Conceptual Topic Aggregation

Klara M. Gutekunst[0009−0002−6416−7873], Dominik Dürrschnabel[0000−0002−0855−4185], Johannes Hirth[0000−0001−9034−0321], and Gerd Stumme[0000−0002−0570−7908]

[1] Knowledge & Data Engineering Group, University of Kassel, Germany
[2] Interdisciplinary Research Center for Information System Design, University of Kassel, Germany
klara.gutekunst@uni-kassel.de
{duerrschnabel, hirth, stumme}@cs.uni-kassel.de

Abstract. The vast growth of data has rendered traditional manual inspection infeasible, necessitating the adoption of computational methods for efficient data exploration. Topic modeling has emerged as a powerful tool for analyzing large-scale textual datasets, enabling the extraction of latent semantic structures. However, existing methods for topic modeling often struggle to provide interpretable representations that facilitate deeper insights into data structure and content.
In this paper, we propose FAT-CAT, an approach based on Formal Concept Analysis (FCA) to enhance meaningful topic aggregation and visualization of discovered topics. Our approach can handle diverse topics and file types – grouped by directories – to construct a concept lattice that offers a structured, hierarchical representation of their topic distribution. In a case study on the ETYNTKE dataset, we evaluate the effectiveness of our approach against other representation methods to demonstrate that FCA-based aggregation provides more meaningful and interpretable insights into dataset composition than existing topic modeling techniques.

Keywords: Topic Modeling · FCA · NLP · Text Mining · Hierarchical Topic Representation · Semantic Analysis.

1 Introduction

Current methods such as LDA, BERTopic, and Top2Vec can be used for the identification of topics in documents. However, these techniques do not leverage the intrinsic hierarchical structure of the topics encapsulated in the documents. One attempt in the realm of FCA [1] is to overcome this shortcoming by creating document-topic relations and use the concept lattice as a hierarchical representation. However, as they work on independent and homogenous documents, which are exclusively of textual nature, this method lacks the ability to aggregate information.

In this work, we introduce FCA-based Aggregation for Topics using Conceptual Analysis and Taxonomies (FAT-CAT), which integrates Natural Language Processing (NLP) with FCA, leveraging their complementary strengths

to structure and visualize relationships within data. Thereby, we can incorporate heterogenous data, consisting of different data types and content. The main contributions of this work are:

(a) Our approach integrates non-textual image data.
(b) We employ FCA to aggregate topic information effectively, facilitating insight combination across multiple directories.
(c) FAT-CAT is designed to operate with minimal training requirements, utilizing pre-trained, off-the-shelf models and thereby removing the necessity for extensive training.
(d) We demonstrate through a case study on a real-world dataset, how the approach can support a human analyst.
(e) We provide an open-source implementation, accessible on GitHub[3].

The structure of this paper is as follows: In Section 2, we review relevant literature. Section 3 presents foundations, followed by Section 4, which details our proposed methodology FAT-CAT. In Section 5, we present three baseline strategies. Section 6 outlines implementation details of our approach. In Section 7, we present a case study on a real-world dataset, including results and an empirical assessment of our approach alongside the baselines. Finally, we conclude with Section 8 and outline potential directions for future research.

2 Related Work

The authors of [2] apply FCA to a Twitter dataset to extract insights into the topics present. In their methodology, tweets serve as objects, while words occurring within the tweets act as attributes. They assume that tweets forming a formal concept, i.e., those sharing the same terminology, correspond to the same topic. Upper concepts in the resulting concept lattice represent general topics, whereas lower-level concepts reflect more specific themes. However, a key limitation of their approach is the excessive number of concepts generated, making interpretation challenging. To mitigate this, they introduce upper and lower frequency thresholds, discarding terms whose occurrence falls outside these predefined bounds, arguing that the former are too general and the latter too specific for meaningful analysis. While their approach provides valuable means as how to derive topic structures from text data, it relies purely on term frequency as attributes rather than leveraging the semantic relationships inherent in topic models, as we do in this work.

The authors of [3] employ FCA to structure topic representations derived from a topic model. They construct two matrices: One capturing document-topic probabilities obtained through their ILDA model and another representing topic-word probabilities, restricted to the top ten words per topic. Unlike our approach, which automatically derives hierarchical relationships, their method

[3] `https://github.com/KlaraGtknst/text_topic` (29.01.2025) & `https://github.com/KlaraGtknst/clj_exploration_leaks` (29.01.2025)

Table 1: Document-topic relation

Documents \ Topics	t_1	t_2	$\cdots$	t_n
d_1	$w_{1,1}$	$w_{1,2}$	$\cdots$	$w_{1,n}$
d_2	$w_{2,1}$	$w_{2,2}$	$\cdots$	$w_{2,n}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
d_l	$w_{l,1}$	$w_{l,2}$	$\cdots$	$w_{l,n}$

Table 2: Term-topic relation

Terms \ Topics	t_1	t_2	$\cdots$	t_n
s_1	$\hat{w}_{1,1}$	$\hat{w}_{1,2}$	$\cdots$	$\hat{w}_{1,n}$
s_2	$\hat{w}_{2,1}$	$\hat{w}_{2,2}$	$\cdots$	$\hat{w}_{2,n}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
s_l	$\hat{w}_{l,1}$	$\hat{w}_{l,2}$	$\cdots$	$\hat{w}_{l,n}$

relies on manual categorization of topics into broader thematic groups. These manually curated categories are subsequently visualized using a FCA-inspired order diagrams to reflect topic generalization. In contrast, our approach seeks to infer hierarchical topic structures automatically, without requiring manual intervention.

The authors of [4] also leverage FCA to structure topic hierarchies. They utilize LDA for topic modeling, encoding documents based on the distribution of topics. Similar to our approach, they introduce a thresholding mechanism to retain only relevant topics per document. However, a major divergence lies in their reliance on HowNet, a linguistic knowledge base. Their primary objective is to facilitate precise querying, using a concept lattice to provide a more nuanced representation of documents beyond mere word-level indexing. To manage the complexity of the lattice, they eliminate concepts whose attribute sets exhibit excessively high or low support. While effective for information retrieval, their approach differs from ours in fundamental ways. The focus of our approach is uncovering hierarchical relationships within topic structures across the dataset, rather than enhancing document retrieval. We do not discard high-support concepts, as they provide crucial insights into overarching themes.

Our approach builds on the work of the authors of [1] who propose extracting two relational structures from topic models. The first structure is a weighted document-topic relation, i.e., which topics t_i are present in which documents d_j. The second structure is a weighted topic-word relation, i.e., which words w_k are present in which topics t_i. These relations are subsequently used to derive incidence matrices, as visualized in Tables 1 and 2.

To compute a binary relation from the document-topic relation, they apply a thresholding approach, while for the term-topic relation, they retain only the top n words per topic. Subsequently, the authors compute the formal concepts of high support. Their approach enables a structured representation of topic hierarchies, which they demonstrate on academic papers from machine learning conferences. The approach lacks the possibility to compare subsets of the whole dataset, which might stem from an organization of files into directories. By contrast, our approach addresses this limitation by explicitly incorporating cross-directory topic structures, allowing for a more comprehensive analysis of thematic relationships spanning multiple data sources. Furthermore, our method is not restricted to only textual data.

3 Foundations

Formal Concept Analysis (FCA) The mathematical theory of FCA [5] is a method to analyze relational data in a hierarchical structure. The triple $\mathbb{K} \coloneqq (G, M, I)$ defines the *formal context*. G is the set of *objects*, M is the set of *attributes*, and $I \subseteq G \times M$ is the *incidence relation*, which describes when an object g has an attribute m. The object derivation $(\cdot)' : \mathcal{P}(G) \to \mathcal{P}(M)$ defines for any subset $A \subseteq G$ as $A' \coloneqq \{m \in M \mid \forall g \in A : (g, m) \in I\}$. The attribute derivation $(\cdot)' : \mathcal{P}(M) \to \mathcal{P}(G)$ is defined for any subset $B \subseteq M$ as $B' \coloneqq \{g \in G \mid \forall m \in B : (g, m) \in I\}$. The same notation $(\cdot)'$ is used for both derivation operators. A formal concept of $\mathbb{K}$ is a pair (A, B) where $A \subseteq G$ and $B \subseteq M$ satisfy the conditions $A' = B$ and $B' = A$. The set $\mathfrak{B}(\mathbb{K})$ denotes all formal concepts of $\mathbb{K}$. The concept lattice $\underline{\mathfrak{B}}(\mathbb{K}) \coloneqq (\mathfrak{B}(\mathbb{K}), \leq)$ is defined by the set of all formal concepts with the order relation $(A, B) \leq (C, D)$ if and only if $A \subseteq C$.

Image Captioner An image captioner is a tool to generate textual descriptions from images. In this paper, we specifically employ the Generative Image-to-text Transformer (GIT)[4] model [6]. It is a straightforward generative model comprising an image encoder and a text decoder. The image encoder is a Contrastive Learning-pretrained, Swim-like vision transformer, which outputs a flattened 2D representation of the image. This vector is then projected into D dimensions before being passed to the text decoder, a transformer module, which generates coherent textual descriptions. This pre-trained model has the ability to generate meaningful captions without the need for custom training, enabling us to seamlessly integrate visual information into our topic modeling process.

Top2Vec The workflow of the pre-trained topic model Top2Vec [7], which is employed in this study, is illustrated in Figure 1. Top2Vec's ability to embed texts into the same vector space as words and topics is particularly advantageous, as it allows for semantic similarity calculations between texts, words, and topics. The model encodes texts using a multilingual Universal Sentence Encoder (USE) model [8], supporting 16 languages, which is crucial for working with diverse datasets containing multiple languages. The documents are first embedded into a 300-dimensional space, then reduced to 5 dimensions and finally clustered into topics.

TITANIC Computing concept lattices for large datasets presents a significant challenge, as the size of a concept lattice can grow exponentially with respect to the size of the context. Consequently, complete visualizations of large concept lattices may obscure important hierarchical relationships within the data, hindering human interpretation. To address this issue, [9] introduced the TITANIC algorithm, which computes iceberg concept lattices. An iceberg concept lattice includes only the top-level concepts, specifically those corresponding to frequent attribute sets.

[4] `https://huggingface.co/microsoft/git-base` (29.01.2025)

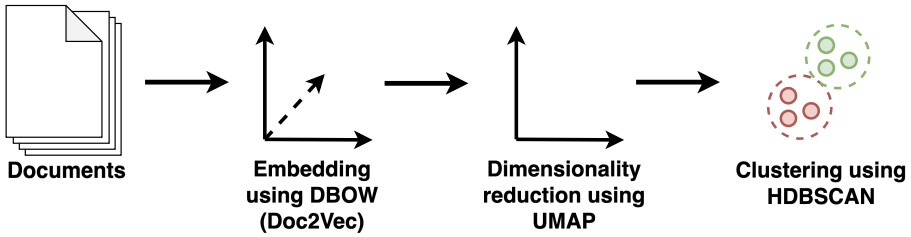


Fig. 1: Visual representation of the Top2Vec workflow for deriving text clusters.

The authors define a minimum support threshold, $minsupp \in [0, 1]$, below which attribute sets are not considered frequent. An attribute set $B \subseteq M$ is deemed frequent in $\mathbb{K}$ if its support

$$supp(B) = \frac{|B'|}{|G|}$$

meets or exceeds the $minsupp$ threshold. A concept (A, B) is considered frequent if its intent B constitutes a frequent attribute set. The iceberg concept lattice of a context $\mathbb{K}$ comprises all such frequent concepts.

4 Method

The *FCA-based Aggregation for Topics using Conceptual Analysis and Taxonomies (FAT-CAT)* methodology introduced in this paper outlines a systematic workflow for deriving the directory-topic concept lattice of a dataset. First, we extract the text from the data files. The processing of text-related fields follows a distinct workflow, as shown in Figure 2.

By employing a topic model, specifically the Top2Vec model, we extract the topics of the documents. While the model typically returns only one topic

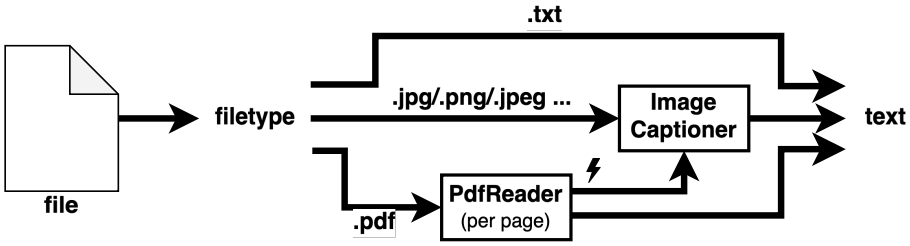


Fig. 2: Visual representation of the workflow for extracting text from the data files. If a page of a PDF document contains no text or raises an error during extraction, the image captioner is used to generate a textual description of the image of that page.

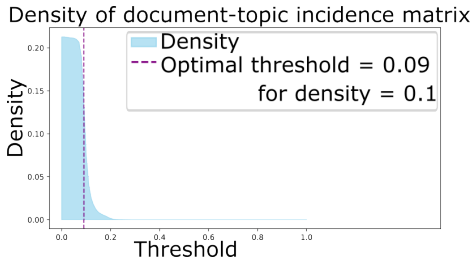


Fig. 3: Computation of threshold δ such that the resulting thresholded binary document-topic incidence $\mathcal{D}_\delta$ has a density of $\frac{|\mathcal{D}_\delta|}{|D \times T|} = 0.1$.

per document, we retrieve the top ten topics for each document to capture a broader range of thematic content. Additionally, the model provides topic scores, topic words, and word scores, which we leverage to derive document-topic relationships, as proposed by [1] (cf. Table 1). These topics are assigned real values $w_{i,j} \in \mathbb{R}_{\geq 0}$, which indicate the relevance of a respective topic t_i to document d_j. Next, these weights are row-normalized. The binary document-topic incidence matrix is then created by applying a threshold δ, which helps in determining the most relevant topics to each document. If $w_{i,j} \geq \delta$, topic t_i is present in document d_j. The threshold is the first value such that the density of thresholded weights is below a certain value. The computation is visualized in Figure 3. With accordance to [1], we set the density threshold to 0.1 to obtain a reasonably sparse binary incidence. After δ is determined, we obtain the binary document-topic incidence matrix

$$\mathcal{D}_\delta(i,j) = \begin{cases} 1 & \text{if } \ w_{i,j} \geq \delta, \\ 0 & \text{else.} \end{cases}$$

This matrix serves as the foundation for understanding how topics are distributed across different directories.

Compared to the approach of Hirth and Hanika [1], we aggregate the binary document-topic incidence matrices from all directories to gain a broader understanding of the dataset's topic structure as a whole. The workflow is visualized in Figure 4.

We adopt the TITANIC algorithm as it generates interpretable concept lattices from large datasets, making it well-suited for our study. For each directory, we compute an iceberg concept lattice. We focus on high-support concepts because their attribute sets represent topics that are consistently present across many documents in a directory. These frequent topics capture broader themes that are widely distributed throughout the respective directory, making them ideal for constructing a representative set of attributes. The minimum support is a parameter that should be chosen according to the degree of detail in which the data ought to be explored. According to the experience of the authors, a starting parameter of 0.9 is recommendable. The sequence of frequent concepts over multiple directories is then used to build the incidence matrix presented in

Table 3. This aggregation allows for a hierarchical investigation of the dataset, where more general topics appear at the top of the hierarchy and more specific topics are found towards the bottom.

Table 3: Weighted directory-topic relation. The weights $\tilde{w}_{i,j} \in \{0, 1\}$ represent whether a topic t_j is present in a directory $\hat{d}_i$.

Directories \ Topics	t_1	t_2		t_n
$\hat{d}_1$	$\tilde{w}_{1,1}$	$\tilde{w}_{1,2}$	$\cdots$	$\tilde{w}_{1,n}$
$\hat{d}_2$	$\tilde{w}_{2,1}$	$\tilde{w}_{2,2}$	$\cdots$	$\tilde{w}_{2,n}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\hat{d}_k$	$\tilde{w}_{k,1}$	$\tilde{w}_{k,2}$	$\cdots$	$\tilde{w}_{k,n}$

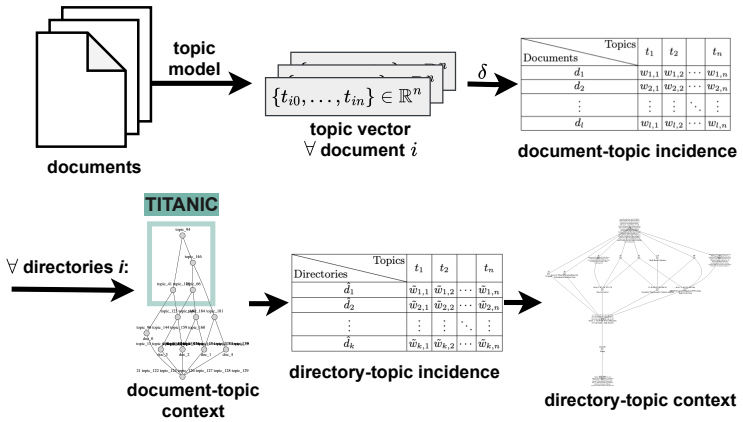


Fig. 4: Workflow for deriving the directory-topic concept lattice.

5 Baselines

The following sections outline common topic visualization strategies serving as baselines for our approach. We begin with word clouds, a simple, frequency-based visualization baseline of the most common words in a document. Next, we describe the Embedding - Dimension Reduction (EDR) approach that clusters low-dimensional embeddings of the documents by directories. We then explore Clustering Named Entitys (CNE).

Word clouds An initial approach for visualizing topics involves using word clouds, a simple and intuitive method for displaying the most frequent words in a document. A word cloud is a visual representation of text data where words are displayed in varying sizes based on their frequency or importance. It helps identify key themes and trends in textual content at a glance. Since we are interested in the content of the directories, we generate word clouds for each directory in the dataset. The Figures 10a and 10b show examples of word clouds.

Embedding - Dimension Reduction (EDR) The second baseline approach EDR, whose workflow is illustrated in Figure 5, aims to cluster documents by directories to investigate whether the files within a directory exhibit semantic similarity in terms of content. To achieve this, the text from the files is embedded with the Sentence-BERT (SBERT) model in a vector space such that distance correlates with semantic dissimilarity. Subsequently, dimensionality reduction is employed to enable visualization via a scatter plot. For this baseline we tried the three different dimensionality reducing algorithms Principal Component Analysis (PCA), Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP), and t-Distributed Stochastic Neighbor Embedding (t-SNE) (cf. Figure 11) and settled on t-SNE, as the other two empirically lacked the ability to clearly separate of the data points. The directory names serve as class labels for the files, i.e., they give the color of the presented points. To make the method applicable to large-scale datasets, the directory names are preprocessed to reduce the number of classes. Specifically, directory names that contain more than 50% numerical characters are replaced with the label "numbers", while any remaining directory names containing numbers have the numbers stripped. Directory names with fewer than five alphabetic characters are replaced with the label "chars".

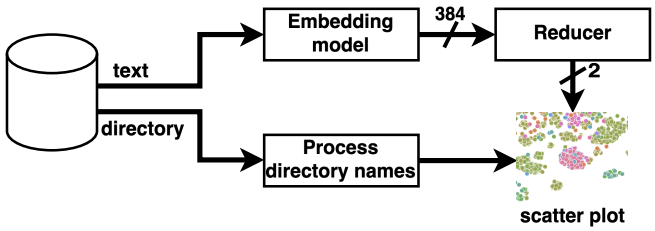


Fig. 5: Illustration of the dimensionality reduction process for document embeddings colored by directory labels. The embedding model SBERT omits 384 dimensional embeddings for each text. They are reduced to two dimensions using PCA, UMAP or t-SNE. The directory names are preprocessed to group similar ones and reduce the number of classes since they are used to color the points in the scatter plot.

Clustering Named Entitys (CNE) The baseline CNE builds on Named Entity Recognition (NER), which identifies and classifies entities into predefined categories, such as organizations, locations, dates, and other relevant terms [10]. Thereby, NER provides a method for categorizing text. The NER workflow, as shown in Figure 6, begins with truncating the text to meet the input length constraints of the `nlp` object. The NER model is then applied to the text, restructuring the identified Named Entitys (NEs) so that each entity category functions as a key mapping to a list of corresponding NEs. This structure allows for the efficient retrieval of named entities by category, facilitating their use in subsequent analyses.

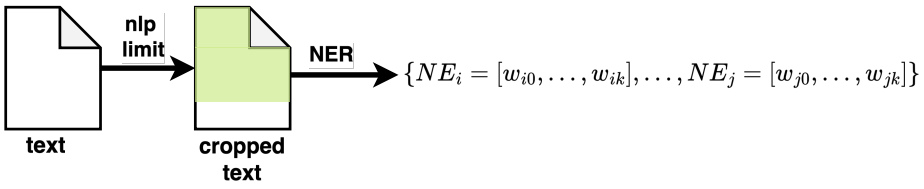


Fig. 6: Visual representation of the workflow from text processing to the structure storing NEs for a given document.

In this baseline, we disregard the hierarchical directory structure and focus on the clustering of NEs for distinct categories. The approach of clustering of NEs has been proposed by [10]. The workflow of this baseline is illustrated in Figure 7. First, NER is applied to all texts. While [10] replaces similar NEs with a single version, referred to as soft entity linking, we omit this step, though it may improve clustering results.

Subsequently, NEs are embedded using a language model that produces embeddings comparable via cosine similarity. This study employs the pre-trained Word2Vec model. For each NE category i, the $N = 50$ most frequent NEs are extracted. A symmetric $N \times N$ similarity matrix is computed for the most frequent NEs within each category, using cosine similarity as a similarity measure. Each row of the similarity matrix is treated as the feature vector of a given NE. These NEs are then clustered using the k-Means algorithm, with $k = 5$. Both N and k are hyperparameters which can be optimized.

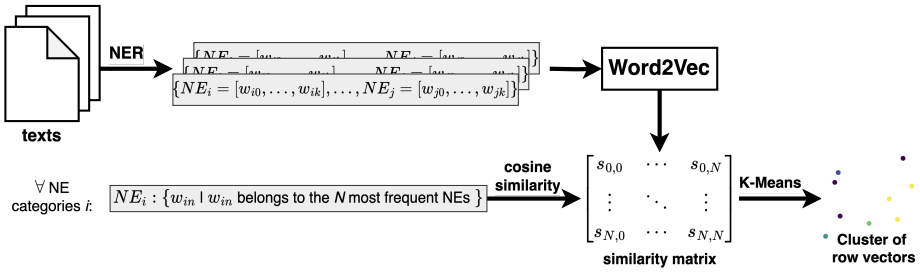


Fig. 7: Illustrates the workflow for deriving a clustering of the NE categories i.

6 Implementation

This section details the implementation of the key components of our approach. Several FCA tools used in this work are implemented by the `conexp-clj` library [11]. We derive iceberg concept lattices via the TITANIC algorithm. We choose a minimum support threshold of 0.9. For text visualization via word clouds, we utilize the Python library `wordcloud`[5] which is employed with its default parameters. To extract NEs, we employ the pre-trained `en_core_web_sm` pipeline from the spaCy library, serving as the `nlp` object. This lightweight model, trained on English text, provides a general-purpose pipeline for tagging, parsing, lemmatization, and NER[6]. However, this implementation is constrained to a maximum of 10^6 tokens[7].

7 Case Study

To assess the applicability of our approach FAT-CAT, we apply our method alongside three baselines to the ETYNTKE dataset. The next section describes the dataset, followed by an evaluation of our approach against baseline techniques.

7.1 Dataset

The dataset utilized in this study, referred to as Everything You Need To Know Ever (ETYNTKE)[8], comprises approximately 101 GB of images, text documents, and other file types uploaded in 2015. It covers a diverse range of topics, including politics, weaponry, and computer science, providing a rich and varied foundation for extracting semantic insights. This diversity is particularly suited for the application of our FCA approach, as it enables the exploration of patterns across multiple domains, showcasing the versatility of the methodology.

[5] `https://github.com/amueller/word_cloud` (01.02.2025)
[6] `https://spacy.io/models` (13.02.2025)
[7] `https://spacy.io/api/language` (30.01.2025)
[8] `https://archive.org/details/ETYNTKE` (29.01.2025)

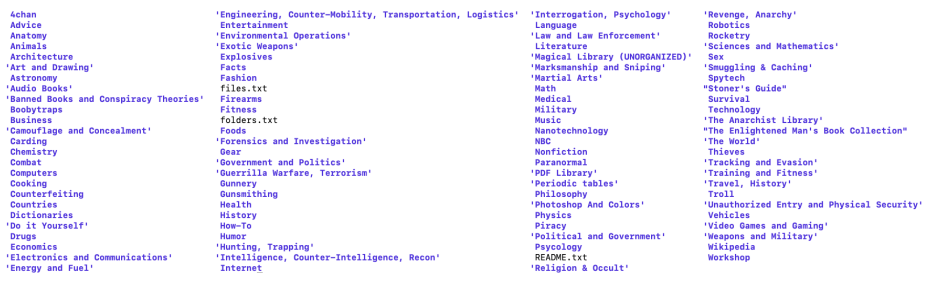


Fig. 8: The top-level directory structure of the ETYNTKE dataset, excluding the README.txt file.

Structurally, the dataset follows a hierarchical organization, with files arranged in folders and subfolders. The top-level directory structure is illustrated in Figure 8. This organization presents an opportunity to investigate whether the directory structure accurately reflects the content of the files, which has not been examined in previous analyses. The presence of malformed files, which cannot be opened or read, adds an additional layer of complexity, making the dataset an ideal challenge for testing the robustness and adaptability of our approach.

The large size, heterogeneity, and hierarchical structure of ETYNTKE make it an ideal candidate for applying our FAT-CAT methodology, as we can leverage the organizational structure in our approach. Its variety of content, along with the potential inconsistencies in its organization, provides a comprehensive testbed for exploring how NLP and FCA can work together to extract meaningful insights and structure complex datasets.

7.2 FAT-CAT

The lattice resulting from the ETYNTKE dataset is illustrated in Figure 9. Topics, represented by numbers above the nodes, propagate downward to transitive adjacent nodes. Consequently, every node connected to and below node 94 also contains topic 94. Topics associated with higher nodes are more common among directories, whereas lower nodes contain more specific topics found in fewer directories. The uppermost node represents topics shared across all directories. In this case, no such topic exists. Ranges of consecutive topic numbers are denoted by their lowest and highest IDs, separated by a hyphen. A subset of topic words for a selection of topic IDs is given in Table 4. As directories propagate upwards, lower nodes correspond to fewer directories.

A motif is a statistically significant subgraph or pattern [13]. The presence of such structural patterns, which can be identified in Figure 9, including both contranominal and ordinal motifs, suggests promising opportunities for analysis. In accordance with expectations, the directories at the bottom of the graph, which are considered the most specific in terms of topic categorization, predominantly contain subdirectories. It is noteworthy that the only directory present at both the highest level of the directory hierarchy and the bottom node is

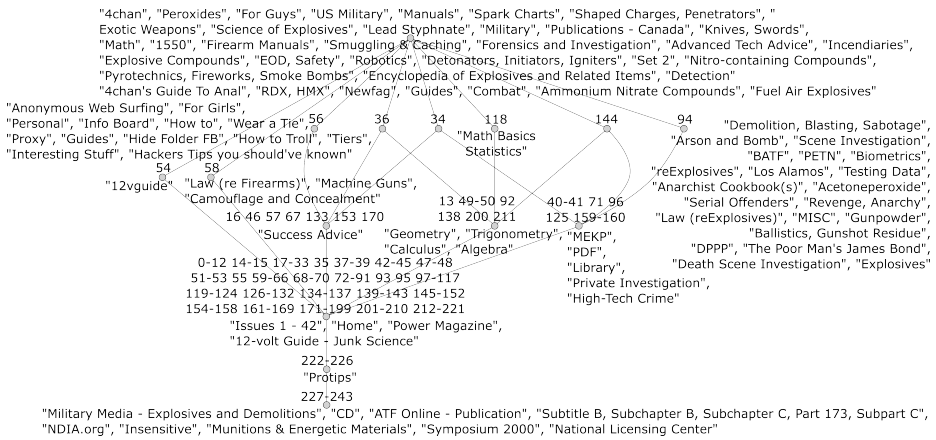


Fig. 9: Concept lattice representing the directory-topic incidence of the dataset ETYNTKE. The numbers correspond to topic IDs, with each topic characterized by 50 words, as determined by the Top2Vec model. A subset of the topic IDs is mapped to corresponding terms in Table 4. The text entries below the nodes represent directory names.

the `Military` directory. With respect to content, some directories exhibit intriguing thematic associations. The `High-Tech Crime` directory, for instance, comprises topics related to explosives, while the `Success Advice` directory additionally features content concerning chemicals. The top node of the ordinal motif, which aggregates the incoming edges, introduces philosophical content. Consequently, the content of directories positioned at the lowermost levels of the graph, which predominantly comprise subdirectories, encompasses a greater proportion of philosophical subjects in comparison to directories positioned at more elevated levels within the graph and, in some cases, the directory hierarchy of the dataset.

Our approach FAT-CAT enables us to explore large datasets without the need for manual inspection of each document. The hierarchical structure of the concept lattice offers insights into the relationships between topics, facilitating a deeper understanding of how different themes interconnect across directories. However, it is important to note that while this approach offers valuable insights, the visual representation may not be intuitive to all users, particularly those unfamiliar with concept lattices or FCA. Furthermore, the topics are only represented by their IDs which are not meaningful to the reader. Due to the nature of the topic model used to derive the topics, it is the only possibility to display the topic IDs, because the space is too limited to display 50 words per topic. Unfortunately, this necessitates the manual translation of the topic IDs to words by looking them up in an additional JSON file. Despite these limitations, the method provides a powerful way to explore and analyze datasets in a structured and scalable manner.

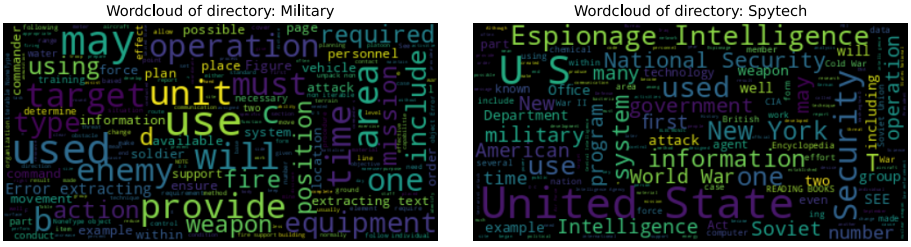
Table 4: Subset of translated topic IDs and their top 5 descriptive words for the directory-topic concept lattice.

Topic ID	Topic Words				
34	gun	weapons	rifles	pistol	firearm
36	terminological	nonverbal	linguist	codeword	longtext
54	microelectron	microkernels	microparticle	microphysics	microcosmic
56	pythonpath	pythonmac	python_version	python_object	pythonpowered
58	nuclear	hazardous	explosive	nanotoxicity	chernobyl
94	project	fundraiser	donation	volunteering	charity

7.3 Baselines

Word clouds These word clouds not only provide an overview of the directory's content based on the most frequent terms in the directory but also allow identifying directories noteworthy of further inspection. Intuitively, if the words of a word cloud do not align with the expected content of the directory name, or if a word cloud includes terms of particular relevance, one can consider the directory in further data mining. While word clouds are intuitive and simple to interpret, they have limitations. They only display the most frequent words, which may not fully represent the semantic relationships between terms. Moreover, the quality of the word clouds is dependent on the quality of the text extraction process (cf. Figure 2). Furthermore, by inspection of the word clouds for the directories `Military` and `Spytech`, respectively displayed in Figure 10a and 10b, it becomes evident that it is complicated to compare the content of different directories.

Embedding - Dimension Reduction (EDR) Ideally, the dimensionality reduction algorithm should separate the clusters of documents associated with different directories. Although t-SNE provides the best separation among the reducer algorithms, the large number of directories results in too many classes to distinguish, as shown in Figure 11. Consequently, this approach is not well-suited for datasets of this size, since it is difficult to extract meaningful clusters from the reduced embeddings, as the colors become too similar.



(a) `Military` (b) `Spytech`

Fig. 10: Wordcloud visualisations for selected directories.

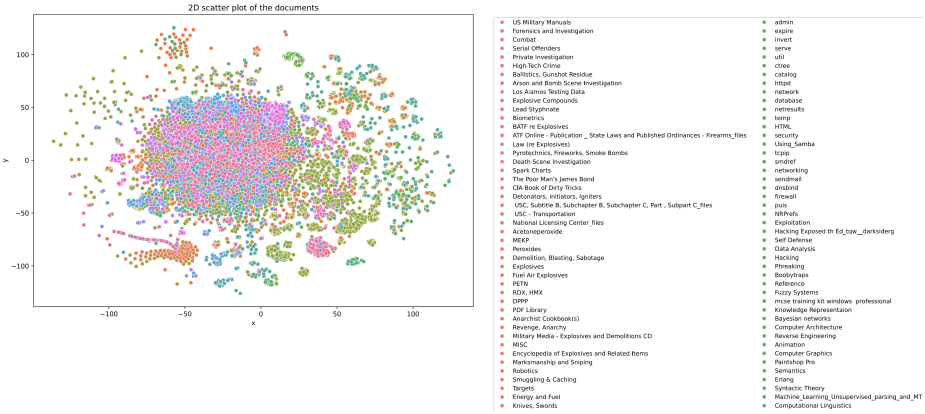


Fig. 11: Scatter plot of document embeddings reduced to two dimensions using t-SNE, colored by directory labels.

Clustering Named Entitys (CNE) This baseline is based on entity categories, in contrast to our FAT-CAT, which relies on the hierarchical directory structure of the dataset. A visualization of the clustering of the NE category `LANGUAGE` is shown in Figure 12. It reveals that the approach effectively separates programming languages, such as *C++*, from spoken languages. Most language families[9] are grouped together. However, as with the NER process, entities are often incorrectly categorized, e.g., *False*. Although some patterns in the class members can be observed, neither the topological clusters nor the clustering results provide clear or accurate information. Overall, this approach does not yield meaningful insights into the data's structure.

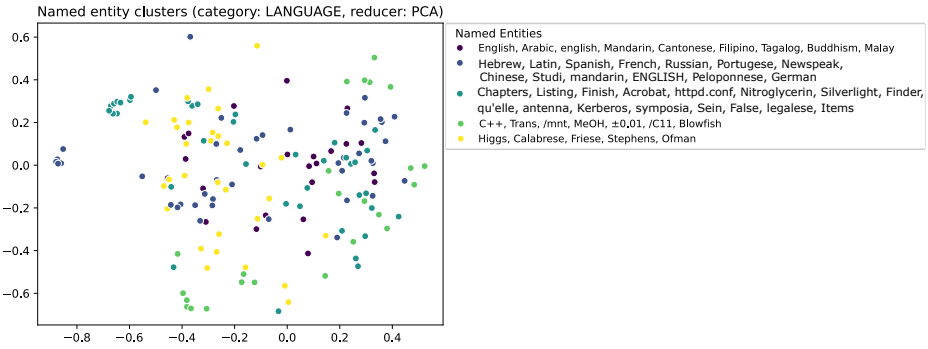


Fig. 12: Clustering of NEs of the category `LANGUAGE`.

[9] `https://www.omniglot.com/writing/langfam.htm?` (01.02.2025)

7.4 Empirical Comparison

In this section, we compare different approaches for analyzing directory content, assessing their strengths and limitations. Our evaluation demonstrates that the proposed FCA-based method FAT-CAT offers superior insights into the dataset's structure.

Word clouds provide an intuitive overview by highlighting frequently occurring terms, enabling the identification of noteworthy directories. However, they fail to capture semantic relationships and depend heavily on the text extraction quality. Additionally, comparing multiple directories using word clouds is challenging due to their unstructured nature. In contrast, embedding-based visualization techniques allows directly comparing directories based on their content. Not only does the quality of the reducers affect the results, but the large number of directories also complicates the interpretation of the visualization. This approach is therefore ineffective for large-scale datasets. NER-based clustering of category entities suffers from frequent misclassification. Moreover, the resulting clusters neither exhibit clear patterns nor provide meaningful structural insights into the content of directories, making this method unreliable for data analysis. Our proposed FCA-based approach enables a hierarchical exploration of directory topics without requiring manual document inspection. Although the visualization is less intuitive for non-experts and topics are represented by numerical IDs rather than meaningful terms, this method remains the most effective for uncovering the relationships between directory topics. Among the evaluated methods, FAT-CAT proves to be the most effective for structured data exploration.

8 Conclusion and Future Work

In this paper, we proposed a method for analyzing structures in an unknown dataset of diverse content, and compared it to three baseline methods. In the case study, we found that our approach FAT-CAT is better suited for exploring and visualizing topics across multiple directories than the baseline techniques. While word clouds offer a quick overview, and embedding-based methods attempt clustering, neither approach provides sufficient clarity for the large dataset of the case study. NER clustering is unreliable due to misclassification. In contrast, FCA enables hierarchical and systematic topic analysis, making it the preferred method despite requiring expertise for interpretation.

Currently, FAT-CAT's results do not provide any insight into the dataset's directory hierarchy. However, this structural information could be highly valuable when considered alongside the topic hierarchy of the directories. Therefore, future work could explore integrating both sources of information to enhance the analysis.

In this work, we employed Top2Vec [7] as the topic modeling technique. While newer models, such as C-Top2Vec [12], have been introduced, we opted for Top2Vec due to its ability to represent documents using a single-vector format. This approach allows for topic aggregation by counting the occurrences

of topics across the documents in a directory. In contrast, C-Top2Vec supports multi-vector document representation, which would provide a more granular description of individual documents and the dataset as a whole. Future work could explore and evaluate the performance of C-Top2Vec in comparison to Top2Vec for this purpose.

References

1. Hirth, J. and Hanika, T.: The Geometric Structure of Topic Models. In: arXiv (2024).
2. Cigarrán et. Al.: A step forward for Topic Detection in Twitter: An FCA-based approach. In: Expert Systems with Applications (2016).
3. Wang, K. and Wang, F.: Topic-Feature Lattices Construction and Visualization for Dynamic Topic Number. In: Journal of Systems Science and Information (2021).
4. Wang et. Al.: Dynamically constructing semantic topic hierarchy through formal concept analysis. In: Multimedia Tools and Applications (2023).
5. Ganter, B. and Wille, R.: Formal concept analysis: mathematical foundations. Springer Nature, (2024).
6. Wang et. Al.: GIT: A Generative Image-to-text Transformer for Vision and Language. arXiv (2022).
7. Dimo Angelov: Top2Vec: Distributed Representations of Topics. In: arXiv (2020).
8. Yang, Yinfei, et. Al.: Multilingual universal sentence encoder for semantic retrieval. In: arXiv (2019).
9. Stumme et. Al.: Computing iceberg concept lattices with Titanic. In: Data & Knowledge Engineering (2002).
10. Alsudais, A. and Tchalian, H.: Clustering Prominent Named Entities in Topic-Specific Text Corpora. In: Twenty-fifth Americas Conference on Information Systems 2019 (2019).
11. Hanika, T. and Hirth, J.: Conexp-Clj: A Research Tool for FCA. In: ICFCA (Supplements), vol. 2378, pp. 70-75, 346 (2019).
12. Angelov, D. and Inkpen, D.: Topic Modeling: Contextual Token Embeddings Are All You Need. In: Findings of the Association for Computational Linguistics: EMNLP 2024 (2024).
13. Hirth et. Al.: Ordinal motifs in lattices, vol. 659, Information Sciences (2024)