase segmentation, we use the AutoPhrase algorithm of Liu et al. [26]. Therefore, we process each dataset with their phrase mining approach and segment multi-word phrases as single tokens automatically. No labeled input phrases are required.

*Application to ATE.* To use our model for the automatic extraction of domain terms, domains correspond to collections. As a result after training our topic model, the most frequent top-k words from the collection-specific word distribution correspond to domain terms. In the field of ATE, typical values for $k$ are 10 to 100 [9]. An application to automatic term extraction would profit from multi-word phrases and their potential to reduce semantic ambiguity.

## 4 EVALUATION

To compare ccLDA with the entropy-based model, we evaluate accuracy of document classification, topic coherence, and perplexity on datasets consisting of either two or three collections from different domains.

### 4.1 Datasets

We present four datasets, which differ in domain, language, number of documents, average document length, and number of collections. Table 1 gives an overview of the dataset sizes.

The patents-papers dataset focuses on the domain of computer science and contains abstracts of ACM papers and U.S. patents. The patents[4] have been published by the United States Patent and Trademark Office (USPTO) between 2001 and 2016 and the papers[5] by Tang et al. [27]. We consider only those patents that contain the term "ACM" in the citations. To show that our approach generalizes to other document types, we evaluate our model on two large British newspapers: The Guardian and The Telegraph. The dataset of articles published in the politics category between 2010 and 2015 forms the largest dataset in our evaluation.

The third dataset consists of Wikipedia articles that have an English, a French, and a German version. We crawled English, French, and German Wikipedia articles about movies produced between 2000 and 2016. The length of such articles differs heavily,

[3] https://hpi.de/naumann/projects/web-science/cross-collection-text-mining/entropy-based-topic-modeling.html

[4] https://www.uspto.gov/learning-and-resources/electronic-bulk-data-products

[5] https://aminer.org/citation

**Table 1: Collections with their respective number of documents $D$ and average number of words per document $W/D$ (after stop word removal).**

| Dataset | Collection | D | W/D |
|---|---|---|---|
| Patents-Papers | Patents | 14031 | 71 |
| | Papers | 16998 | 85 |
| Newspapers | Guardian | 30774 | 310 |
| | Telegraph | 30749 | 273 |
| Wikipedia | English | 2927 | 324 |
| | French | 2927 | 291 |
| | German | 2927 | 302 |
| Traveler Forum | India | 1432 | 199 |
| | Singapore | 1179 | 187 |
| | UK | 1580 | 288 |

because the multi-lingual versions are no direct translations of each other. Therefore, for each triple of an English, a French, and a German Wikipedia article, we calculate the minimum number of words and reduce all three articles to this text length. If the minimum number of words is less than 300 (before stop word removal), the articles are discarded to ensure that the topic model learns only on articles of sufficient length. Longer texts make sure that there are enough topic-specific words to learn a document's topics.

From the paper that proposes ccLDA, we reuse a dataset[6] crawled from the online platform lonelyplanet.com [23]. At this platform, there are three separate forums for Singapore, India, and the UK, each having thousands of threads. Each document in Paul's dataset is the concatenation of all messages in one forum thread.

Table 2 lists the four datasets with their natural number of topics, entropy thresholds (depending on the number of collections), and proportion of collection-specific words $\gamma$. We determine the natural number of topics per dataset with a preliminary experiment following Arun et al.'s approach [1]. Their approach views LDA as a matrix factorization method and determines the natural number of topics for a given dataset by minimizing KL-Divergence of Singular value distributions. The hyperparameter $\gamma$ is estimated as described in Section 3.5 and gives a first impression of the linguistic contrasts to expect in each dataset. Not surprisingly, the proportion of collection-specific words is highest in the multi-lingual Wikipedia dataset, whereas it is lowest in newspaper articles. Regarding the number of topics, patents and papers have the by far highest topical diversity.

## 4.2 Experimental Setup

The pre-processing consists of phrase segmentation, tokenization of multi-word phrases as single tokens[7], and removal of stop words based on stop word lists. The topic distribution priors, $\alpha$, are fixed and uniform except for one background topic. According to Paul and Girju, updating $\alpha$ or other hyperparameters at runtime does not largely affect the sampling procedure [23]. We set $\beta$ and $\delta$ to 0.01 and $\gamma_0$ and $\gamma_1$ to 1.0, corresponding to symmetric distributions with equal probability of occurrence of collection-specific and collection-independent words. The Gibbs sampling runs for a burn-in period of 200 iterations. 10 samples, separated by lags of 10 iterations, are averaged for the final result.

After the sampling process, we measure per-topic coherence separately for the collection-independent word distribution $\varphi$ and each collection-specific word distribution $\sigma$. Furthermore, we measure mixed topic coherence, language model perplexity, and document classification accuracy. We split each dataset into 90% training set and 10% test set and use 10-fold cross-validation.

**Table 2: Number of topics, entropy thresholds, and respective estimated hyperparameters $\gamma$ (proportion of collection-specific words).**

| Dataset | #Topics | Entropy | $\gamma$ |
|---|---|---|---|
| Patents-Papers | 260 | 0.918 | 0.632 |
| Newspapers | 155 | 0.918 | 0.184 |
| Wikipedia | 25 | 0.946 | 0.784 |
| Traveler Forum | 100 | 0.946 | 0.742 |

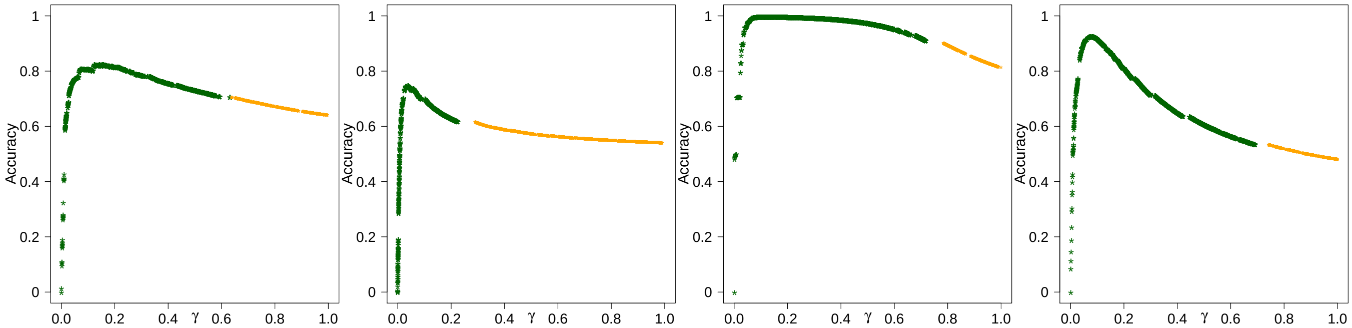
**Table 3: Topic model comparison regarding accuracy (Acc), topic coherence (TC), and perplexity (Perpl).**
**Asterisks denote statistically significant improvement as determined by a paired two-tailed t-test at 0.05 level.**

| | ccLDA | | | Entropy-Based | | |
|---|---|---|---|---|---|---|
| Dataset | Acc | TC | Perpl | Acc | TC | Perpl |
| Patents-Papers | 0.61 | 0.404 | 827 | 0.70* | 0.412* | 797* |
| Newspapers | 0.54 | 0.436 | 4151 | 0.61* | 0.493* | 4042* |
| Wikipedia | 0.70 | - | 5150 | 0.92* | - | 4955* |
| Traveler Forum | 0.45 | 0.390 | 771 | 0.58* | 0.413* | 785 |

## 4.3 Document Classification

The evaluation of the separation of collection-specific and collection-independent words is based on a document classification task. For this task, each topic model predicts the collections of test documents given the documents' words. Rather than only outputting the most likely collection per document, topic models assign a probability to each collection. This probabilistic classification allows a more detailed evaluation regarding each topic model's degree of certainty. The document classification accuracy corresponds to the probability assigned to the correct collection normalized by the sum of probabilities assigned to all collections. We calculate the average accuracy across all test documents to obtain a document classification accuracy for each topic model. This evaluation task allows for a comparison of ccLDA and the proposed topic model.[8]

[6] http://cmci.colorado.edu/~mpaul/downloads/ccdata.php
[7] To ensure a fair comparison, we incorporate multi-word phrases also into the ccLDA baseline approach.
[8] We do not consider document classification approaches as baselines, because the goal of this evaluation is to compare different cross-collection topic models.




**(a) Patents-Papers** **(b) Newspapers** **(c) Wikipedia** **(d) Traveler Forum**

**Figure 2: Document classification accuracy for all possible proportions of collection-specific words, $\gamma$. Collection-specific words in our entropy-based model are colored in green, collection-independent words in yellow.**

CcLDA calculates the probability $P(c|d)$ for collection $c$ given document $d$ as:

$$P(c|d) = \prod_{w \in d} \sum_{c} \mathbb{L}(w|\theta_d, c)$$

We use the word likelihood $\mathbb{L}(w|\theta_d, c)$ and compute the product of the likelihoods of all words in a document per collection $c$. The classification accuracy for document $d$ is $\frac{P(c_{correct}|d)}{\sum_c P(c|d)}$.

Table 3 lists all evaluation results. On the Wikipedia dataset, document classification accuracy of our entropy-based model is 31% higher than ccLDA's accuracy. On the other datasets, we achieve 29% (traveler forum), 15% (patents-papers), and 13% (newspapers) higher accuracy. Especially on datasets with stronger linguistic contrasts, such as the multi-lingual Wikipedia dataset, our approach outperforms ccLDA and provides a better separation of collection-specific and collection-independent words. Comparing the different datasets, both approaches achieve the highest accuracy on the Wikipedia dataset. This dataset contains many language-specific words. It confirms that language classification of Wikipedia articles is easier than, for example, classification of newspaper articles as Guardian or Telegraph articles.

Figure 2 visualizes the influence of hyperparameter $\gamma$ on the document classification accuracy of our model. The number of possible, meaningful hyperparameter settings is the number of distinct entropy values for words in the entire vocabulary. Thus, vocabulary size $V$ is an upper limit for the number of thresholds. Each possible threshold corresponds to one hyperparameter setting and splits the vocabulary into two sets of collection-specific and collection-independent words. For each possible threshold, we compute the topic model's document classification accuracy. The larger the proportion of collection-specific words, the more words our model uses to classify a document. The largest gap in each graph corresponds to hapax legomena. We can either choose to consider all hapax legomena as collection-specific words or consider all hapax legomena as collection-independent words. There is no choice in between, because all these words have the same entropy value. The smaller gaps in the graph correspond to words that occur exactly twice (dis legomena), three times (tris legomena), and so on. These words make up a large proportion of all words in the corpus, which is why their entropy value is more frequent.

Additionally, Figure 2 shows that we could achieve the highest document classification accuracy with $\gamma \approx 0.1$. However, such $\gamma$ considers words as collection-specific only if they improve document classification accuracy. Incorrectly, the word's frequency distribution would not be considered. For example, consider one word that occurs in every document of one collection and never in the other collection. This single, collection-specific word achieves optimal document classification accuracy. Thus, all other words would be incorrectly considered collection-independent regardless of their frequency distribution.

## 4.4 Topic Coherence

The evaluation of topic coherence compares ccLDA and the entropy-based model with regard to their capability to cluster words by semantic similarity within one collection and across multiple collections. Especially, we evaluate the capability to correctly align topics of different collections despite different word distributions. For automatic evaluation of topic coherence, we use the Palmetto[9] library, which implements the $C_V$ measure [25].

However, current topic coherence measures consider only single word distributions per topic and their topic representation in the form of top words. Current measures cannot handle multiple word distributions per topic. Therefore, we present an extension of topic coherence measures for cross-collection topic models, which we also published online. Our approach considers topic representations (which are sets of top words) of the collection-independent word distribution $\varphi$ and each collection-specific word distribution $\sigma$. Per topic, we use the union of these representations as a single topic representation across the collection-independent and each collection-specific topic-word distribution. The coherence of this union can be measured with current topic coherence measures. We call this measure "mixed topic coherence", because it mixes the word distributions and thereby allows to evaluate the topical alignment of the different word distributions.

However, measuring the topic coherence based on word co-occurrences has its limitations. The $C_V$ measure assumes that words that never co-occur in the reference dataset are not topically coherent. This assumption does not hold for datasets with

[9]https://github.com/AKSW/Palmetto/

strong linguistic contrast. For example, words from collections with different languages do not co-occur although they can form a coherent topic. Therefore, the $C_V$ measure cannot evaluate topic coherence of topic models trained on our multi-lingual dataset. For all other datasets, we obtain a single coherence value per topic model by averaging all its topics' coherence values. In the evaluation, ccLDA and the proposed entropy-based model form almost equally coherent topic representations with slightly higher topic coherence of our approach, especially on the newspapers dataset (13% improvement).

## 4.5 Perplexity

With test set perplexity, we evaluate a topic model's ability to generalize from training data to unseen test data. Perplexity focuses on the generative aspect of topic models to predict word probabilities for unseen documents. Lower perplexity is better.

We use the "fold-in" method used by Paul et al. for the evaluation of ccLDA to learn topic probabilities $\theta$ of test documents. "Fold-in" means that the evaluation keeps all topic model parameters fixed as they have been learned on the training data. Gibbs sampling estimates only the per-document topic distributions on the test set. This method has been introduced by Hofmann [13]. Given the topic probabilities $\theta_d$ and the collection $c$ of a document $d$, the likelihood $\mathbb{L}$ of a word $w$ in $d$ is calculated as:

$$\mathbb{L}(w|\theta_d, c) = \sum_z P(z|\theta_d)[P(\neg x)P(w|z, \neg x) + P(x)P(w|z, c, x)].$$

In this formula, $x$ is a binary variable that denotes whether $w$ is collection-specific ($x$) or collection-independent ($\neg x$). In ccLDA, this variable depends on collection $c$ and topic $z$ but not on the word itself. In contrast, our model estimates $x$ for each word in the vocabulary and thus learns the proportion of collection-specific words $\gamma$ per dataset.

$P(x)$ is the probability that a collection-specific word is sampled and is equivalent to $\gamma$ in our approach. $P(w|z, x)$ is the probability that in particular word $w$ is sampled if a collection-specific word from topic $z$ is sampled. The perplexity of a model $m$, with $M$ being the total number of words in all test documents, is calculated as:

$$\mathbb{P}(m) = 2^{-\frac{1}{M}\prod_w \mathbb{L}(w|\theta_d, c)}.$$

Both models achieve comparable perplexity with slight advantages of the entropy-based model, especially on the patents-papers dataset (4% improvement). On the traveler forum dataset, ccLDA achieves lower perplexity than the entropy-based model, although this difference is not significant.

## 4.6 Example Topics

As an example, Table 4 shows the top-5 words of each specific and independent word distribution our model learned on the patents-papers dataset. For reasons of clarity, we limit the number of topics to 25. Topic 0, the stop word topic, correctly identifies patent-specific stop words, such as "apparatus" and paper-specific stop words, such as "approach". The word "datum", for example, is represented in the patent-specific and paper-specific word distributions, which means its collection-specific frequencies differ significantly and therefore cannot be combined in a single, collection-independent frequency. Collection-independent words, such as "method" and "system", occur with similar frequency in both collections. Another example, topic 5, deals with computer graphics (collection-independent top words) and our model shows that the focus of patents ("pixel" or "display") is rather low-level, while papers focus on high-level algorithmic problems in the context of image processing ("algorithm", "detection", "recognition"). Topic 2 is a security topic and its collection-specific words reveal that most documents about computer "attacks" are scientific papers, but no patents.

The original ccLDA is based on word co-occurrences only, which is particularly problematic if documents across collections share only few words, since they are then considered unlikely to belong to the same topic. English, French, and German Wikipedia movie articles, for example, share less than 10% of their words. For comparison, the Guardian and the Telegraph articles share 37%. As a consequence, ccLDA exhibits poor topic coherence and topic alignment across multi-lingual collections.

Table 5 compares two Indian movie topics learned by the entropy-based model and ccLDA. Both models are capable to assign English, French, and German words to respective collection-specific word distributions. However, ccLDA considers "khan" to be the most frequent word of the collection-independent distribution and the English-specific distribution. This mistake reveals ccLDA's deficiency to separate collection-specific and collection-independent words correctly. In contrast, the entropy-based model guarantees that words are either collection-specific or collection-independent. The word "bollywood" is not collection-independent, because its frequency is not uniformly distributed across all collections. Instead, the word occurs more frequently in French articles with this topic (1.1%) than in English articles (0.7%). Table 5 further indicates that ccLDA does not align topics correctly: The German-specific word distribution falsely contains sports related words and some words are not even German.

Table 6 lists two topics about the military conflict in Syria and Iraq learned by ccLDA and the entropy-based model on the newspapers dataset. The entropy-based model provides a better separation of collection-specific and collection-independent words. For example, the entropy-based model assigns "international" to both collection-specific word distributions, because the word occurs with significantly different frequencies in both collections, which also holds for "iraq", "syrian", and "assad". In both collection-specific word distributions, "assad" is the seventh most frequent word, but the difference of relative frequencies (1.2% compared to 1.4%) is comparably large so that our model represents the word with two collection-specific frequencies instead of a collection-independent frequency. In contrast, ccLDA assigns "britain" to the collection-independent and the Telegraph-specific word distribution. Representing "britain" in a collection-independent and a collection-specific distribution at the same time is inappropriate. The entropy-based model reveals that the Telegraph and The Guardian differ in usage of the words "isil" and "isis" and that the politician William Hague is mentioned more frequently in The Telegraph for this topic.

Table 7 compares transportation topics from the traveler forum corpus. In the entropy-based model, the word "luggage" occurs in the India-specific and the Singapore-specific topic representation, because it is very frequent in this topic in both collection. However,

**Table 4: Entropy-based model with 25 topics for the patents-papers dataset.**

| ID | Patent-Specific Topic Representation | Collection-Independent Topic Representation | Paper-Specific Topic Representation |
|---|---|---|---|
| 0 | datum include computer apparatus determine | **system method base provide process** | model paper approach result datum |
| 1 | task module tool script work | **software agent group activity environment** | student learning design computer learn |
| 2 | message security user electronic access_control | **key secure signature authentication encryption** | security scheme attack protocol access_control |
| 3 | circuit design test block programmable | **logic integrate gate cell delay** | test fault design circuit testing |
| 4 | server client request user content | **web_service application network system platform** | user mobile cloud architecture technology |
| 5 | pixel display render unit frame | **image color region camera method** | feature object detection recognition algorithm |
| 6 | user content template business knowledge | **system digital provide media_content base** | science computer technology community conference |
| 7 | database table result attribute schema | **query search_result retrieval xml relational** | datum algorithm database optimization mining |
| 8 | output input power second clock | **signal frequency voltage supply current** | power game low consumption circuit |
| 9 | packet queue connection protocol switch | **network route traffic communication flow** | sensor_nodes wireless protocol mobile |
| 10 | channel terminal transmit receive wireless | **communication transmission network receiver rate** | channel performance antenna interference wireless |
| 11 | partition cost constraint algorithm estimate | **cell region method route threshold** | water temperature thermal measurement study |
| 12 | display user_interface device graphical | **virtual interactive provide environment visual** | user_interface interaction human device |
| 13 | thread transaction memory request lock | **resource system application scheduling distribute** | memory parallel performance architecture processor |
| 14 | device electronic mobile user location | **customer network service product system** | market risk patient health price |
| 15 | document content input electronic feature | **text word language character annotation** | document gene semantic recognition feature |
| 16 | element entry table hash classification | **network rule set base machine** | fuzzy classification neural_network learning learn |
| 17 | item segment list matrix determine | **function method vector point space** | problem matrix equation solution linear |
| 18 | program execution instruction execute compile | **source_code language class logic type** | program programming specification formal semantics |
| 19 | object reference interface garbage pointer | **management system process policy change** | business knowledge project organization development |
| 20 | node message graph plurality leaf | **tree path number set time** | algorithm problem graph bound complexity |
| 21 | stream datum audio frame content | **video filter signal code media** | estimation algorithm noise propose scheme |
| 22 | storage memory datum device block | **system control controller volume primary** | robot vehicle design simulation robotic |
| 23 | instruction cache processor memory register | **address window bit trace set** | book guide include server learn |
| 24 | model object threedimensional modeling polygon | **surface point shape mesh flow** | simulation finite_element fluid property |

**Table 5: Example topics from the Wikipedia dataset.**

| **Entropy-Based Model** | | | **ccLDA** | | |
|---|---|---|---|---|---|
| khan judy arjun kabir singh maya ali estevez vidya hindus | | | khan hitch bragg poojaclu cole vidya karan maya anjali | | |
| **English** | **French** | **German** | **English** | **French** | **German** |
| indian | inde | adam | khan | mariage | chucky |
| india | indien | indien | indian | raj | team |
| award | portail | hochzeit | india | carter | carter |
| bollywood | bollywood | liebe | film | inde | spieler |
| love | composee | verliebt | love | kapoor | cole |
| kapoor | musique | tagebuch | award | ronan | spiel |
| mitty | kapoor | brahm | raj | sam | ella |
| raj | interpretee | hideko | family | amoureux | duke |
| english | technique | keat | shah | relation | kannibalen |
| role | indienne | tiger | kapoor | pierre | jennifer |

**Table 6: Example topics from the newspapers dataset.**

| **Entropy-Based Model** | | **ccLDA** | |
|---|---|---|---|
| syria britain support country libya military force attack regime government british | | russia country military attack britain syria support russian world international | |
| **Telegraph** | **Guardian** | **Telegraph** | **Guardian** |
| isil | iran | britain | force |
| mr_hague | iraq | isil | state |
| syrian | syrian | president | president |
| iraq | isis | force | group |
| air_strikes | iraqi | state | leader |
| refugee | international | america | power |
| assad | assad | british | cameron |
| international | western | leader | situation |
| oil | civilian | nation | america |
| stop_the_war | humanitarian | terrorist | official |

**Table 7: Example topics from the traveler forum dataset.**

| **Entropy-Based Model** | | | **ccLDA** | | |
|---|---|---|---|---|---|
| flight airport time hour check fly arrive book leave airline | | | flight fly check hour airline airport arrive time luggage bag | | |
| **UK** | **India** | **Singapore** | **UK** | **India** | **Singapore** |
| heathrow | license | taxi | ticket | mumbaus | changi |
| train | chandigarh | changi | train | domestic | terminal |
| eurostar | luggage | terminal | book | mumbai | time |
| london | car | luggage | time | bangkok | luggage |
| ticket | bike | transit | allow | international | transit |
| frills | drive | changi_airport | eurostar | trivandrum | the_airport |
| paris | reliable | the_airport | london | jet | transit_hotel |
| global_guide | valid | immigration | connection | air | free |
| paddington | storage | transit_hotel | global_guide | arrival | changi_airport |
| option | enfield | shuttle | travel | direct | immigration |

it is not collection-independent because its frequency distribution differs significantly from a uniform distribution with probabilities 0.018 in $C_{India}$, 0.029 in $C_{Singapore}$, and less than 0.01 in $C_{UK}$. In ccLDA, the word "time" is falsely collection-specific *and* collection-independent, occurring in the UK-specific, the Singapore-specific, and the collection-independent topic representation.

## 5 CONCLUSIONS AND FUTURE WORK

We presented a probabilistic, entropy-based topic model for multiple domain-specific text collections. Our approach incorporates multi-word phrases and is the first topic model that precisely distinguishes collection-specific and collection-independent words. To compare our results to the state-of-the-art in cross-collection topic modeling, ccLDA, we evaluated document classification accuracy, topic coherence, and language model perplexity. We extended topic coherence measures from single-collection to cross-collection topic models and therefore allow to consider the alignment of word distributions in the evaluation.

Experiment results demonstrate the robustness of the proposed approach on a variety of text collections across different domains. Our model outperforms ccLDA on all three quality measures on four evaluated datasets of two and three collections. Furthermore, our model provides superior topic representations due to its clear-cut separation of collection-specific and collection-independent words. This separation is achieved by splitting the vocabulary into two sets of collection-specific and collection-independent words according to an entropy-based estimation of each word's termhood. Whereas words with similar frequency across all collections are considered collection-independent, words with significantly different frequency per collection are considered collection-specific.

Possible applications are bias detection in newspaper articles or the revelation of regional and cultural differences in traveler forum posts. Furthermore, topic-based search becomes possible in patents and scientific papers with collection-specific vocabulary and in multi-lingual document collections, such as Wikipedia. While collection-specific word distributions reveal linguistic contrasts, collection-independent word distributions bridge the gap between collections by generalizing from domain-specific language.

For future work, an interesting extension of our approach is to calculate word entropy not only per corpus but per topic. This extension considers words that are collection-specific in one topic and collection-independent in another topic. Furthermore, the number of collections in our evaluation is limited to two and three. Whether this number has a crucial effect needs to be studied in more detail and besides entropy, alternative measures of skewness could be tested. Another possible path is further clustering of collection-specific words per-topic. Opinion words are one example from related work that focuses on specific groups of words. A clustering of words according to their function, role, or semantic context in a text document would be even more interesting.

## REFERENCES


[1] Rajkumar Arun, Venkatasubramaniyan Suresh, C Veni Madhavan, and M Narasimha Murthy. 2010. On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Proc. of the Pacific-Asia Conf. on Advances in Knowledge Discovery and Data Mining (PAKDD)*. Springer, 391–402.

[2] Yang Bao, Nigel Collier, and Anindya Datta. 2013. A partially supervised cross-collection topic model for cross-domain text classification. In *Proc. of the Int. Conf. on Information and Knowledge Management (CIKM)*. ACM, 239–248.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022.

[4] Jing-Shin Chang. 2005. Domain specific word extraction from hierarchical Web documents: A first step toward building lexicon trees from Web corpora. In *Proc. of the SIGHAN Workshop on Chinese Language Processing*. ACL, 64–71.

[5] Changyou Chen, Wray Buntine, Nan Ding, Lexing Xie, and Lan Du. 2015. Differential topic models. In *Transactions on Pattern Analysis and Machine Intelligence*. IEEE, 230–242.

[6] Jacob Eisenstein, Amr Ahmed, and Eric P Xing. 2011. Sparse additive generative models of text. In *Proc. of the Int. Conf. on Machine Learning (ICML)*. ACM, 1041–1048.

[7] Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R. Voss, and Jiawei Han. 2014. Scalable Topical Phrase Mining from Text Corpora. *Proc. of the VLDB Endowment* 8, 3 (Nov. 2014), 305–316.

[8] Yi Fang, Luo Si, Naveen Somasundaram, and Zhengtao Yu. 2012. Mining contrastive opinions on political texts using cross-perspective topic model. In *Proc. of the Int. Conf. on Web Search and Data Mining (WSDM)*. ACM, 63–72.

[9] Denis Fedorenko, Nikita Astrakhantsev, and Denis Turdakov. 2013. Automatic recognition of domain-specific terms: an experimental evaluation., In Proc. of the Spring Researchers Colloquium on Databases and Information Systems. *SYRCoDIS* (2013), 15–23.

[10] Haidong Gao, Siliang Tang, Yin Zhang, Dapeng Jiang, Fei Wu, and Yueting Zhuang. 2012. Supervised cross-collection topic modeling. In *Proc. of the Int. Conf. on Multimedia (MM)*. ACM, 957–960.

[11] Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proc. of the National Academy of Sciences (PNAS)* 101, suppl 1 (2004), 5228–5235.

[12] Yulan He. 2016. Extracting Topical Phrases from Clinical Documents. In *Proc. of the Nat. Conf. on Artificial Intelligence (AAAI)*. AAAI Press, 2957–2963.

[13] Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proc. of the Int. Conf. on Research and Development in Information Retrieval (SIGIR)*. ACM, 50–57.

[14] Shoaib Jameel and Wai Lam. 2013. An Unsupervised Topic Segmentation Model Incorporating Word Order. In *Proc. of the Int. Conf. on Research and Development in Information Retrieval (SIGIR)*. ACM, 203–212.

[15] Kyo Kageura and Bin Umino. 1996. Methods of automatic term recognition: A review. *Terminology* 3, 2 (1996), 259–289.

[16] Noriaki Kawamae. 2014. Supervised N-gram Topic Model. In *Proc. of the Int. Conf. on Web Search and Data Mining (WSDM)*. ACM, 473–482.

[17] Chunyu Kit. 2002. Corpus tools for retrieving and deriving termhood evidence. In *Proc. of the East Asia Forum of Terminology*. 69–80.

[18] Ralf Krestel and Padhraic Smyth. 2013. Recommending patents based on latent topics. In *Proc. of the Conf. on Recommender Systems (RecSys)*. ACM, 395–398.

[19] Jey Han Lau, Timothy Baldwin, and David Newman. 2013. On Collocations and Topic Models. *ACM Transactions on Speech and Language Processing (TSLP)* 10, 3 (July 2013), 10:1–10:14.

[20] Sujian Li, Jiwei Li, Tao Song, Wenjie Li, and Baobao Chang. 2013. A novel topic model for automatic term extraction. In *Proc. of the Int. Conf. on Research and Development in Information Retrieval (SIGIR)*. ACM, 885–888.

[21] Robert V. Lindsey, William P. Headden, III, and Michael J. Stipicevic. 2012. A Phrase-discovering Topic Model Using Hierarchical Pitman-Yor Processes. In *Proc. of the Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. ACL, 214–222.

[22] Jialu Liu, Jingbo Shang, Chi Wang, Xiang Ren, and Jiawei Han. 2015. Mining quality phrases from massive text corpora. In *Proc. of the Int. Conf. on Management of Data (SIGMOD)*. ACM, 1729–1744.

[23] Michael Paul and Roxana Girju. 2009. Cross-cultural analysis of blogs and forums with mixed-collection topic models. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, 1408–1417.

[24] Julian Risch and Ralf Krestel. 2017. What Should I Cite? Cross-Collection Reference Recommendation of Patents and Papers. In *Proc. of the Int. Conf. on Theory and Practice of Digital Libraries (TPDL)*. Springer, 40–46.

[25] Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proc. of the Int. Conf. on Web Search and Data Mining (WSDM)*. ACM, 399–408.

[26] Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. 2018. Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering* (2018).

[27] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnetminer: Extraction and mining of academic social networks. In *Proc. of the Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD)*. ACM, 990–998.

[28] Xuerui Wang, Andrew McCallum, and Xing Wei. 2007. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Proc. of the Int. Conf. on Data Mining (ICDM)*. IEEE, 697–702.

[29] Andrew T Wilson and Peter A Chew. 2010. Term weighting schemes for latent dirichlet allocation. In *Proc. of the Conf. of the North American Chapter of the Association for Computational Linguistics (NAACL)*. ACL, 465–473.

[30] Limin Yao, David Mimno, and Andrew McCallum. 2009. Efficient methods for topic model inference on streaming document collections. In *Proc. of the Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD)*. ACM, 937–946.

[31] ChengXiang Zhai, Atulya Velivelli, and Bei Yu. 2004. A cross-collection mixture model for comparative text mining. In *Proc. of the Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD)*. ACM, 743–748.

[32] Duo Zhang, Qiaozhu Mei, and ChengXiang Zhai. 2010. Cross-lingual latent topic extraction. In *Proc. of the Annual Meeting on Association for Computational Linguistics (ACL)*. ACL, 1128–1137.

[33] Jingwei Zhang, Aaron Gerow, Jaan Altosaar, James Evans, and Richard Jean So. 2015. Fast, flexible models for discovering topic correlation across weakly-related collections. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, 1554–1564.

[34] Liyi Zheng, Tian Nie, Ichiro Moriya, Yusuke Inoue, Takakazu Imada, Takehito Utsuro, Yasuhide Kawada, and Noriko Kando. 2014. Comparative topic analysis of japanese and chinese bloggers. In *Proc. of the Int. Conf. on Advanced Information Networking and Applications Workshops (WAINA)*. IEEE, 664–669.