

---

# TECoMINER: TOPIC DISCOVERY THROUGH TERM COMMUNITY DETECTION

---

TECHNICAL NOTE

**Andreas Hamm**  
**Jana Thelen**  
**Rasmus Beckmann**  
**Simon Odrowski**

{Andreas.Hamm, Rasmus.Beckmann, Simon.Odrowski@dlr.de}  
Think Tank  
German Aerospace Center DLR  
Cologne, Germany

March 18, 2021

## ABSTRACT

This note is a short description of TeCoMiner, an interactive tool for exploring the topic content of text collections. Unlike other topic modeling tools, TeCoMiner is not based on some generative probabilistic model but on topological considerations about co-occurrence networks of terms. We outline the methods used for identifying topics, describe the features of the tool, and sketch an application, using a corpus of policy related scientific news on environmental issues published by the European Commission over the last decade.

**Keywords** Information extraction, text mining, NLP, word communities, modularity, topic detection, topic modelling, topic visualization, environmental policy

## 1 Introduction

The rapidly increasing amount of electronically available texts has augmented the importance of automatized unsupervised methods for text exploration and analysis. A very typical task is to identify the themes which are latent in massive text collections. Not only does this help to obtain a quick overview over the content of the text collection, but it also enables a structured analysis of relationships and developments reflected in the texts.

In contrast to supervised text classification, the discovery of topics is quite an open-ended endeavor, leaving an important role to subject-related interpretation by domain experts. Our own experience as an interdisciplinary team involved in monitoring and assessing huge text collections has led us to leave the well-trodden path of probabilistic topic models and explore the possibilities of detecting topics as communities in term networks. We have applied this approach to several text corpora [1] (scientific publications, political texts, RSS news feeds) and found it to be advantageous with regard to the ease of topic interpretation and to the command over topic granularity.

TeCoMiner is a software tool designed for users who want to apply the **Term Community** approach to topic detection to their corpora of interest. Before we describe the tool and demonstrate its application we will give an overview over the underlying methods.

## 2 Related Work

During the last two decades probabilistic topic models have dominated the topical analysis of text corpora, seminally influenced by [2]. Based on popular software packages and more specialized advanced methods, this has led to

many applications in areas as diverse as scientometric publication analysis, social media monitoring, literature studies, historical text exploration, and the social sciences. For an extensive overview we refer to [3].

The question of how to evaluate topic quality is problematic. For the applications mentioned above, topic interpretability by humans is the key measure of success, but this is a concept which is not easy to grasp in a computational way. Various metrics of topic coherence [4] have been studied as indicators of interpretability. Recently it was suggested to involve word embeddings in the assessment of topic coherence [5]. We found this approach useful in our own comparative experiments with various topic models [1], and it has influenced the way we present topics in TeCoMiner.

Visualization methods play an important role in aiding topic interpretation; [6, 7] are only two out of many examples designed for depicting probabilistic topic models for human consumption. TeCoMiner uses similar visualizations but is geared towards network based topic detection.

The idea that topics can be interpreted as communities in networks which embody the relations between words and documents of a corpus came up in various forms [8, 9, 10, 11]. In these settings, topics can be found by applying one of the many methods of community detection [12]. We will explain below how we framed this approach in a way that leads to enhanced interpretability and controlled granularity of topics.

### 3 Methods

While the generative approach of probabilistic topic modeling follows the sound methodology of statistical machine learning with a high potential for insights into the genesis of a corpus, it is built on certain assumptions about document generation and prior distributions that are disputable. In contrast, here we take a more phenomenological position when exploiting observed word co-occurrences in the corpus documents for topic detection. This can be done by studying the characteristics of co-occurrence networks of terms. Topics then show up as communities of strongly connected terms. However, we suggest that it needs careful pre- and postprocessing steps for achieving results that are competitive or even better compared to probabilistic topic models. Here we briefly describe the core elements of our processing pipeline (Figure 1). More details can be found in [13].

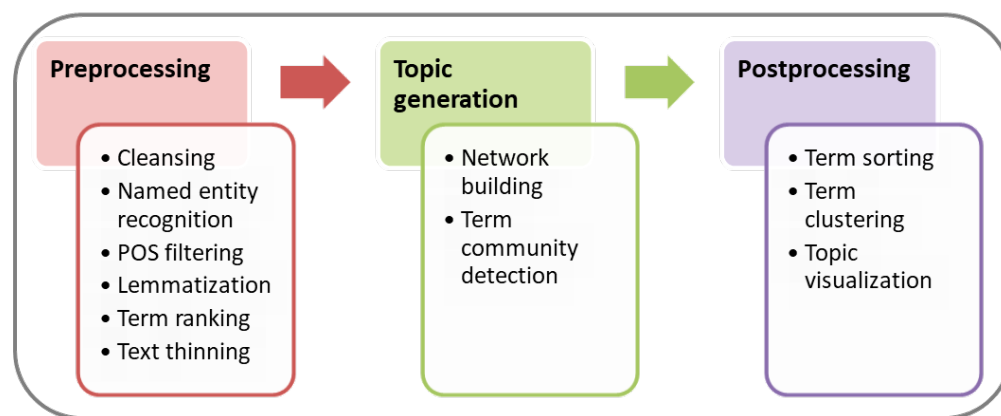


Figure 1: Topic identification pipeline in TeCoMiner

#### 3.1 Term Ranking

Starting from a corpus of raw texts, we first follow standard NLP preparation procedures: Removal of unwanted tokens and stopwords; POS filtering that retains only adjectives, nouns, and proper nouns; lemmatization.

With regard to multiword expressions we experimented with various approaches. While merging words which form a unique expression is beneficial for the interpretation of term communities, the confusing effect of word combinations that just appear to be common by statistical observations without bearing a special meaning can be more harmful. We therefore consider only multiword combinations that show up in named entity recognition.

In preparation of a viable term network it is essential to reduce the number of terms retained in each document even further. For this purpose, we introduce a term ranking in each document and keep only the top-ranked terms. This is related to unsupervised keyword extraction which aims at finding those terms of a document that are most significant for its content. We have developed an approach - posIdfRank - which combines several ideas [14]: the PageRank inspired

voting based on local word neighborhood associations introduced in [15], a weighting according the absolute position in the text [16], and counterbalancing the influence of unspecific words by the inverse document frequency [17].

Technically, we obtain the ranking values for the  $n$  terms of a document as the stationary distribution of a Markov chain  $x_{t+1} = Gx_t$  on the  $n$ -dimensional space of terms with a transition matrix

$$G_{ij} = \alpha \frac{\text{Idf}_j f_{ij}}{\sum_{k=1}^n \text{Idf}_k f_{ik}} + (1 - \alpha) \frac{(1 + \text{pos}_i)^\beta \text{Idf}_i}{\sum_{k=1}^n (1 + \text{pos}_k)^\beta \text{Idf}_k},$$

where  $\text{Idf}_i$  is the inverse document frequency of term number  $i$ ,  $\text{pos}_i$  is the earliest position of that term and  $f_{ij}$  is the frequency how often the terms number  $i$  and  $j$  lie in the same neighborhood window of size  $w$ .  $\alpha$ ,  $\beta$ , and  $w$  are tuning parameters which we chose as  $\alpha = 0.9$ ,  $\beta = -0.9$  and  $w = 11$  after experiments with documents with pre-assigned keywords [14].

Each document is then thinned by keeping only the top  $P$  percent of terms according to this ranking, where we choose  $P$  between 10 and 80 depending on the average length of the corpus documents.

### 3.2 Term Community Detection

We define a network with all the terms contained in the thinned documents as vertices  $\mathcal{V}$ . The edge weight  $A_{ij}$  between two vertices  $v_i$  and  $v_j$  is defined to be the number of documents in which  $v_i$  and  $v_j$  appear together. Pruning edges between rare terms is an option for large corpora with long texts.

Communities in networks are, intuitively speaking, vertex groups with strong linkage inside the groups but only loose connections to other groups. Comparing various community detection approaches we found that modularity optimization produces particularly good topics with respect to interpretability [1].

In TeCoMiner we use the generalized modularity definition introduced in [18]: We call a partition  $\mathcal{C} = \{C_s, s = 1, \dots, m\}$  of  $\mathcal{V}$  a candidate of communities. The generalized modularity,  $\mathcal{H}_\gamma(\mathcal{C}) = \mathcal{I}(\mathcal{C}) - \gamma \mathcal{J}(\mathcal{C})$ , compares for a candidate partition  $\mathcal{C}$  the fraction of edge weights inside of candidate communities,  $\mathcal{I}(\mathcal{C}) = \frac{1}{2m} \sum_{i,j} A_{ij} \delta_{c(i)c(j)}$ , for the given graph on the one hand with the expected fraction of edge weights inside candidate communities for a random network with the same degree distribution,  $\mathcal{J}(\mathcal{C}) = \frac{1}{(2m)^2} \sum_{i,j} k_i k_j \delta_{c(i)c(j)}$ , on the other hand; here  $k_i = \sum_j A_{ij}$  denotes the weighted degree (weighted number of edges) of vertex  $v_i$ ,  $m = \frac{1}{2} \sum_i k_i$  is the weighted total of edges in the network,  $c(i)$  enumerates the candidate community of vertex  $v_i$ , and  $\delta_{ij}$  is Konecker's delta. The partition which maximizes  $\mathcal{H}_\gamma(\mathcal{C})$  describes the optimal communities in the sense of high intra-group linkage compared to the expectations in a random situation. The parameter  $\gamma$  influences how strongly one values the gain of additional intra-group edge weights. With  $\gamma = 0$  one does not compare to the random situation at all and therefore the optimal solution is one all-embracing community. With  $\gamma \rightarrow \infty$  intra-group links practically do not get rewarded, so that the extreme partition into one-vertex communities appears as the optimal solution. Hence,  $\gamma$  can be used as a resolution parameter: smaller values lead to a few big communities, larger values to many small communities.

Maximizing the modularity is known to be NP-hard. However, there exist efficient greedy algorithms for finding local modularity maxima. In TeCoMiner we use the Leiden algorithm [19].

### 3.3 Term Community Presentation

Community detection in the term co-occurrence network partitions all terms into topics. Consequently, a topic is typically characterized by some hundreds or even thousands of terms. Intrinsically, this list of terms does not come with any order. This is different from the situation in probabilistic topic models where the model produces a probability distribution over topic words.

Therefore, we introduce two criteria of how to structure the topic terms when presenting them for human interpretation: a rating of terms for sorting them according to significance, and a stratification of terms into layers of semantically similar terms.

In subsection 3.1 we have already introduced `posIdfRank` as a method for finding the most significant terms per document. Now we need a measure for significance within the whole corpus. A naive average over the `posIdfRanks` of a term from all documents in which it occurs would be unfair because of the very different document frequencies of terms. A more appropriate way of rating is the Bayesian average as known from scoring systems. Concretely, we rank a term  $a$  by calculating  $x(a) = \frac{0.3C + s(a)}{C + d(a)}$  where  $d(a)$  is the number of documents containing  $a$ ,  $s(a) = \sum_u x_u(a)$  with  $x_u(a)$  equal to 3, 2 or 1 depending on whether  $a$  is among the top 5, 10 or 15 percent of terms in document  $u$  according to `posIdfRank`, and  $C$  is the sum of the number of all unique terms per document divided by the number of unique terms in the whole corpus, see [1] and [13].

We discover semantically similar terms by mapping all topic terms into a 300-dimensional vector space with a pre-trained fastText word embedding [20]. In this space we form groups of semantically related terms by agglomerative clustering based on Euclidean distance.

We use these two structuring principles for an easy-to-grasp visualization of topic terms in the form of a stratified word cloud which shows the terms in sizes according to their significance ranking positions and in colored horizontal strata which bring together semantically related terms.

## 4 Demonstration

We discuss how users can readily apply term community topic detection based on the methods described in the previous section with TeCoMiner, a software tool we wrote in Python utilizing in particular the packages pandas, spaCy, python-igraph, scikit-learn, gensim, wordcloud, Bokeh, HoloViews, and Panel. Here we briefly describe its features, which were developed in close cooperation between data scientists and end users, and introduce our demonstration case.

### 4.1 TeCoMiner Web Application

TeCoMiner runs as a single-page application in a web browser. There are two work phases: first, uploading and preprocessing of corpora through the **Add corpus** feature (with an option to change the parameters  $\alpha$ ,  $\beta$ ,  $w$  and  $P$  mentioned in subsection 3.1), second, interactive topic detection and analysis within those corpora. Preprocessed corpora can be looked at in several views presented as tabs: *Model*, *Topic*, *Document*, and, depending on the corpus, further tabs for metadata connected to the documents—in the present example the tabs *Time* and *Theme*.

The **Model tab** (Figure 2) provides an overview of the current community topic model in the form of a two-dimensional t-SNE plot [21]. The dots represent documents; in a model with  $N$  topics they are first placed in an  $N$ -dimensional space according to the proportions each topic contributes to the document and are then mapped to two dimensions via t-SNE. The color of each dot is chosen depending on the topic with highest proportion in that document. The title of each document can be displayed through mouse-over. Large unicolored clusters represent the most dominant topics; scanning the titles involved gives a first impression of what those topics deal with. Dense clusters hint at sharp topics. Coalescing clusters indicate topical relationships.

On the Model tab it is also possible to generate a new topic model after choosing a value for the resolution parameter  $\gamma$  (see subsection 3.2).

Topics can be analyzed in more detail by selecting TeCoMiner’s **Topic tab** (Figure 3). For a topic chosen from a drop-down list, the left side of the screen shows a stratified word cloud of the topic terms as described in subsection 3.3. With this form of presentation the user can recognize at a glance the subject-related common ground in about 100 terms. Highly ranked words stand out by size, and the colored horizontal strata group related terms.

On the right side of the screen, there is a list of up to 30 documents in which the selected topic takes a proportion of more than 15 percent, sorting the documents with highest proportion to the top.

The **Document tab** (Figure 4) shows the full text of a document and highlights in it all terms that belong to any topic with a proportion of at least 10 percent. Different topics are marked in different colors. Documents can be selected either by entering their file name or by opening them from the document list of the Topic tab. On the one hand, this tab is particularly useful for understanding and interpreting topics if word clusters and document titles have not been sufficiently insightful. On the other hand, it also provides deeper insights into how different topics relate to each other.

With the **Time tab** it is possible to follow the temporal evolution of topics. This tab presents as line charts the time series of the accumulated proportion which the (multi-)selected topics have in all the documents published in each single month.

The **Theme tab** refers to pre-assigned thematic tags (like “Biodiversity”, “Climate change”, and “Sustainable mobility”) which are provided with the documents of the present corpus. It visualizes the connection between these tags and the detected topics, which is useful as a consistency check and as support for topic interpretation.

Next to the interactive features, on the **Download tab** TeCoMiner offers an Excel export of the stratified topic term clouds and the topic distribution per document.



Figure 2: TeCoMiner Model tab, with a graphical overview of the model and the option to recalculate models with varying resolutions. Here: view of a resolution-1 model of the EU Science for Environment Policy News Alert 2011–2020 corpus

## 4.2 Application Case

We show some results derived from a collection of 1463 articles scraped from the European Commission’s *EU Science for Environment Policy News Alert*<sup>1</sup> from April 2011 until May 2020. They summarize environmental research studies for a political audience or decision-makers in general, therefore using a non-technical and accessible language. The article length varies from 100 to 1400 words.

Preprocessing including thinning with  $P = 33.3\%$  of that corpus takes hardly more than 10 minutes on an Intel Core i7 PC. This produces a term network of 16 152 vertices and 1 149 078 edges. Calculation of a term community topic model including visualizations runs in 20 seconds so that users can easily explore topics of variable granularity in an interactive manner.

<sup>1</sup><https://ec.europa.eu/environment/integration/research/newsalert/>



**TeCoMiner**

Model Topic Document Time Theme Download Add corpus

Document filename  
465na2\_en

Topic\_21: 37.4%  
Topic\_27: 19.0%  
Topic\_59: 11.5%

Does environmental noise lead to depression and anxiety? People who are annoyed by environmental noise are also more likely to suffer from depression and anxiety, a new, large-scale study from Germany suggests. The results do not prove that noise causes mental health issues but suggest a possible link, which the study's authors are exploring further. Of all the types of noise considered in the study, aircraft noise was reported to be the most annoying. Noise, such as traffic and industrial noise, is now recognised as a serious environmental problem and is regulated in Europe under the EU's Environmental Noise Directive.

It is associated with a number of cardiovascular health problems, including heart disease, heart failure and stroke.

It is also well known that noise can cause annoyance, which can be accompanied by negative, stress-related emotions, such as irritability, distress and exhaustion. However, very little research has considered whether this annoyance and potential stress could lead to mental health disorders. Therefore, this study investigated whether there is a link between noise annoyance and depression and anxiety. It also explored the annoyance levels caused by different sources of noise.

The researchers analysed questionnaires completed by 14 635 residents, aged 35–74, in and around the city of Mainz, Germany, between 2007 and 2012. Part of this area is in the flight path of nearby Frankfurt Airport, one of the busiest airports in the world. The questionnaires asked the residents how annoyed they had been in recent years (rated on a five-point scale, from not annoyed to extremely annoyed) by six different types of environmental noise: road traffic, aircraft, railways, industrial construction, neighbourhood indoor noise, and neighbourhood outdoor noise.

The results show that 20.7 % of participants reported no annoyance to the sources of environmental noise, 26.6 % slight annoyance, 25 % moderate annoyance, 17.3 % strong annoyance and 10.5 % extreme annoyance. Of the six types of noise considered, aircraft noise was the most problematic. Nearly 60 % of the population reported being annoyed by it to some degree, and 6.4 % were extremely annoyed by it. Results in Table 1 show annoyance levels caused by the noise sources.

Source of noise Percentage of participants affected (slightly, moderately, strongly or extremely annoyed) Percentage of participants extremely annoyed Aircraft 59.9 % 6.4 % Road traffic 43.5 % 1.9 % Neighbourhood outdoor 31.8 % 1.2 % Neighbourhood indoor 19.6 % 0.9 % Industrial construction 19.6 % 0.9 % Railway 15.8 % 0.7 % Table 1: annoyance caused by six sources of environmental noise among study participants. Note: the study does not relate these figures to noise exposure levels. Continued on next page.

They asked the participants to indicate whether they suffered symptoms of depression and anxiety, and the researchers assigned a score for each condition. Participants were also asked if they had ever received medical diagnoses of depression or anxiety. They found that indicators of depression and anxiety increased steadily with levels of annoyance to the noise.

Average depression scores increased from 3.5 (out of a possible total of 27) among the 'no annoyance' group, to 5.1 for the 'extreme annoyance' group. The percentage of each group with a depression score of 10 or more (a clinically significant level of depression) increased from 6.1 % of the 'no annoyance' group through to 12 % of the 'extremely annoyed' group. The percentage of the population with medical diagnoses of depression was also higher with each level of annoyance, for instance, 10.1 % of the 'no annoyance' group and 14.8 % of the 'extremely annoyed' group had been diagnosed with depression by a doctor.

Average anxiety scores steadily increased from 0.7 (out of a possible total of 6) in the 'no annoyance' group, to 1.1 among the 'extreme annoyance' group. The percentage of each group with a clinically significant anxiety score of 3 or more increased from 4.5 % of the 'no annoyance' group through to 10 % of the 'extreme annoyance' group. 6.3 % of the 'no annoyance' group had been diagnosed with anxiety disorders, but the figure was 9.9 % for the 'extreme annoyance' group. The study did not assess actual noise levels, just personal responses to noise. It also points out the possibility that people who are already depressed or anxious may be more sensitive to noise and, therefore, report higher annoyance; it is not necessarily the case that noise annoyance leads to mental health issues. However, the association between annoyance and mental health disorders in these data is very strong and the researchers say their results are 'compatible' with the hypothesis that annoyance leads to stress, which in turn can lead to depression and anxiety, or worsen existing symptoms. They are, therefore, conducting regular follow-up assessments with the participants to explore the possible relationship between noise and mental health further.

Figure 4: Document tab, showing highlighted topic terms within a document. Here: A document with title “Does environmental noise lead to depression and anxiety?” Three topics have a share of more than 10 percent in this document: *noise pollution* (Topic 21; red), *health issues* (Topic 27; orange), and *urban living conditions* (Topic 59; green).

## Acknowledgement

We are grateful to Mark Azzam for stimulating discussions.

## References

- [1] Jana Thelen. Methoden der Netzwerkanalyse im Topic Modeling. Master's thesis, Department of Mathematics and Computer Science, University of Cologne, 2020. <https://elib.dlr.de/141146/>.
- [2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.
- [3] Jordan Boyd-Graber, Yuening Hu, and David Mimno. Applications of topic models. *Foundations and Trends® in Information Retrieval*, 11(2-3):143–296, 2017.
- [4] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15*. ACM Press, 2015.
- [5] Anjie Fang, Craig Macdonald, Iadh Ounis, and Philip Habel. Using word embedding to evaluate the coherence of topics from twitter data. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval - SIGIR '16*. ACM Press, 2016.

- 
- [6] Carson Sievert and Kenneth Shirley. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. Association for Computational Linguistics, 2014.
- [7] Kostiantyn Kucher, Rafael M. Martins, and Andreas Kerren. Analysis of VINCI 2009-2017 proceedings. In *Proceedings of the 11th International Symposium on Visual Information Communication and Interaction - VINCI '18*. ACM Press, 2018.
- [8] Hassan Sayyadi and Louiqa Raschid. A graph analytical approach for topic detection. *ACM Transactions on Internet Technology*, 13(2):1–23, dec 2013.
- [9] Andrea Lancichinetti, M. Irmak Sirel, Jane X. Wang, Daniel Acuna, Konrad Kording, and Luís A. Nunes Amaral. High-reproducibility and high-accuracy method for automated topic classification. *Physical Review X*, 5(1), 2015.
- [10] Tommy Dang and Vinh The Nguyen. ComModeler: Topic Modeling Using Community Detection. In Christian Tominski and Tatiana von Landesberger, editors, *EuroVis Workshop on Visual Analytics (EuroVA)*. The Eurographics Association, 2018.
- [11] Martin Gerlach, Tiago P. Peixoto, and Eduardo G. Altmann. A network approach to topic models. *Science Advances*, 4(7):eaq1360, 2018.
- [12] Santo Fortunato and Darko Hric. Community detection in networks: A user guide. *Physics Reports*, 659:1–44, 2016.
- [13] Andreas Hamm and Simon Odrowski. Term-community-based topic detection with variable resolution. Preprint, 2021.
- [14] Andreas Hamm. Complex word networks - comparing and combining information extraction methods. <https://elib.dlr.de/127501/>, May 2019. Contributed to SPCS2019, Stockholm.
- [15] Rada Mihalcea and Paul Tarau. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain, 2004. Association for Computational Linguistics.
- [16] Corina Florescu and Cornelia Caragea. PositionRank: An unsupervised approach to keyphrase extraction from scholarly documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2017.
- [17] Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972.
- [18] Jörg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Physical Review E*, 74(1), 2006.
- [19] V. A. Traag, L. Waltman, and N. J. van Eck. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1), 2019.
- [20] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.
- [21] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.