Topical Sequence Profiling

Tim Gollub*, Nedim Lipka†, Eunyee Koh†, Erdan Genc*, and Benno Stein*
* Bauhaus-Universität Weimar
<first name>.<last name>@uni-weimar.de
† Adobe Systems
lipka@adobe.com, eunyee@adobe.com

***Abstract*—This paper introduces the problem of topical sequence profiling. Given a sequence of text collections such as the annual proceedings of a conference, the topical sequence profile is the most diverse explicit topic embedding for that text collection sequence that is both representative and minimal. Topic embeddings represent a text collection sequence as numerical topic vectors by storing the relevance of each text collection for each topic. Topic embeddings are called explicit if human readable labels are provided for the topics. A topic embedding is representative for a sequence, if for each text collection the percentage of documents that address at least one of the topics exceeds a predefined threshold. If no topic can be removed from the embedding without loosing representativeness, the embedding is minimal. From the set of all minimal representative embeddings, the one with the highest mean topic variance is sought and termed as the topical sequence profile. Topical sequence profiling can be used to highlight significant topical developments, such as raise, decline, or oscillation. The computation of topical sequence profiles is made up of two steps, topic acquisition and topic selection. In the first step, the sequence's text collections are mined for representative candidate topics. As a source for semantically meaningful topic labels, we propose the use of Wikipedia article titles, whereas the respective articles are used to build a classifier for the assignment of topics to documents. Within the second step the subset of candidate topics that constitutes the topical sequence profile is determined, for which we present an efficient greedy selection strategy. We demonstrate the potential of topical sequence profiling as an effective data science technology with a case study on a sequence of conference proceedings.**

I. Introduction

The capability to mine and visualize insights from data has become the basis for competition and growth of companies, and it causes an increasing demand for data science experts and technology [10]. In this paper, we reveal data science technology for the analysis of sequences of text collections. Interesting sequences of this kind may be daily business news feeds, the social media mentions of a company over time frames, or the collected annual proceedings of a research field. Our working hypothesis is that in cases such as those mentioned, statistical insights into the topic distribution of both the individual text collections and the topic development over the sequence is of a high value for the respective stakeholders; however, the amount of potentially relevant topics often prohibits a comprehensive examination. To provide stakeholders with a selection of topics that is informative on the one hand and small enough to be surveyed quickly on the other, we introduce the problem of topical sequence profiling. Taking the annual proceedings of a research field as illustrative example, the goal of topical sequence profiling is to showcase research topics that peak as "hot topic" in distinct years but show a significant decline throughout the remaining years. We argue that, in contrast to topics that never peak or that constantly belong to the "usual suspects", especially from these topics valuable insights can be expected.

A. Problem Definition

The problem of topical sequence profiling can be stated as follows. Given a sequence of text collections $\mathcal{D}, \mathcal{D} = (D_1, D_2, \ldots, D_n)$, where each $D \in \mathcal{D}$ is a set of documents, find the most diverse, explicit topic embedding $\mathbf{T}^*$ for the sequence that is both minimal and representative. $\mathbf{T}^*$ is called the topical sequence profile of $\mathcal{D}$.

A topic embedding $\mathbf{T}$ can be considered as a matrix that represents each $D \in \mathcal{D}$ as a column of k topics,

$$\mathbf{T} = \begin{array}{c} \\ t_1 \\ \vdots \\ t_k \end{array} \begin{array}{c} \begin{array}{ccc} D_1 & \cdots & D_n \end{array} \\ \begin{bmatrix} \mathbf{T}_{11} & & \mathbf{T}_{1n} \\ & \ddots & \\ \mathbf{T}_{k1} & & \mathbf{T}_{kn} \end{bmatrix} \end{array}.$$

The matrix entries $\mathbf{T}_{ij}$ are denoted as *topic coverage* and correspond to the percentage of documents in D_j that are relevant for topic t_i. To compose a topic embedding $\mathbf{T}$, first a set of topics has to be acquired and assigned to the documents in $\mathcal{D}$. Subsequent to the presentation of related work in Section II, an algorithm for this topic acquisition step that utilizes Wikipedia articles as a topic resource and employs text classification to label documents with these topics is introduced in Section III-A. Once a topic set has been acquired and assigned, topic embeddings can be composed by selecting specific topic subsets.

Table I illustrates our concept for the visualization of topic embeddings, which depicts the rows $T_{i:}$ of an embedding in the form of bar charts. In order to become interpretable for users, the topics have to be made *explicit*, i.e., a semantically meaningful label must accompany each bar chart. While our approach utilizes Wikipedia titles as explicit topic labels, latent topic models such as LDA [1] or doc2vec [9] could be employed in combination with topic labeling.

A topic embedding should reveal insights into the topic distribution of each text collection. We call a topic embedding *representative* for a sequence, if for every $D \in \mathcal{D}$, the percentage of documents covered by at least one of the

Table I
CONCEPT FOR THE VISUALIZATION OF TOPICAL SEQUENCE PROFILES. THE SEQUENCE FLOWS FROM LEFT TO RIGHT, THE TOPICS FROM TOP TO BOTTOM IN DESCENDING ORDER OF THEIR DIVERSITY. THE HEIGHT OF EACH OF THE BARS IN THE CELLS IS PROPORTIONAL TO THE TOPIC COVERAGE VALUES $\mathbf{T}_{ij}$. COLORS ENCODE RAISING (GREEN), DECLINING (RED), AND OSCILLATING (GRAY) TOPICS.

Topic	D_1	D_2	D_3	D_4
Topic 1	5%	6%	15%	20%
Topic 2	14%	6%	15%	5%
Topic 3	15%	12%	11%	8%
Coverage	15%	15%	15%	15%

embedded topics exceeds a predefined threshold $c \in [0, 100]$. Since one document may be relevant for multiple topics, the collection coverage for D_j is in general not the sum over column $T_{:j}$ but less or equal. We state the collection coverage for every $D \in \mathcal{D}$ in the last row of our visualization.

To address the requirement of a topic embedding to be of a manageable size, the notion of *minimality* is introduced: A topic embedding is minimal if no topic can be removed without losing representativeness. From all representative minimal topic embeddings we are interested in the instance $\mathbf{T}^*$ that on average contains the *most diverse* topics. Section III-B introduces an effective greedy strategy to the optimization problem of finding $\mathbf{T}^*$ based on a representative non-minimal topic embedding $\mathbf{T}$. As a measure of topic diversity, we propose the variance of the topic distributions. Our choice is motivated by the fact that in order to achieve high topic variance, the topic coverage for the individual text collections must deviate significantly from the mean. This happens if the topic coverage peaks for some text collections and is low elsewhere.

To conveniently spot raising, declining, or oscillation topics in our visualization, we apply linear regression to the topic coverage values of each topic and color-code the slope of the resulting regression curve (cf. Table I). A positive slope (raise) is encoded by green bars, a negative slope (decline) by red bars. The lighter the color, the steeper the slope. Zero slope is encoded by gray bars and, in case the topic's diversity is high, represents an oscillating topic.

The potential of topical sequence profiling as a data science tool is highlighted with a case study on the basis of conference proceedings in Section IV, and we close with a discussion of our contributions in Section V.

II. RELATED WORK

Although, to the best of our knowledge, we are the first to study the problem of topical sequence profiling, a close relation exists to the task of labeling a clustering of documents, for which we point out the state of the art in this section.

Table II
COMPARISON OF TOPICAL SEQUENCE PROFILING AND CLUSTER LABELING WITH RESPECT TO DESIRED TOPIC LABEL PROPERTIES.

Property	Cluster Labeling	Sequence Profiling
Unique	✓	✗
Summarizing	✓	✓
Expressive	✓	✓
Discriminating	✓	✗
Contiguous	✓	✗
Irredundant	✓	✓
Minimal	✓	✓
Representative	✓	✓
Diverse	✗	✓

The problem of cluster labeling can be framed as follows. Given a clustering $\mathcal{C} = \{C_1, \ldots, C_n\}$, where each $C \in \mathcal{C}$ is a set of documents, find a set of explicit topics (the cluster labels) that characterize each of the clusters. Obviously, the above sequence $\mathcal{D}$ can be interpreted as a clustering, and a topic embedding $\mathbf{T}$ may be used to represent the (binary) assignment of labels to clusters. Moreover, both problems include a topic acquisition step that facilitates the composition of $\mathbf{T}$. The most obvious way to acquire explicit topics is to extract (key-) words [5], [14], phrases [3], [17], or queries [6] from the documents in the text collections that are relevant with respect to a retrieval model. More recently, sophisticated parsing technologies have been used to extract only noun phrases [13] or named entities [16], which closely resemble the typical pattern of topics used in library classification systems. The main disadvantage of these approaches is that the label of a relevant topic may not appear in a document, or at least not in a statistically significant way [2]. To overcome this problem, the use of external knowledge resources as a source for explicit topics can be considered state of the art. Proposed resources are thesauri such as WordNet [19], linked open databases such as Dbpedia [8], or encyclopedias such as Wikipedia [2], [12], [15], [18]. Depending on the resource, different classification strategies are proposed, which decide whether or not an external topic is relevant for a document. For example, Carmel et al. formulate search queries from document keyphrases against Wikipedia and classify the top articles as relevant topics. By contrast, the Wikipedia-based approach that we apply for topic acquisition is adopted from the ESA retrieval model, which relies on the cosine similarity between a document and an article for relevance assessments [4].

What distinguishes topical sequence profiling from cluster labeling are the properties topics should satisfy. While we strive for a set of topics that is minimal, representative, and diverse, the desired properties of cluster labels are different. Meyer zu Eißen and Stein [14] have compiled a set of commonly accepted properties, which are listed in Table II; a formal specification of the respective semantics is detailed in their paper. Though half of the properties coincide with our definition of minimality and representativeness, the properties unique, discriminating, and contiguous collide with our definition of diversity.

III. Approach

The approach for computing the topical sequence profile for a sequence $\mathcal{D}$ comprises two steps. In the first step, the topic acquisition step, a comprehensive set of explicit topics is determined, and the topic coverage of each text collection is assessed for each of the topics. The result of the topic acquisition step is a topic embedding $\mathbf{T}$ that is representative but not minimal. In the second step, topics are removed from $\mathbf{T}$ with the objective to find the most diverse minimal topic embedding $\mathbf{T}^*$.

A. Topic Acquisition

As pointed out in Section II, there are several ways for obtaining a set of explicit topics that are tailored to a collection of documents. In line with the state of the art, our approach of choice is to consider the titles of Wikipedia articles as explicit topics. To decide whether a document is relevant for a Wikipedia topic or not, the cosine similarity between the vector space representations of the Wikipedia article and the document is computed under the BM25 model [11]. To make the binary decision upon relevance, which is needed to compute the topic coverage values $\mathbf{T}_{ij}$, we adopt an effective unsupervised technique from the field of similarity graph sparsification [7]. The main idea of the approach is to compute for every topic an expected similarity score based on the aggregated vector representation of the whole sequence. Only if the similarity score of a document exceeds the expected value, the document is classified as being relevant for the topic.

With more than five million English articles, the pairwise computation of similarities between Wikipedia articles and sequence documents is inefficient for large sequences, and an efficient strategy is desired that determines a subset of articles that contain the topics of $\mathbf{T}^*$ with full recall and acceptable precision. To this end, we reuse the aggregated vector representation of the sequence and determine its most similar Wikipedia articles. The rationale is that if the text collections of the sequence have a common general topic domain (such as a common research field), the most similar articles of the aggregated vector should reveal this. Using these articles as seeds, we can traverse the Wikipedia link graph until a representative topic embedding $\mathbf{T}$ is obtained that is tailored to the sequence. To optimize the quality of the traversal, we consider only links that have been clicked at least ten times according to a recently released Wikipedia clickstream dataset [20].

B. Topic Selection

Given the representative topic embedding $\mathbf{T}$ of the topic acquisition step, the optimization problem of the topic selection step is to determine the subset of topics in $\mathbf{T}$ that maximizes the average topic diversity and satisfies the minimality property:

$$\begin{aligned} \text{maximize} \quad & \frac{1}{k}\sum_{i=1}^{k} \mathrm{Var}(\mathbf{T}_{i:}) \\ \text{subject to} \quad & \mathbf{T} \text{ is minimal} \end{aligned}$$

Note that the above optimization problem is an instance of the set cover problem, and hence it cannot be solved efficiently for large sequences.[1] Here we present, in form of Algorithm 1, an efficient greedy strategy to find an approximate solution. First, the rows (topics) of the given topic embedding $\mathbf{T}$ are sorted by diversity in ascending order, and $\mathbf{T}^*$ is initialized with $\mathbf{T}$. Then, for each of the k topics in $\mathbf{T}$, starting with the least diverse topic, it is checked whether the removal of the topic still yields a representative topic embedding. If so, the topic is removed from $\mathbf{T}^*$. After applying this procedure, $\mathbf{T}^*$ is minimal, and since the topics are removed in ascending order of their diversity, the algorithm strives for maximizing the average topic diversity.

Algorithm 1 Greedy Topic Selection

Input: Topic Embedding $\mathbf{T}$
Output: Topic Embedding $\mathbf{T}^*$

```
sortAscending(T)
T* ← T
for i = 1; i ≤ k; i = i + 1 do
    if representative(T* \ T*_{i:}) then T* ← T* \ T*_{i:}
return T*
```

IV. Case Study

Due to the complexity of the task, a thorough evaluation of the usefulness of topical sequence profiling in practical scenarios would require an extensive user study. In this paper, however, we resort to a case study on the basis of a sequence of SIGIR conference proceedings to gather first empirical insights into the performance of our approach. We choose SIGIR proceedings since we, and likely the reader, are familiar with the information retrieval research domain. What we wish to obtain is a topical sequence profile that (1) consists of reasonable information retrieval research topics that are objectively representative, and that (2) show an interesting development over the years. We consider proceedings from 2007 to 2015 which results in the following sequence.

$\mathcal{D}$	Year	# Papers	$\mathcal{D}$	Year	# Papers
D_1	2007	198	D_6	2012	216
D_2	2008	193	D_7	2013	205
D_3	2009	193	D_8	2014	226
D_4	2010	214	D_9	2015	193
D_5	2011	232			

Topic Acquisition. First we obtain a small set of seed Wikipedia articles that match the general topic domain of the sequence to facilitate an efficient acquisition of a representative topic embedding $\mathbf{T}$. For this purpose, the BM25 vector space representations for all papers in $\mathcal{D}$ are aggregated. In terms of the cosine similarity with this aggregated vector, the ten most similar Wikipedia articles obtained are:

1) Concept Search (0.678)
2) Information Retrieval (0.593)

[1] If $P \neq NP$.

3) Human-Computer Information Retrieval (0.588)
4) Web Query Classification (0.582)
5) Enterprise Search (0.549)
6) Search engine technology (0.540)
7) Document retrieval (0.539)
8) Cognitive models of information retrieval (0.524)
9) Federated search (0.524)
10) Web search query (0.518)

Starting from these ten articles, the Wikipedia link graph is traversed in a breadth first manner and every sequence document is classified against the visited articles. With a collection coverage threshold c of 80%, the traversal stopped with a representative topic embedding $\mathbf{T}$ after visiting 1261 Wikipedia articles, which indicates that link graph traversal based on seed articles leads to significant efficiency improvements over the classification against the whole Wikipedia.

Topic Selection. The topical sequence profile $\mathbf{T}^*$ obtained after applying Greedy Topic Selection to $\mathbf{T}$ with c set to 60% is illustrated in Table III. Since each of the proceedings contain about 200 documents, a topic coverage of one percent roughly corresponds to two papers that have been assigned to a topic in a specific year. The topical sequence profile consists of 19 topics. I.e., altogether, these topics are representative for each of the years and no topic can be removed without loosing representativeness. Though some of the topics may be interpretable only after looking up the respective Wikipedia article (Table III includes hyperlinks to the articles), a reasonable selection of information retrieval research can be observed. "Library Classification" is declining and the most diverse topic, followed by "Query" and the raising topics "Search engine results page" and "Endeca". The topics "InnoDB" and "Hidden Markov model" show oscillating behavior. The collection coverage values at the bottom reveal that the document coverage for 2014 prevents the removal of further topics (reaches c).

V. Discussion

With topical sequence profiling, we contribute a new research problem for the analysis and visualization of sequential text collections. In contrast to cluster labeling, sequence profiles aim at revealing representative topics that are subject to significant changes in terms of their coverage throughout a sequence of text collections. A larger evaluation is in preparation and could not be presented here, but the shown case study revealed that the computation of topical sequence profiles is efficient and produces promising results. For practical applications, we observe that through interactive topical sequence profiles, which update after users explicitly remove or pin topics from the profile, the perceived quality can be further increased. Due to the efficiency of the greedy topic selection algorithm, the updating of a profile in response to user interaction can be achieved instantaneously. Further, the post-acquisition of topics based on a given topical sequence profile seems worth considering. Looking again at the SIGIR profile in Table III, it appears that the post-acquisition of a raising topic that contributes to the collection coverage of 2014 would help balancing both the collection coverage values as well as the ratio of raising, declining, and oscillating topics.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3:993–1022, 2003.

[2] D. Carmel, H. Roitman, and N. Zwerdling. Enhancing Cluster Labeling Using Wikipedia. In *Proceedings of the 32nd international ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 09)*, pages 139–146, New York, NY, USA, 2009. ACM.

[3] N. Erbs, I. Gurevych, and M. Rittberger. Bringing order to digital libraries: From keyphrase extraction to index term assignment. *D-Lib Magazine*, 19(9/10), 2013.

[4] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, volume 7, pages 1606–1611, 2007.

[5] F. Geraci, M. Pellegrini, M. Maggini, and F. Sebastiani. Cluster Generation and Cluster Labelling for Web Snippets: A Fast and Accurate Hierarchical Solution. In *Proceedings of the 13th Symposium on String Processing and Information Retrieval (SPIRE 06)*, pages 25–36, 2006.

[6] T. Gollub, M. Hagen, M. Michel, and B. Stein. From Keywords to Keyqueries: Content Descriptors for the Web. In C. Gurrin, G. Jones, D. Kelly, U. Kruschwitz, M. de Rijke, T. Sakai, and P. Sheridan, editors, *36th International ACM Conference on Research and Development in Information Retrieval (SIGIR 13)*, pages 981–984. ACM, July 2013.

[7] T. Gollub and B. Stein. Unsupervised Sparsification of Similarity Graphs. In H. Locarek-Junge and C. Weihs, editors, *Classification as a Tool for Research. Selected papers from the 11th IFCS Biennial Conference and 33rd Annual Conference of the German Classification Society (GFKL)*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 71–79, Berlin Heidelberg New York, 2010. Springer.

[8] I. Hulpus, C. Hayes, M. Karnstedt, and D. Greene. Unsupervised graph-based topic labelling using dbpedia. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 465–474, New York, NY, USA, 2013. ACM.

[9] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1188–1196, 2014.

[10] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers. Big data: The next frontier for innovation, competition, and productivity. Technical report, McKinsey Global Institute, 2011.

[11] S. Robertson and H. Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, Apr. 2009.

[12] U. Scaiella, P. Ferragina, A. Marino, and M. Ciaramita. Topical clustering of search results. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 223–232, New York, NY, USA, 2012.

[13] J. Stefanowski and D. Weiss. Comprehensible and accurate cluster labels in text clustering. In *Large Scale Semantic Access to Content (Text, Image, Video, and Sound) (RIAO 07)*, pages 198–209, Paris, France, France, 2007. Le Centre de Hautes Etudes Internationales d'Informatique Documentaire.

[14] B. Stein and S. Meyer zu Eißen. Topic Identification: Framework and Application. In K. Tochtermann and H. Maurer, editors, *4th International Conference on Knowledge Management (I-KNOW 04)*, Journal of Universal Computer Science, pages 353–360, Graz, Austria, July 2004. Know-Center.

[15] Z. S. Syed, T. Finin, and A. Joshi. Wikipedia as an ontology for describing documents. In *ICWSM*, 2008.

[16] H. Toda and R. Kataoka. A Clustering Method for News Articles Retrieval System. In *Proceedings of the 14th International Conference on World Wide Web (WWW 05) – Special Interest Track and Posters*, pages 988–989, New York, NY, USA, 2005. ACM.

[17] P. Treeratpituk and J. Callan. An experimental study on automatically labeling hierarchical clusters using statistical features. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 707–708, New York, NY, USA, 2006. ACM.

[18] B. Wei, J. Liu, Q. Zheng, W. Zhang, C. Wang, and B. Wu. Df-miner: Domain-specific facet mining by leveraging the hyperlink structure of wikipedia. *Knowledge-Based Systems*, 77:80 – 91, 2015.

[19] T. Wei, Y. Lu, H. Chang, Q. Zhou, and X. Bao. A semantic approach for text clustering using wordnet and lexical chains. *Expert Systems with Applications*, 42(4):2264 – 2275, 2015.

[20] E. Wulczyn and D. Taraborelli. Wikipedia clickstream. *figshare*, 2015.

Table III

TOPICAL SEQUENCE PROFILE FOR THE PROCEEDINGS OF THE SIGIR CONFERENCE FROM 2007 TO 2015. THE SEQUENCE FLOWS FROM LEFT TO RIGHT, THE TOPICS, IN DESCENDING ORDER OF THEIR DIVERSITY, FROM TOP TO BOTTOM. THE HEIGHT OF EACH CELL IS PROPORTIONAL TO THE COVERAGE OF THE TOPIC IN THE RESPECTIVE YEAR. COLORS ENCODE RAISING (GREEN), DECLINING (RED), AND OSCILLATING (GRAY) TOPICS.

Topic	2007	2008	2009	2010	2011	2012	2013	2014	2015
Library classification	14.6%	12.4%	11.9%	10.7%	6.0%	11.1%	5.9%	2.7%	5.7%
Query	14.1%	12.4%	15.0%	17.8%	12.1%	11.6%	9.3%	6.2%	6.2%
Search engine results page	10.6%	8.3%	14.0%	12.1%	19.0%	13.4%	13.2%	16.7%	16.1%
Endeca	9.1%	6.7%	11.9%	10.3%	15.5%	13.4%	13.7%	13.6%	17.1%
Extended Boolean model	10.1%	13.0%	6.7%	12.1%	7.8%	5.1%	5.9%	3.5%	7.8%
Database search engine	10.1%	8.8%	14.5%	10.3%	17.7%	11.1%	13.2%	15.5%	14.0%
InnoDB	14.1%	11.4%	17.1%	14.5%	9.9%	12.5%	8.8%	13.6%	16.6%
World Wide Web	10.1%	10.9%	8.8%	11.2%	13.8%	5.1%	8.3%	7.0%	6.2%
Taxonomy for search engines	8.1%	5.7%	8.3%	5.6%	13.4%	9.7%	9.8%	12.4%	9.8%
Full text database	9.1%	10.4%	5.7%	8.9%	3.4%	4.6%	4.4%	5.0%	3.6%
Natural language programming	5.6%	12.4%	6.2%	6.1%	5.6%	4.6%	6.8%	3.1%	7.3%
Probabilistic relevance model	10.1%	10.4%	7.3%	10.3%	8.2%	5.6%	7.8%	2.7%	7.3%
Classification	10.6%	11.4%	8.8%	8.4%	6.0%	9.3%	7.3%	3.9%	5.7%
Text segmentation	9.6%	12.4%	5.7%	7.5%	6.9%	6.9%	5.9%	6.2%	4.7%
Latent Dirichlet allocation	7.1%	11.9%	10.4%	8.4%	7.8%	5.6%	8.8%	4.7%	7.3%
Data mining	6.6%	11.9%	9.3%	7.9%	13.4%	10.2%	12.7%	9.7%	9.8%
Hidden Markov model	7.6%	9.3%	6.2%	8.9%	5.2%	4.6%	6.3%	5.8%	10.9%
Cosine similarity	7.1%	6.7%	7.3%	6.1%	5.6%	2.3%	3.9%	4.3%	9.3%
Gensim	7.1%	13.0%	9.8%	7.5%	8.2%	9.3%	7.8%	8.5%	6.2%
Coverage	75%	76%	78%	74%	72%	67%	67%	60%	71%