Multi-view learning for hierarchical topic detection on corpus of documents

Juan Camilo Calero Espinosa

Universidad Nacional de Colombia
Facultad de Ingenieria, Departamento de Ingenieria de Sistemas e Industrial.
Bogotá, Colombia
2021

Multi-view learning for hierarchical topic detection on corpus of documents

Juan Camilo Calero Espinosa

Tesis presentada como requisito parcial para optar al título de:
Magister en Ingeniería de Sistemas y Computaciǿn

Director:
Ph.D. Luis Fernando Niño V.

Línea de Investigación:
Procesamiento de lenguaje natural
Grupo de Investigación:
Laboratorio de investigación en sistemas inteligentes - LISI

Universidad Nacional de Colombia
Facultad de Ingenieria, Departamento de Ingenieria en Sistemas e Industrial.
Bogotá, Colombia
2021

To my parents Maria Helena and Jaime.
To my aunts Patricia and Rosa.
To my grandmothers Lilia and Santos.

Acknowledgements

To Camilo Alberto Pino, as the original thesis idea was his, and for his invaluable teaching of multi-view learning. To my thesis advisor, Luis Fernando Niño, and the Laboratorio de investigación en sistemas inteligentes - LISI, for constantly allowing me to learn new knowledge, and for their valuable recommendations on the thesis.

Abstract

Topic detection on a large corpus of documents requires a considerable amount of computational resources, and the number of topics increases the burden as well. However, even a large number of topics might not be as specific as desired, or simply the topic quality starts decreasing after a certain number. To overcome these obstacles, we propose a new methodology for hierarchical topic detection, which uses multi-view clustering to link different topic models extracted from document named entities and part of speech tags. Results on three different datasets evince that the methodology decreases the memory cost of topic detection, improves topic quality and allows the detection of more topics.

Keywords— Named entities, POS tagging, topic detection, multi-view clustering, graph fusion

Resumen

La detección de temas en grandes colecciones de documentos requiere una considerable cantidad de recursos computacionales, y el número de temas también puede aumentar la carga computacional. Incluso con un elevado nùmero de temas, estos pueden no ser tan específicos como se desea, o simplemente la calidad de los temas comienza a disminuir después de cierto número. Para superar estos obstáculos, proponemos una nueva metodología para la detección jerárquica de temas, que utiliza agrupamiento multi-vista para vincular diferentes modelos de temas extraídos de las partes del discurso y de las entidades nombradas de los documentos. Los resultados en tres conjuntos de documentos muestran que la metodología disminuye el costo en memoria de la detección de temas, permitiendo detectar màs temas y al mismo tiempo mejorar su calidad.

Keywords— Entidades nombradas, etiquetado gramatical, detección de temas, agrupamiento multi-vista, fusión de grafos

Esta tesis de maestría se sustentó el 14 de mayo de 2021 a las 08:00 am,
y fue evaluada por los siguientes jurados:

Sergio Gonzalo Jiménez Vargas, Ph.D.
Grupo de investigación de lingüística Computacional
Instituto Caro y Cuervo

Elizabeth León Guzmán, Ph.D.
Profesora Asociada
Departamento de ingeniería de Sistemas e Industrial
Universidad Nacional de Colombia, Sede Bogotá

Contents

List of Tables

List of Figures

1. Introduction

A part-based representation of the world is important since there are psychological [1] and physiological [2] [3] evidences that the brain represents the world based on perceptions of its parts [4]. Arranging and representing data on smaller parts can help to understand their underlying structure. In scenarios like natural language processing, this can be done by modeling a dataset as a hierarchy of topics. In addition to serving as a representation of the data, this hierarchy has a broad range of potential applications, e.g., it can help in tasks like summarization, guided browsing, categorization, trending topic identification, among others.

Organizing topics in a hierarchy might not be necessary, but this can improve the categorization quality as more granularity is available at the time of performing some tasks thanks to the provision of different level of categories, e.g. when categorizing new data, there might not be enough features for some data, making it difficult to fit the data in any category of the lowest level, but with more general categories in higher levels, data might fit in some of them. In a huge corpus of documents, a myriad of topics can be extracted, and in the beginning of a search, users might just want to see general topics, and descend on the topic tree as search results get narrower. For these reasons, hierarchical topic detection can be an important tool for the processing of big collections of documents as lots of topics can be created and organize them in hierarchies brings additional value to the topics.

It is common to perform data cleansing to datasets before topic extraction. Although techniques such as stemming, lemmatization or stop-word removal perform really well, in some cases, with highly noisy data, more aggressive filters might be needed. For example, text extracted from PDF files or from images (such as scanned documents) can have lots of meaningless words, random symbols, inherent words to some formats (e.g., HTML tags, URLs, table formats, etc.), non-printable characters, words in different languages, etc. In other cases, noisy data is not the problem, but just certain types of words are desired to process, e.g., in a dataset from a newspaper containing political profiles, a topic model of just people names can be more interesting or revealing than a topic model with all the words, or a corpus about animals might not have many names of people, but rather a lot of nouns, locations and adjectives instead. Thus, depending on the dataset, focusing on just the important words can reduce noisy data and increase the effectiveness of topic modeling.

For this reason, we perform topic detection on just specific types of entities, thus reducing drastically the input data for the topic model. With no extra components in the topic model, the required resources for running it are significantly lower. Most topic models are based on a bag of words and co-occurrences of words, and in certain scenarios, detecting co-occurrences of e.g. just people or organizations, might be more relevant than identifying co-occurrences of all the words in the dataset. But still detecting relations of people for instance with something in common but not co-occuring in any document is not yet resolved; so, we propose multi-view topic clustering where additional relations between the topics can be found. For example people doing the same things on the same places can be linked together even if no document talks about them at the same time, as the verbs and places in them do co-occur.

Access to academic documents in research repositories, such as technical reports, final papers, thesis, and research articles, should exploit the semantic relationships between the themes and terms used. Depending on the interest of the user, it should be possible to have hierarchies of automatically constructed topics that facilitate the construction of bibliographies for further research. In this thesis, we propose a way to improve hierarchical topic detection using additional NLP and multi-view techniques. The methodology is tested on a set of academic documents in the Universidad Nacional de Colombia institutional repository and also on a set of Wikipedia pages.

We propose a 14 step methodology depicted in figure 1.1 with a corpus of documents as input data. If documents are not in plain text, a step of text extraction is performed using Apache Tika[1], which can extract metadata and text from over a thousand different file types (such as PPT, XLS, DOC, PDF, etc.).

In the end, we get a bag of topic models that are linked through clustering, allowing to use them jointly to improve the usefulness of topic detection as demonstrated in [27]. This bag of topic trees are summarized, visualized an evaluated through a proposed heuristic. The 14 steps of the methodology are:

1. Text extraction via Tika. If documents are already in plain text, this is not necessary.
2. Data cleansing (removal of non printable characters, HTML tags, empty, damaged or encrypted files, encoding errors, etc.)
3. Language detection filter. This is important as pre-trained models for named entity recognition (NER) and Part-of-speech (POS) tagging just perform well on the language they were trained on.
4. Part-of-speech tagging. A new dataset containing the same documents is created per tag, but documents just have the words belonging to that tag.

[1] https://tika.apache.org/

5. Named entities recognition. A new dataset containing the same documents is created per entity, but documents just have the words belonging to that entity.
6. Lemmatization and stop-words filter per tag and entity sub-dataset.
7. Top N words selection per tag and entity sub-dataset according to an average TF-IDF score.
8. Topic detection via HLTA [45] per tag and entity sub-dataset.
9. Evaluation of the different topic models.
10. Topic view extraction.
11. Topic multi-view clustering via GFSC [52].
12. Topic fusion.
13. Fusion topic evaluation.
14. Visualization of topics.

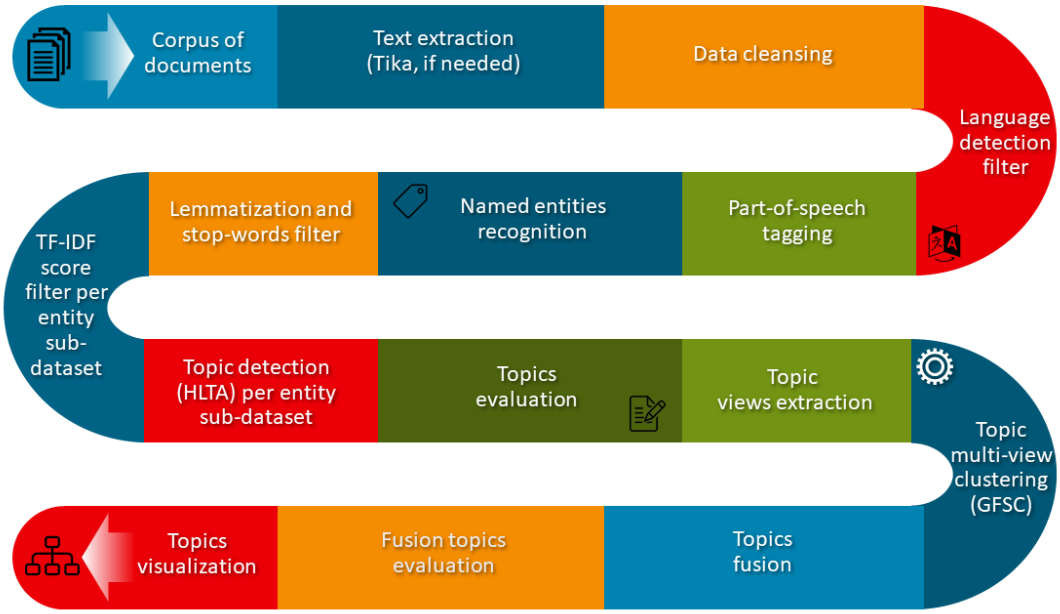

Figure 1.1.: Methodology diagram.

In the following chapters, we elaborate on the steps of the proposed methodology. Chapter 2 explains the theoretical framework of the thesis. Chapter 3 presents the datasets used in this work, and depending on the dataset, the applied data cleansing (steps 1 and 2) is shown. NER and POS tagging (steps 3, 4, 5, and 6) are explained in this chapter as well along with the detection of the bag of entity topics. The applied multi-view clustering is described in chapter 4 (steps 10 and 11). Topic fusion and evaluation (steps 12 and 13) are explained in chapter 5, and finally, some visualization of topics (step 14) are illustrated in appendixes A, C, and D.

Summarizing, in this work we present a new methodology for creating a bag of entity topics which is much more memory-efficient than traditional methodologies. And also, we propose a new approach for clustering topics using multi-view learning, that used together with a graph fusion algorithm, it is possible to evaluate, visualize and find relations among topics in an easier way.

2. Theoretical Background

2.1. Hierarchical topic detection

In the field of automatic topic detection, a topic is a set of words that tend to co-occur with high frequency in a corpus of documents. Using this set of words and the latent space inferred from the data, a soft clustering of the documents is possible. Some models are called mixed-membership models, because the sum of the probabilities of the topics present in a document must be 1. On the other hand, models where a document can belong to several topics with a probability of 1, whereby the sum of probabilities does not have to be 1, are called multi-membership models.

One of the most popular topic model is Latent Dirichlet Allocation (LDA) [5], which is a generative and mixed-membership model, where each document is made from a list of topics β. The topic distribution vector (θ_d) for each document is drawn from a Dirichlet distribution. The word distribution (ϕ_t) of each topic is also drawn from a Dirichlet distribution. The generative process consists that for each word in a document, a topic is selected based on a multinomial distribution, and then, using that topic, a word is sampled following another multinomial distribution. Given a corpus of documents, the generative process is inferred using Gibbs sampling or variational inference.

The purpose of hierarchical topic detection (HTD), is to detect topics with different abstraction level, that is to say, to create a tree of topics where topics that are in the low levels of the tree are as specific as possible, and as higher levels are visited, more general topics are found. This gives the opportunity to organize information with different levels of granularity, which can be useful in a wide range of applications, such as information retrieval, search engines, social network ans sentiment analysis, and in general, problems where finding latent variables is paramount, making HTD and important tool in other domains like image and audio processing, genomics, robotics, among others.

Several proposed HTD models are based on LDA, among these methods, the nested Chinese restaurant process (nCRP) [6, 12], Pachinko allocation model (PAM) [11, 13] and nested hierarchical dirichlet process (nHDP) [42] are included. The principal drawback of these methods, is that either for performance considerations, or for the properties of the model, it is required for the user to provide the structure of the hierarchy, namely, the number

of nodes at each level, and the number of levels, which are usually set to 3 levels due to computational costs.

The authors of [45], presented a new multi-membership model, Hierarchical Latent Tree Analysis (HLTA), and unlike LDA-based models, HLTA does not uses a document generation process, rather, it is based on hierarchical latent tree models (HLTMs).
Hierarchical latent class models have been used for cluster analysis [9] and the term "latent tree model" was introduced in [16, 17], which refers to a tree-structured Bayesian network. In the case of HLTA, the latent variables at the first level represent word co-occurrence patterns, while the latent variables at higher levels represent co-occurrence patterns of the latent variables discovered at the level below. This is done by using a set of novel algorithms, among the most important, "Build Islands" and "Bridge Islands" on the pointwise mutual information (PMI) of each pair of words or latent variables at each level. The model is optimizes making use of expectation–maximization (EM) or through a faster method called progressive EM.

Another important aspect of HLTA, is that the authors presented the results with topic quality metrics that are independent from the latent space behind the model. This is important as shown in the study performed in [19], topic models that are only focused on metrics like held-out likelihood, may create more irrelevant topics semantically speaking.
Based on two metrics, topic coherence score [26] and topic compactness score [46], HLTA outperforms other topic modeling methods.

The topic coherence score is applied to the M words $W^{(t)} = \{w_1^{(t)}, ..., w_M^{(t)}\}$ of a topic t. Equation 2.1 illustrates how the score is calculated, where $D(w_i^{(t)})$ represents the number of documents containing word $w_i^{(t)}$, and $D(w_i^{(t)}, w_j^{(t)})$ returns the number of documents containing both words $w_i^{(t)}$ and $w_j^{(t)}$.

$$Coherence\,(W^{(t)}) = \sum_{i=2}^{M} \sum_{j=1}^{i-1} log \frac{D(w_i^{(t)}, w_j^{(t)}) + 1}{D(w_i^{(t)})} \tag{2.1}$$

The compactness score is based on the similarity of the M words $W^{(t)} = \{w_1^{(t)}, ..., w_M^{(t)}\}$ of a topic t. The compactness score is given by equation 2.2, where $S(w_i^{(t)}, w_j^{(t)})$ represents the similarity between words $w_i^{(t)}$ and $w_j^{(t)}$:

$$Compactness\,(W^{(t)}) = \frac{2}{M(M-1)} \sum_{i=2}^{M} \sum_{j=1}^{i-1} S(w_i^{(t)}, w_j^{(t)}) \tag{2.2}$$

The similarity can be calculated in different ways, in this case, it comes from the cosine similarity of words in a vector representation from a word2vec model [34, 35, 36] trained on

part of Google News dataset (about 100 billion words).

The compactness score of a topic model is calculated as the average of all the topics compactness scores. The same is applied to the coherence score, where it is calculated as the average of all the topics coherence scores. In both cases, the greater the score, the better. However, generally, the more words used, the lower the scores, consequently, as proposed in [45], only four words are used per topic at the time of evaluation to prevent from putting at disadvantage topics with a large number of words and to follow the evaluation protocol established in that work. If a topic has fewer than four words, it is omitted in the evaluation step.

Topic modeling and named entity recognition (NER) have been used together. The study in [10] presents a new method called CorrLDA2 that was derived from LDA, where word topics contain a distribution over words and over entity topics. The authors of [33] present another version of CorrLDA2, The Entity-Centered Topic Model (ECTM), which models entity topics as a mixture of word topics, ECTM differs mainly from CorrLDA2 in the sampling order of entities and words. In [37], topic detection is not performed on the named entities, instead they are able to detect events using topic clustering and named entities together. In [41], a real time event detection method is presented using named entities and clustering, which splits documents using the entities they contain. This tackles the problem that topic detection and tracking (TDT) systems were designed without regard of noise, spam or real-time performance, specially in big datasets like news on Twitter.

Several topic models have integrated domain knowledge into the topic detection algorithm. The authors of [15] present concept topic model (CTM), which combines LDA with semantic concepts, while the authors of [18] add constrains to LDA where words should or should not be in the same topic according to a domain knowledge. The study in [22] extends LDA by adding word features as supplementary information. In [29], users provide a set of seed words that they consider characteristic of the underlying topics. GK-LDA is introduced in [31] as a general knowledge based model, and MDK-LDA [32] uses prior knowledge from several domains. In [44], a new entity based topic model is presented that additionally incorporates ontologies as the background knowledge into the model. Finally, interactive topic modeling (ITM) [38] allows users to interactively encode feedback into the topic models.

In the aforementioned models, the named entities, entity topics or domain knowledge have been used as additional components in the topic detection model to increase topic quality, but this also adds complexity to an already expensive model, making it even more demanding in time and physical resources. Nonetheless, each entity topic is still of great value by itself, and using a bag of topic models instead of building a golden list of topics can be more useful as more topics are available and thus more granularity is allowed into the topics. Although,

visualizing a lot of topics can be laborious, it is not a major drawback as in most scenarios, topics are used as backend for some applications and for the system using them, the more topics available, the more beneficial it can be, specially if topics are organized on hierarchies, as they enable different levels of granularity and branch filtering. To avoid visualizing a big amount of topics at the same time, a summarized version of the bag of topic models can be created for quick user review.

2.2. Multi-view data

Multiple view data is very common in real world applications. Often, a lot of data is retrieved from different information sources or from different measuring methods, and using a single-view representation cannot comprehensively describe the information of all the items [49]. One solution is to concatenate all multiple views into a single representation and applying a single-view model directly, but this can bring over-fitting problems and ignore the specific statistical properties of each view [49]. In the clustering field, the authors of [8] demonstrated that several multi-view clustering algorithms for text data significantly outperform those based on a single-view.

2.3. Multi-view clustering

Among the multi-view clustering algorithms the ones based on spectral-clustering, have shown promising results. The work in [52] proposed Multi-Graph Fusion for multi-view Spectral Clustering (GFSC), a novel method for integrating graph learning, graph fusion and spectral clustering into a single model, which are mutually optimized on an iterative strategy. GFSC results shows improvements in different performance metrics with widely used datasets.

Some co-training [24] and co-regularized [25] methods for multi-view spectral clustering search for a graph that is consistent across all the views. These methods use the eigenvectors acquired from one view to updated the graph of other view. But one problem of these models is that they tend to have a high variance error. For this reason, several algorithms [28, 39, 50] have been proposed to perform graph learning and improve the quality of the obtained graph making use of a property of the data known as self-expressiveness, which states that each data sample can be expressed as a linear combinations of other data samples. This property has allowed the discovery of low-dimensional manifolds of high-dimensional data representations [21, 40, 43]. But one problem of the graph obtained using this property, is that is not optimized for clustering, for this reason, GFSC obtains a consensus graph from all

views employing the self-expressiveness property, and at the same time, this graph looks to have k-connected components, accomplish by minimizing the sum of the k lowest eigenvalues of the graph Laplacian matrix.

2.4. Multi-view learning on LDA

Multi-view learning has been used with LDA for multimodal categorization and word understanding for robots [20]. Robots can observe an object from different viewpoints, data from audio and haptic sensors as available as well. Using a simile with a bag of words model, where a document, word and topic correspond to an image (scene), image feature and a category, respectively, the robot is able to differentiate and label some objects [14].
There are techniques that use a bag of multimodal LDA models at the same time to improve categorization. Each model pays special attention to a certain modality, making it possible to infer unobserved properties of objects thanks to the connection between the categories in each model [27]. This shows that topic detection can be used in different fields, making it paramount to keep improving topics quality and multi-view learning is a way to do this.

3. Experimental setup

3.1. Datasets

Datasets like academics documents in an university repository have a clear division of documents based on manually created topics, or documents about life forms can be organized according to the organisms taxonomy. In datasets like this, where relations between documents are common and subsets of documents can have several similarities, topics detected automatically should be able to express meaningful themes to users that will help them gain more insight about the information contained in the dataset and complement existing topics. For this reason, academic documents and organisms pages present on Wikipedia are the datasets selected for the experiments in this work.

3.1.1. Wikipedia organisms category

This dataset consists of all the pages of some subcategories of the Wikipedia organisms category[1], and all the pages of the subcategories of the Organisms subcategories, and so forth in a recursive manner.
The traversal of the Wikipedia Organisms category can expand many levels on the spanning tree, and after a certain level, a really big number of nodes have been expanded. On the other hand, the categories on Wikipedia are not acyclic, therefore, a history of the visited nodes is necessary to avoid loops on the traversal. Expanding recursively a single Organisms subcategory can take several hours or days, but as pages can belong to many categories, the traversal on a single category can contain many pages from its sibling categories as Wikipedia protocol for managing categories is not rigorous; therefore, to increase the download speed, only certain subcategories from the Organisms subcategories were taken into account. Appendix I shows the link of each of the categories downloaded recursively. This dataset is available on Google Drive[2].

[1] https://en.wikipedia.org/wiki/Category:Organisms
[2] https://drive.google.com/file/d/1JeazLcV13f9x4WQakTfxfqiLGkahsoue/view?usp=sharing

Data cleansing

All the pages were downloaded as HTML files, each one under a directory of the category they belong to, and as some pages belong to several categories at the same time, removal of repeated files was necessary.

A cleansing step was necessary to remove the HTML format and convert the pages to plain text. This step consisted of applying regex (regular expression) filters carried out using the Python library for Regular expressions (re[3] module). The main regex filters are the following:

1. $< \neg(< \cup >)^+ >$
   Used to delete html labels such as `<href>, <p>, etc.`
2. $\{\{(\Sigma)^*(style \cup cite)(\Sigma)^*\}\}$
   Avoids some labels proper from Wikipedia such as citations.
3. $==\ See\ also(\Sigma)^+$
   Remove all the content below the *See also* section.
4. $(\ [\ )^+ \cup (\ ]\ )^+ \cup (\ |\ )^+ ...$
   Remove some undesired character.
5. $(< SPACE >)^+$ Replace more than one characters.

Where Σ refers to the alphabet corresponding to all UTF-8 valid characters as well as the token `<SPACE>`. These filters are applied in the order given in the above list.
Figure 3.1 shows an example of a fragment from one of the original texts and the output text after applying the filters. Even though these filters cover most HTML tags, cleansing is not perfect as some complex HTML tags still remain, for instance, the removal of nested tables markup is not ideal, nevertheless, this is a good behavior since we want to test the pipeline with imperfect data as it is common in real life situations.

Exploratory data analysis

Table 3.1 shows the number of pages per category, and the corresponding histogram is shown in figure I.1. There are a total of 477,181 pages with a size on disk of 2.48 GB. After removing repeated files that were in many categories at the same time, only 296,042 files were left with a size on disk of 954 MB. And after cleaning the files, this size was reduced to 577 MB.
As we can see, the insects by year of formal description is the category with the highest number of related pages, and categories that describe organisms in a certain century are the ones with the majority of pages, specially the ones related to animals.

[3] https://docs.python.org/3/library/re.html

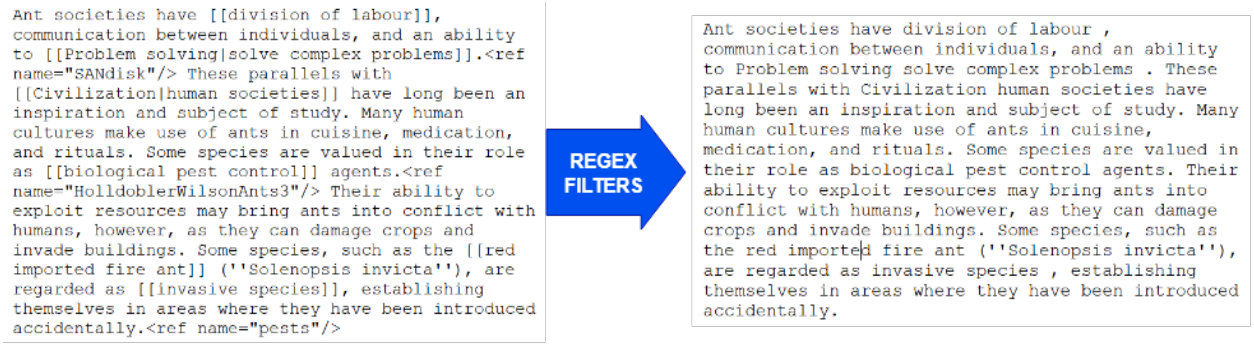

Figure 3.1.: Example of preprocessing using regular expressions to a fragment from the Ant page from Wikipedia.

3.1.2. Universidad Nacional de Colombia thesis abstracts

This dataset consist of downloading all the thesis IDs located at the institutional repository of the Universidad Nacional de Colombia[4], and then, download each thesis metadata as JSON format using its ID. The thesis title and abstract are extracted from the JSON file. Most of these theses are in Spanish, so, it is the language selected for this dataset, and as many documents have several abstracts in different languages, just the first one is selected since normally it is the abstract in Spanish. In total, 16317 abstracts were downloaded with a size on disk of 29.6 MB. This dataset is available on Google Drive[5].

3.1.3. Universidad Nacional de Colombia thesis dataset

This dataset consist of downloading all the thesis IDs located at the institutional repository of the Universidad Nacional de Colombia, and then, downloading each document as PDF using its ID. Some documents are in Microsoft Word format, or the theses are split up into several documents, thus, just the first PDF file was downloaded per thesis.
In total 16260 theses were downloaded, and as these documents are in PDF, a toolkit was need to extract the plain text from them, Apache Tika[6] port on python (tika-python[7]) was the tool selected.
The documents as PDF have an space on disk of 66.4 GB, and after text extraction, they have a size on disk of 3.7 GB.
Most of the thesis documents are written in spanish, therefore, this is the language selected for this dataset. This dataset is available on Google Drive[8]

[4] https://repositorio.unal.edu.co/
[5] https://drive.google.com/file/d/1Jj1bHsEonWDnPF7RNP6S312GzsFFqBn9/view?usp=sharing
[6] http://tika.apache.org/
[7] https://pypi.org/project/tika/
[8] https://drive.google.com/file/d/1xIHO2jKijttT8zriJ_2EhW5Hfmw1lqpE/view?usp=sharing

Category	Pages	Category	Pages
Afrosoricida	85	Monotremes	35
Animals described in the 18th century	1844	Mammals by year of formal description	6229
Animals described in the 19th century	74878	Molluscs by year of formal description	10403
Animals described in the 20th century	102893	Amphibians by year of formal description	6362
Animals described in the 21st century	20441	Nematodes by year of formal description	5
Archea	283	Multituberculates	127
Bacteria by classification	11351	Opossums	134
Bacteria by year of formal description	6144	Pangolins	22
Bats	4778	Plant genera	10639
Birds by year of formal description	11240	Spiders by year of formal description	5148
Carnivorans	12269	Plants by year of formal description	21182
Cingulates	66	Primate families	31
Colugos	8	Protista	91
Crustaceans by year of formal description	1735	Reptiles by year of formal description	5993
Dasyuromorphs	133	Ptolemaiidans	5
Diprotodonts	372	Rodents	5267
Elephant shrews	23	Shrew opossums	11
Euharamiyids	7	Sirenians	57
Fish by year of formal description	47	Soricomorphs	537
Plant orders	245	Fungi by classification	12376
Fungi by year of formal description	5745	Sponges by year of formal description	95
Insects by year of formal description	130730	Starfish by year of formal description	22
Hyraxes	23	Treeshrews	31
Viruses	7039		

Table 3.1.: Pages downloaded per category.

3.2. Text preprocessing

This component in the methodology has the purpose of preparing the data before topic detection, having the POS tagging and named entities recognition steps as the principal components for detecting relevant content words.

3.2.1. Language detection

Current topic detection methods are independent of language, they are based on a bag of words an co-occurrences, but if there are different languages on a dataset, it is expected that the topics created contain just a language at a time (with the premise of just one language per document), but words in different languages can be written the same, so words on topics from multi-language datasets can have several meanings, creating confusing topics. Also, if a trained model is used to process the data, the right language has to be chosen to get accurate results. A simple approach to handle this is to use a language detection model to filter the documents in the desired language, in this case, we used Google's language detection library [23] port on Python langdetect[9], which has 99% precision for 53 languages using a Naive Bayesian filter.

After applying a Spanish filter to the abstract and thesis datasets, 16042 and 15453 documents were left respectively.

3.2.2. Named entities and POS tags

A corpus of documents contains a myriad of words, and even bigger number of different combinations of topics can be created; though some techniques limits the number or words used, e.g., the top N words of an average TF-IDF (term frequency – inverse document frequency) score, but with current ease of using different NLP models, additional preprocessing can be performed on the data. In this thesis, we used spaCy [47] to extract POS tags and named entities selected by the user according to what they considered important in the dataset. Then, a sub-dataset containing the same documents is created for each tag or entity, but documents just have the words belonging to that tag or entity.

spaCy's English model supports different types of Named entities and POS tags, and for the Organisms dataset, not all of them were considered important and some were discarded, others were merged as just one entity. The used entities are described in table 3.2, a thorough documentation of spaCy can be found at its web page[10]. For ease of use, from now on, entities will be referred to as POS tags, as well as named entities.

[9] https://pypi.org/project/langdetect/
[10] https://spacy.io/api/annotation#named-entities

The Spanish model on spaCy has fewer granularity on the entities and POS tags, hence, just three entities and three tags were used on the abstract and thesis datasets; Table 3.3 describe the entity types used.

After the POS tagging and named entity recognition step, the words are passed through a process of lemmatization, so different conjugation of words (specially verbs and adjectives) are considered the same, also, a filter of stop-words is applied just in case the model mislabels stop-words as some entity or tag.

Type	Description
PERSON	People, including fictional.
ORG	Merge of NORP and ORG: Nationalities, religious or political groups, companies, agencies, institutions, etc.
LOC	Merge of GPE, LOC and FAC: Countries, cities, states, non-GPE locations, mountain ranges, bodies of water, buildings, airports, highways, bridges, etc.
OBJ	Merge of PRODUCT, WORK_OF_ART and LAW: Objects (Not services.), vehicles, foods, titles of books, songs, named documents made into laws, etc.
EVENT	Named hurricanes, battles, wars, sports events, etc.
NOUN	Nouns on singular, mass or plural
VERB	Verbs on base form, past tense, gerund or present participle, non-3rd person singular present, 3rd person singular present, past participle or modal auxiliary.
ADJ	Adjectives

Table 3.2.: Entities and tags used on the Organisms dataset.

Type	Description
PER	Named person or family.
ORG	Name of politically or geographically defined location (cities, provinces, countries, international regions, bodies of water, mountains).
LOC	Named corporate, governmental, or other organizational entity.
NOUN	Nouns
VERB	Verbs
ADJ	Adjectives

Table 3.3.: Entities and tags used on the abstract and thesis datasets.

3.2.3. TF-IDF score

After creating a sub-dataset for each entity, the vocabulary size has decreased considerably on each sub-dataset, but a big number of words still remain and a huge amount of topics would be created if all the vocabulary is used, in addition to increasing the necessary computational resources, thus, a way to control the vocabulary size is necessary. On each sub-dataset, words are sorted based on their average TF-IDF score, and at the time of topic detection, only the top N words are selected.

3.3. Topic detection

Based on the results presented in HLTA with the coherence and compactness scores, and a proof of concept with open source implementations of nCRP, LDA and the author's implantation of HLTA, this last method was the only one able to create more than 400 topics with 30 GB of RAM memory on the organisms dataset with a vocabulary size of 1000 words, accordingly, HLTA is the topic modeling method for the experiments on this work.

Considering that each datasets is split up into smaller sub-datasets counting just a type of entity, these sub-datasets can have different importance according to the vocabulary size N selected for each one of them at the time of topic detection. This parameter N affects the number of topics detected, consequently, the number of topic trees created and their depth. On each sub-dataset, the N words of the vocabulary are selected based on their TF-IDF score in that sub-dataset.

The vocabulary size for topic detection on each entity sub-datasets are described in tables 3.4, 3.5 and 3.6. The concatenation of words (n-grams) is no longer necessary as the entity recognition already returns multi-word strings.

After running HLTA on each sub-dataset, to evaluate the detected topics, in addition to the coherence and compactness scores mention on chapter 2, two new metrics are proposed in this work to do not only perform an intra-topic evaluation, but to perform an inter-topic evaluation as well. The first proposed metric is inter-topic coherence, described on equations 3.1 and 3.2. Equation 3.1 returns a metric of how much words $w^{(a)}$ from topic a tend to co-occur with words $w^{(b)}$ from topic b. On the other hand, equation 3.2 performs this equation to each pair of topics within the same tree level of a topic model $\mathcal{L}$. md refers to the maximum depth of the topic trees present on $\mathcal{L}$, while $\mathcal{L}^k$ represents the topics on level k. The reason to only evaluate topics within the same level, it to avoid the comparison between parents and children and the comparison between general topics with much more specific topics. The idea of this metric, is that the more specific a topic is, the harder it will be for other topics to have words with a high probability of co-occurrence between the two topics.

So, if topics are well separated, this metric should be small.

The metric parent-child coherence on equation 3.3 also use the inter-topic coherence formula 3.1, but only between parents and children. For each parent on $\mathcal{L}$, its inter-coherence with each one of its children is averaged. The idea of this metric is that if parents and children are well related, the words between them should co-occur with high probability, so unlike equation 3.2, in this case the higher the score, the better. In this equation, P is the set of all parents on $\mathcal{L}$, and $C^{(i)}$ refers to the children of topic i.

$$Inter\_Coherence\,(topic\,a,\,topic\,b) = \frac{1}{|w^{(a)}| * |w^{(b)}|} \sum_{i=1}^{w^{(a)}} \sum_{j=1}^{w^{(b)}} log \frac{D(w_i^{(a)}, w_j^{(n)}) + 1}{D(w_i^{(a)} \cup w_j^{(b)})} \tag{3.1}$$

$$Inter\_Coherence\,(\mathcal{L}) = \frac{1}{md} \sum_{k}^{md} \frac{2}{|\mathcal{L}^k|\,(|\mathcal{L}^k| - 1)} \sum_{i}^{\mathcal{L}^k} \sum_{j=i+1}^{\mathcal{L}^k} Inter\_Coherence\,(topic\,i,\,topic\,j) \tag{3.2}$$

$$Parent - Child Coherence\,(\mathcal{L}) = \frac{1}{|P|} \sum_{i=1}^{P} \frac{1}{|C^{(i)}|} \sum_{j=1}^{C^{(i)}} Inter\_Coherence\,(topic\,i,\,topic\,j) \tag{3.3}$$

The evaluation of the detected topics are listed in tables 3.7, 3.8 and 3.9.

Sub-dataset	Vocabulary size
PERSON	600
ORG	800
LOC	2000
OBJ	800
EVENT	1400
NOUN	3000
VERB	1000
ADJ	400
Total	**10000**

Table 3.4.: Vocabulary size used for each organism entity sub-dataset.

Sub-dataset	Vocabulary size
PER	300
ORG	1000
LOC	1000
NOUN	2000
VERB	1000
ADJ	400
Total	**5700**

Table 3.5.: Vocabulary size used for each thesis entity sub-datasets.

Sub-dataset	Vocabulary size
PER	400
ORG	500
LOC	500
NOUN	700
VERB	600
ADJ	300
Total	**3000**

Table 3.6.: Vocabulary size used for each abstract entity sub-datasets.

Sub-dataset	Topics	Topic trees max depth	Coherence	Compactness	Inter-topic coherence	Parent-child coherence
PERSON	203	3	-23.4174	0.3226	-6.8069	-5.1120
ORG	254	3	-26.8115	0.1882	-6.4149	-5.3205
LOC	671	4	-23.2675	0.3393	-5.9399	-4.6048
OBJ	254	3	-25.5791	0.1489	-5.7507	-5.0587
EVENT	459	3	-8.9977	0.0676	-2.7468	-2.3034
NOUN	1062	4	-19.0706	0.1268	-4.2713	-3.6120
VERB	342	3	-17.2069	0.1462	-3.5995	-3.1549
ADJ	164	3	-13.3515	0.2185	-3.6210	-2.5629

Table 3.7.: Organism topics evaluation with a total of **3,409** topics.

Appendix A shows an example of some of the most interesting topic trees per entity obtained from the datasets mentioned above, revealing interesting results, e.g., in the location topics A.6, we can see countries that are geographically close, grouped into the same topics. This is very interesting considering that the topic detection algorithm is not supervised and no geographic knowledge is explicitly introduced. In the people topics A.8, an entire tree

Sub-dataset	Topics	Topic trees max depth	Coherence	Compactness	Inter-topic coherence	Parent-child coherence
PER	115	2	-12.2717	0.4277	-3.8446	-2.3964
ORG	372	3	-17.0408	0.2238	-4.5084	-3.2878
LOC	431	4	-14.231	0.2972	-3.5058	-2.6455
NOUN	1228	5	-8.3957	0.3587	-2.6016	-1.6762
VERB	454	4	-9.5295	0.298	-2.3679	-1.8765
ADJ	289	4	-6.8187	0.2575	-2.6197	-1.4008

Table 3.8.: Thesis topics evaluation with a total **2889** topics.

Sub-dataset	Topics	Topic trees max depth	Coherence	Compactness	Inter-topic coherence	Parent-child coherence
PER	125	2	-8.1211	0.2408	-2.4196	-2.1520
ORG	164	3	-8.3658	0.2088	-2.5182	-2.2405
LOC	168	3	-12.066	0.3126	-3.7495	-2.9980
NOUN	282	3	-15.4832	0.3587	-4.2025	-3.0674
VERB	195	3	-20.6979	0.4382	-4.7419	-4.1922
ADJ	107	2	-16.3447	0.2683	-4.2402	-3.3145

Table 3.9.: Abstract topics evaluation with a total of **1041** topics.

contains just fictional characters, in the majority from The Batman comics; similarly, as in the locations topics, HLTA does not know if a person is fictional or not, and much less if it is from DC comics. The noun topics A.2, evince topics containing just words from the medical field, if all words were used, we might not have topics as specifics as these. The event topics A.5, contain phrases of even 5 words long, doing this with n-grams would be very costly in time and resources, and it would be more probable to get meaningless phrases and omit relevant ones whose concatenations are not frequent in the corpus, besides, if stop-words are filtered before detecting n-grams, it would not be possible to discover multi-word strings such as “The who”, as both words are stop-words, or acronyms like WHO (World Health Organization) would be filtered as well.

With regard to the obtained metrics, there seems to be a trade-off between the compactness and coherence scores, this is very interesting as for the way word2vec was trained, in the end is also telling a co-occurrence score for the words, but with respect to another dataset. Nevertheless, there is a considerable problem with the model trained on Google news, and is that several tokens present on the datasets used on this work are not in the word2vec vocabulary, e.g., the event topics from the organisms dataset are the ones with the worst compactness score, and analyzing the tokens on the event vocabulary, most token are com-

plex n-grams with 3 o more words, e.g. "the_american_revolutionary_war", those n-grams in most scenarios, are not on the word2vec vocabulary, so further work is required to have a reliable compactness score.

Regarding the proposed metrics, there is also a trade-off between the inter-topic coherence and the parent-child coherence. Topic models that performed well on the coherence score, do not perform as good with the inter-topic coherence score, but the results of coherence and parent-child coherence are congruent. This seems to indicate that if the words of a topic have a high co-occurrence probability, there is also a high likelihood that those words will have a high co-occurrence with other topic words.

But still, all these topic trees contain relations have not been identified yet, e.g. linking a people topic with a verb topic can give some insight of what the subjects are doing. In this work, we propose multi-view learning for linking topics.

4. Multi-view topic clustering

A topic has associated some documents with a certain probability, also, a set of words is part of a topic with a probability for each word, and a latent space is available to extract different representations of a topic. Instead of concatenating all representations into a single-view, a multi-view learning approach is used to bring performance improvements. In this work, we propose three topic views:

- Topic documents probability.
- Topic words probability.
- The TF-IDF score of the words present on the topic documents multiplied by the document probability; if a word is present on multiple documents, the highest probability is used.
  The TF-IDF score depends on the entity sub-dataset, but most sub-datasets do not share any word, so, we did not considered splitting this view into a view per entity, it would just add more complexity to the model. In the case that a word is present in different sub-datasets, an average TF-IDF score is calculated.

Consequently, in the end, three topic matrix representations are created, one for each view, and according to [30], such data should be normalized into the range $[-1,\ 1]$ for the task of multi-view clustering.

After the topic detection is complete on each entity sub-dataset, several topic models are now available. Even though these dataset representations might be interesting or just enough for some use case, a model that uses all the topic trees jointly can improve the usefulness of the different entity topics. In this chapter, we describe how multi-view learning can be used to group similar topics from different topic trees.

GFSC is used to group together similar topics, but we do not want to group nodes that are parent or children of each other, thus, only topics within the same node level are clustered, this also means that only topics within the same abstraction level are grouped; therefore, the maximum depth of the topic trees is the number of different clusterings carried out. Table 4.1 shows the number of topics per dataset on each level.

GFSC returns a similarity graph matrix which can be used on a clustering algorithm, in this case, clustering on a projection of the normalized Laplacian. For the dimension of the projection subspace (the number of clusters), a heuristic is used: let n be the number of topics in level i, then, the number of clusters in level i is determined by equation 4.1.

$$min(max(4, \left\lfloor \frac{n}{10} \right\rfloor), n) \tag{4.1}$$

This means that in case every cluster has equal number of elements, each group will have 10 topics. Users may change this heuristic if they want bigger or smaller clusters. Besides, spectral clustering does dot perform very well with many clusters, so, dividing the topics number by 10 might still be a large number.
The strategy to assign labels on the embedding space is discretization [7], users may change it to another strategy like k-means, which is the most popular, but it is also sensitive to random initialization. The code for the experiments carried out in this thesis is available on Github[1], where GFSC was implemented on TensorFlow to facilitate its use with another Python libraries such as NumPy and Scikit-learn, and if necessary, make it possible to run GFSC on GPU.

Appendix B shows the clustering results, and with few clusters, they tend to have similar number of topics, but as the number of cluster increases, the "rich get richer" behavior starts to be notorious (usually on levels 1 and 2). As an example, on image B.3, on level 1 there are 2542 topics and 248 clusters with a cluster containing around 100 topics, much more than other clusters, but this an expected behavior as spectral clustering is better suited for few clusters. Besides, topics in lower levels are more specific than the ones in upper levels, making it harder to detect similarities between topics; therefore, clusters with only one topic can occur, overloading other clusters.

Dataset	Trees	Level 1 topics	Level 2 topics	Level 3 topics	Level 4 topics	Level 5 topics	Unique topic words	Repeated topic words
Organisms	112	2542	662	180	25	0	8870	966
Thesis	112	1881	656	242	88	22	5261	411
Abstracts	93	784	211	46	0	0	2787	198

Table 4.1.: Number of topics on each dataset topic tree levels.

Appendices E, F and G show the results of performing spectral clustering with just a single view at a time. In the abstracts E and organisms G datasets, we can see that in all views, most topics tend to be grouped into a single cluster, concatenating all views into a single one (see figures H.1 and H.3) does not improve the results at all, but when using all views

[1] https://github.com/jccaleroe/GFSC-for-entity-topics

(see figures B.1 and B.3), this behavior is drastically reduced.

In the thesis dataset, the topic-word probability view (figure F.1) has a cluster with the majority of topics in it, but the other two views (figures F.2 and F.3) have more equally distributed clusters. Even though one view is not differentiating topics very well, the multi-view algorithm (figure B.2) still captures the similarities found on the other two views. Concatenating all views (see Appendix H) produces very similar results to just using the best single-view, which exhibits the advantage of using multi-view learning.

5. Graph fusion and evaluation

With each topic belonging to a cluster, now it is possible to improve the categorization of words and documents, e.g., in the task of guided search, the search result of a location might not only have documents containing that locations sorted by an TF-IDF score, also, it will be possible to suggest locations belonging to the same topic, and even show people, events or other entities that are included in the same cluster. In order to validate that topics belonging to a cluster are coherent, we propose a method to fusion the different topic trees based on the clusters each topic belongs to, and evaluate this resultant trees as a single topic model. This graph fusion also facilitates the visualization and summarization of the different entity topic models.

5.1. Graph fusion

Using a bag of topic trees and clusters can be useful in a wide range of applications, but visualizing all these hierarchies and clusters can also be overwhelming, therefore, we propose a heuristic to fusion topics belonging to the same cluster, reducing drastically the number of trees and topics. Algorithm 1 summarizes the details of the heuristic considering $C = \{C^1, ..., C^n\}$ as the set of topic clusters and N as the maximum number of words per topic.

Figures C.1, C.2 and C.3 show all the topic trees for the different datasets, the color represents the cluster in which they were grouped in; as we can see, there are numerous connected components, and associating nodes according to their color, connections between different components can be established.

In figures D.1, D.3 and D.5, the results of interpreting each cluster as a single topic are depicted, the number of connected components are drastically reduced, showing that several relations between different topic trees were identified in the clustering step.

In figures D.2, D.4 and D.6, some fusion topics are shown, topics that start with an ID, are the result of fusioning topics, this ID starts with an "U", then the node level and finally an auto-increment number for the nodes in that level. If a fusion topic is a leave, then, the entity topics that were fusioned to create it are shown as its children.

The fusion of topics is also useful for evaluating the model, as one way to evaluate the clustering is to use the compactness and coherence scores on the topic words of each cluster as if the cluster were a single topic, but these scores are sensitive to the number or words used, hence, just as in [45], at the time of evaluation, only four words are selected per topic. or in this case four words per cluster.

Algorithm 1: Topic fusion

Input: C, N
Result: Summarized topics
foreach $C^v \in C$ **do**
  if $|C^v| = 1$ **then continue**;
  Create new fusion topic U_v;
  foreach $t_i \in C^v$ **do**
    Remove the parent of topic t_i;
    foreach *child* $\in$ *children*(t_i) **do**
      Remove *child*'s parent t_i;
      Set U_v as parent of *child*;
    end
    if t_i *is a leave* **then**
      Set U_v as parent of t_i;
    else
      Delete t_i;
    end
    Set as parent of U_v, the most common parent among the nodes in cluster C^v;
  end
  $W^{(U_v)} \leftarrow \emptyset$ (Words on topic U_v);
  repeat
    foreach $t_i \in C^v$ **do**
      $W^{(U_v)} \leftarrow W^{(U_v)} \cup$ Next top word in t_i;
      if $|W^{(U_v)}| = N$ **then break**;
    end
  until $|W^{(U_v)}| = N$;
end

5.2. Evaluation

Tables 5.1 and 5.2 shows the coherence and evaluation scores for the different datasets. First, all datasets were tested with HLTA, which could only finished with the abstracts dataset due

to memory limit (32 GB of memory where available for each experiment). HLTA obtained the best compactness score for the abstracts with less than half a unit of margin, but the coherence score is more than 4 units below the score of the proposed topic fusion strategy. It must be considered that if a word or phrase is not present in the word2vec model, it will be omitted in the compactness score, and the more words a phrase has, the more unlikely that it will be present in the word2vec model. Long phrases are common in certain types of entities, especially the ones that contain proper names, like people, organizations, locations and events. Thus, the compactness score might not be telling the real value, unless it is guaranteed that all topic words and phases occur in the word2vec model, this might require a meticulous word2vec fine tuning on each dataset that is beyond the scope of this work.

For the entity topic models, the average coherence and compactness scores were calculated using all the entity topics at the same time, resulting in a single score for each dataset, obtaining better results than the proposed fusion of topics and than running HLTA on the whole dataset.

	Abstracts-3k	Theses-5.7k	Organisms-10k
HLTA	-18.9160	Memory limit	Memory limit
Bag of entity topics	-13.9122	-10.8538	-19.6033
Topic fusion	-14.3344	-12.6907	-20.5378

Table 5.1.: Average coherence scores for the different datasets.

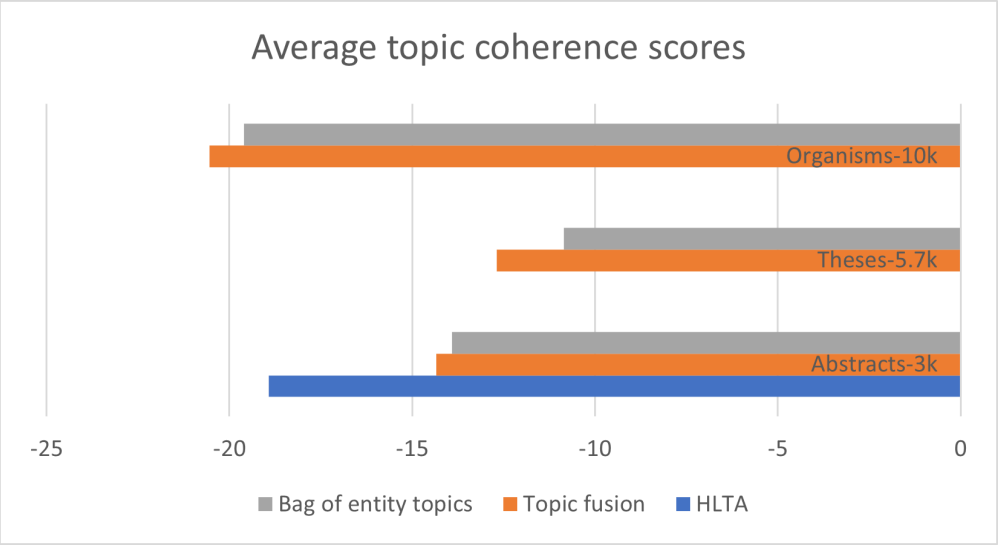

Figure 5.1.: Coherence scores for the different datasets.

The scores for the fusion topics were lower than taking into account all the entity topics, but this is an expected result as the fusion topics now try to model the same patterns the

	Abstracts-3k	Theses-5.7k	Organisms-10k
HLTA	0.3369	Memory limit	Memory limit
Bag of entity topics	0.2942	0.2792	0.1766
Topic fusion	0.2931	0.2679	0.1673

Table 5.2.: Average compactness scores for the different datasets.

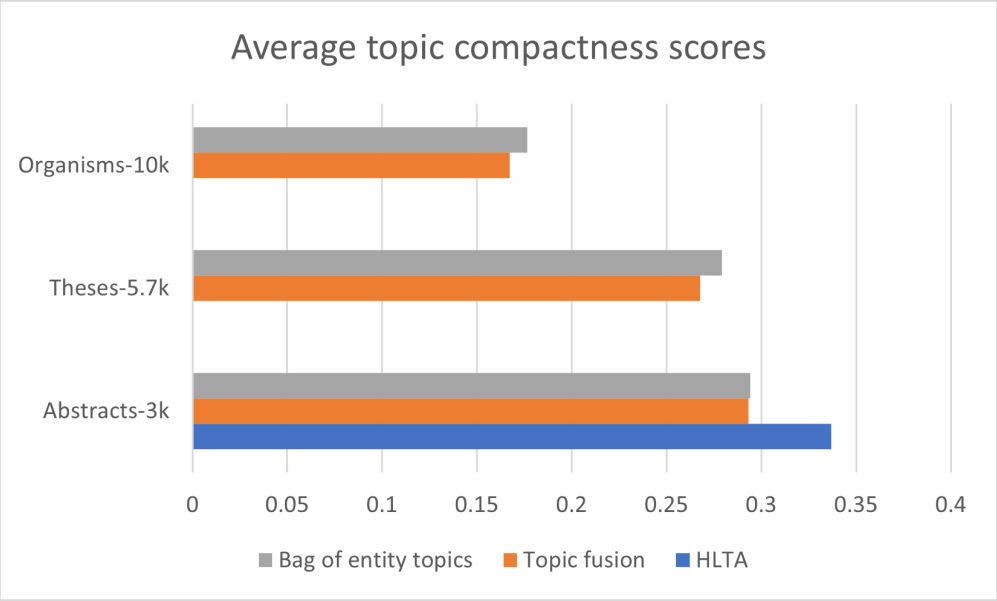

Figure 5.2.: Compactness scores for the different datasets.

entity topics did with a higher number of topics, and yet, for having this big reduction in the number of topics, the differences in the coherence scores are not greater than two units, and in the compactness scores, the differences are not greater than one decimal, which evidence that the clustering algorithm was able to group topics that are related.

On the other hand, with the coherence scores for the abstracts dataset, we can see that the entity topics are around 5 units better than HLTA on the whole dataset. This shows that running hierarchical topic detection on just specific type of words of big datasets decreases the required computational resources and improves topics quality as well.

5.2.1. Inter-topic evaluation

Table 5.3 and image 5.3 show the results of the proposed inter-topic coherence metric, the results for the different experiments on each dataset are very similar, but there is a slightly tendency for the entity topics to be better separated from other topics on the same level. This metric can be further analyzed as this score is originated from the average of the scores

at each tree level, tables 5.4, 5.5 and 5.6 show with more detail how this metrics has obtained.

It is expected for level 1 topics to have the lowest inter-topic coherence as they are the most specific topics and words between them should not co-occur so frequently, by contrast, as higher levels are more general and cover more documents, it is expected for them to have a higher inter-topic coherence than low level topics. In general, we can see this behavior for the different experiments on images 5.4, 5.5 and 5.6. Nevertheless, the opposite behavior is found for the abstracts dataset on the entity topic, and with HLTA, all levels achieved very similar scores, but with the fusion topics, we obtained a more expected outcome.

Regarding the parent-child coherence score, the entity topics always achieved the best results on each dataset, while the fusion topics performed better than HLTA. This proves that by splitting the dataset into multiple bag of words, it is possible to obtain more coherent topics.

	Abstracts-3k	Theses-5.7k	Organisms-10k
HLTA	-4.7010	Memory limit	Memory limit
Bag of entity topics	-4.8197	-2.8474	-5.4412
Topic fusion	-4.6654	-2.8765	-5.3309

Table 5.3.: Average Inter-topic coherence scores for the different datasets.

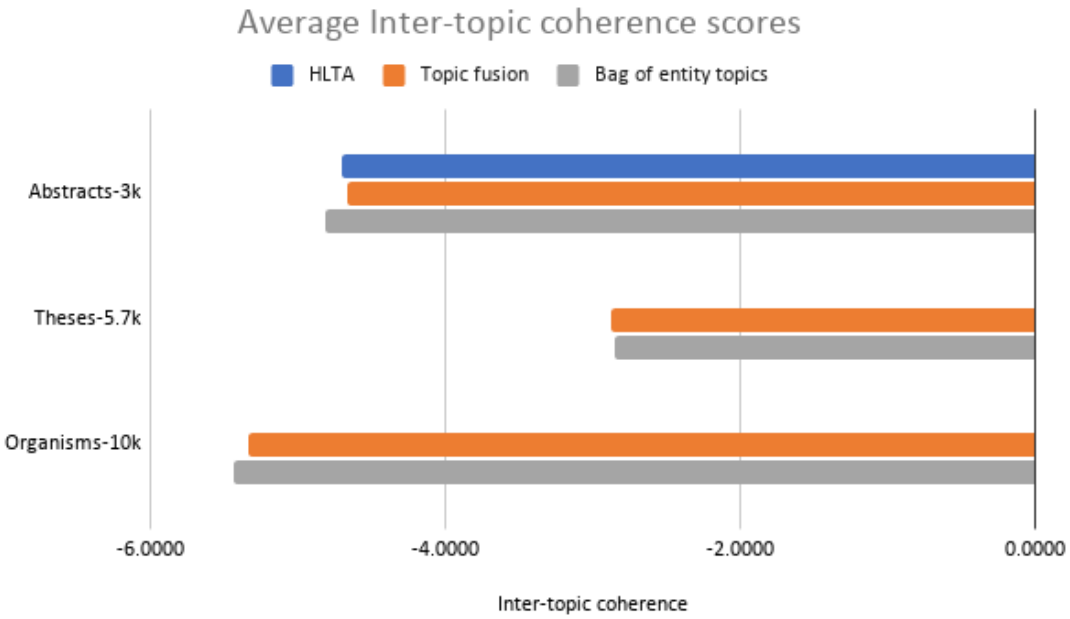

Figure 5.3.: Inter-topic coherence scores for the different datasets.

	Level 1	Level 2	Level 3	Level 4
Abstracts-3k	-4.7386	-4.7151	-4.6494	-4.7009
Theses-5.7k	n/a	n/a	n/a	n/a
Organisms-10k	n/a	n/a	n/a	n/a

Table 5.4.: Inter-topic coherence per tree level for the HLTA topics

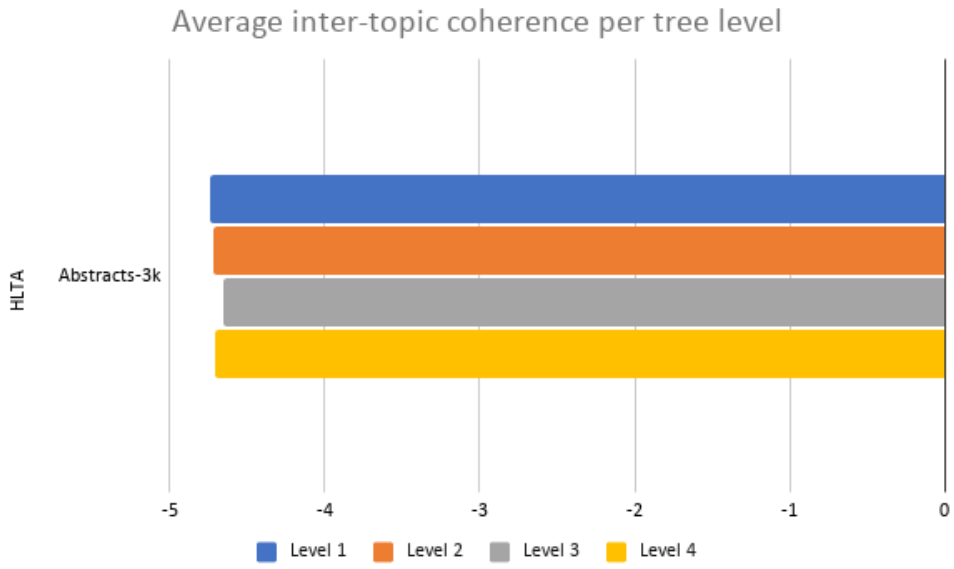

Figure 5.4.: Inter-topic coherence per tree level for the HLTA topics.

	Level 1	Level 2	Level 3	Level 4	Level 5
Abstracts-3k	-4.7647	-4.8143	-4.8802	n/a	n/a
Theses-5.7k	-3.4986	-3.2885	-2.8990	-2.4584	-2.0923
Organisms-10k	-5.8512	-5.7747	-5.6196	-4.5194	n/a

Table 5.5.: Inter-topic coherence per tree level for the entity topics.

	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
Abstracts-3k	-4.7637	-4.7907	-4.6600	-4.4473	n/a	n/a
Theses-5.7k	-3.4992	-3.4088	-3.1092	-2.6277	-2.4599	-2.1544
Organisms-10k	-5.8534	-5.5535	-5.6021	-5.3249	-4.3205	n/a

Table 5.6.: Inter-topic coherence per tree level for the fusion topics.

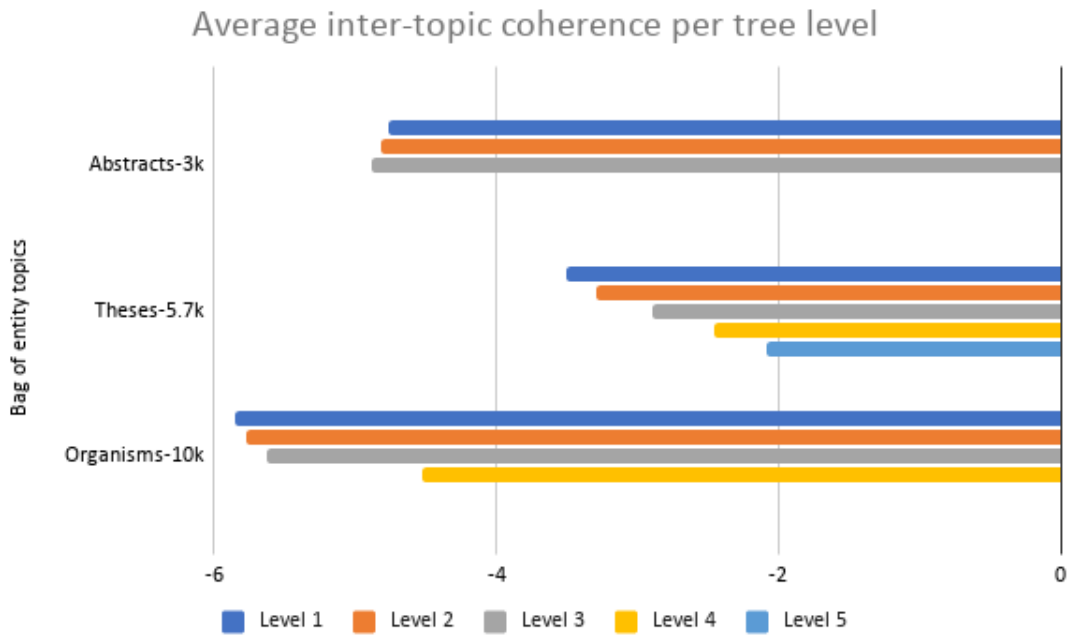

Figure 5.5.: Inter-topic coherence per tree level for the entity topics.

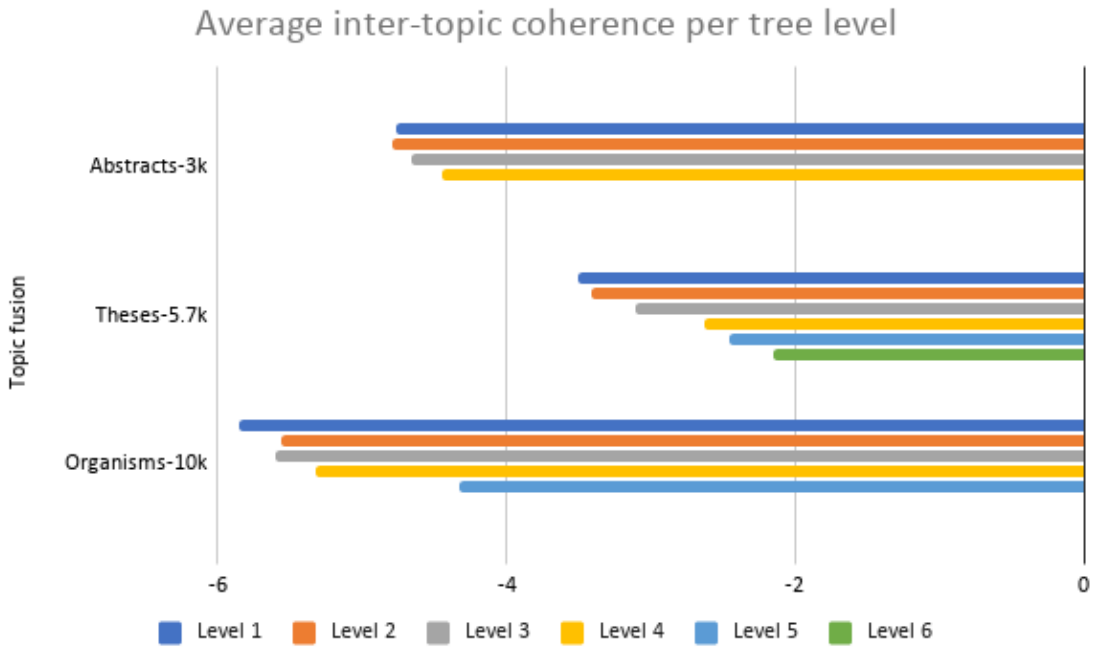

Figure 5.6.: Inter-topic coherence per tree level for the fusion topics.

	Abstracts-3k	Theses-5.7k	Organisms-10k
HLTA	-3.8971	Memory limit	Memory limit
Bag of entity topics	-3.1169	-1.9757	-3.9262
Topic fusion	-3.5564	-2.3912	-4.3167

Table 5.7.: Average parent-child coherence scores for the different datasets.

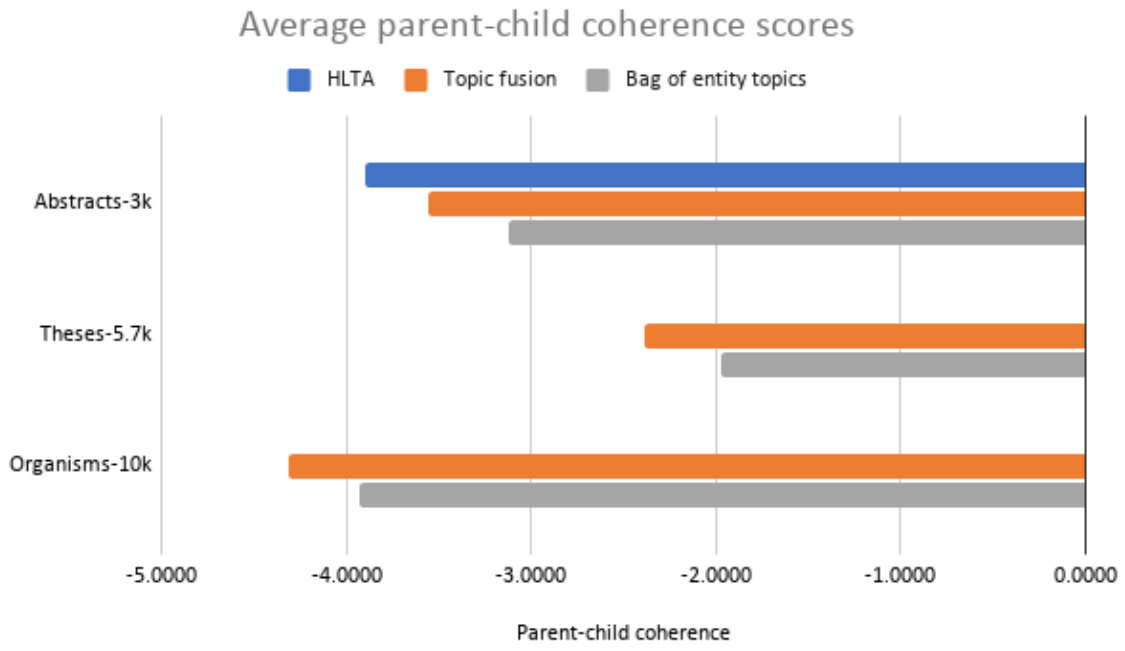

Figure 5.7.: Parent-child coherence scores for the different datasets.

5.3. Empirical evaluations

In figure D.6, some topics from the organisms datasets are presented, e.g., topic *U_1_107* talks about plants, and as its children show that it is created from different types of entities; we can see adjectives (e.g., evergreen, warm, native, ornamental, etc.), nouns (e.g., mountain, autumn, seed, germination, sepal, etc.), verbs (e.g., plant, bloom, branch, etc.) and organizations like the University of California Press, which has subjects for environmental and science studies. All these topics are related to plants, and even though they are from different topic trees, the clustering step has been able to relate them.

The same analysis can be done for the fusion topics on the abstract and thesis datasets. The abstracts topics shown in figure D.2 seem to talk about scientific procedures, e.g. topic *U_1_35* has verbs and adjectives in Spanish like isolate, identify, choose, try, bacterial, similar, etc. And nouns about similar concepts, e.g., bacterium, infection, peptide, control, loss,

among others.

On the other hand, on figure D.4, we have topics about economic affairs from the thesis dataset, e.g., topic *U_1_87* has organizations such as Banco de la República (Central Bank of Colombia), BVC (Bolsa de Valores de Colombia, or Colombia Stock Exchange in English), DIAN (Dirección de Impuestos y Aduanas Nacionales, or National Directorate of Taxes and Customs in English), Financial Superintendent of Colombia, etc. The nouns and verbs are related to economic concepts as well, e.g. remuneration, financing, inflation, accounting, loan, patrimony, liquidate, worsen, narrow down, etc.

Based on these experiments, with multi-view clustering was possible to identify similar topics from different topic models, but topics quality can still be improved as many words are mislabeled in the entity recognition step. This can be handled by fine-tuning the POS and NER models on each dataset, but this would require additional annotated examples that are beyond the scope of this work.

5.4. Topic models initial reviews

In order to get additional validation to the experiments carried out in this work, an exploratory survey was performed on 30 people, including undergraduate and graduate students as well as graduates from the Universidad Nacional de Colombia. The survey contains 6 topic trees from the different datasets, two topics were selected from the HLTA model, two from the entity topics, and two from the proposed fusion of topics. The respondents were asked to answer how meaningful the topics were for them on a scale from 1 to 10, where 1 means not meaningful at all, and 10 means very meaningful. Additional questions were made to detect if respondents had previous experience analyzing topic trees generated automatically.

Figure 5.8 and table 5.8 shows the average of the responses, the fusion of topics and the HLTA topics had similar results with a slightly advantage for the topic fusion. Entity topics had a better reception with almost two points ahead of HLTA. These results are coherent with the metrics obtained with the compactness and coherence scores.

Among the comments, people said that with the entity topics they could inferred a concrete context, and with the other topics, they could not identify a main topic. Another point to highlight is that in all the questions, people with previous experience on topic detection, on average had better reviews than the ones without experience, this might be because they are aware that not meaningful topics can occur with automatic topic detection. A summary

of the results can be found at Google docs[1].

	Average	Average with experience	Average without experience
HLTA	4.23	5.2	4.04
Topic fusion	4.98	5.5	4.88
Entity topics	6.08	6.7	5.96

Table 5.8.: Survey results for topic meaningfulness based on a scale from 1 to 10.

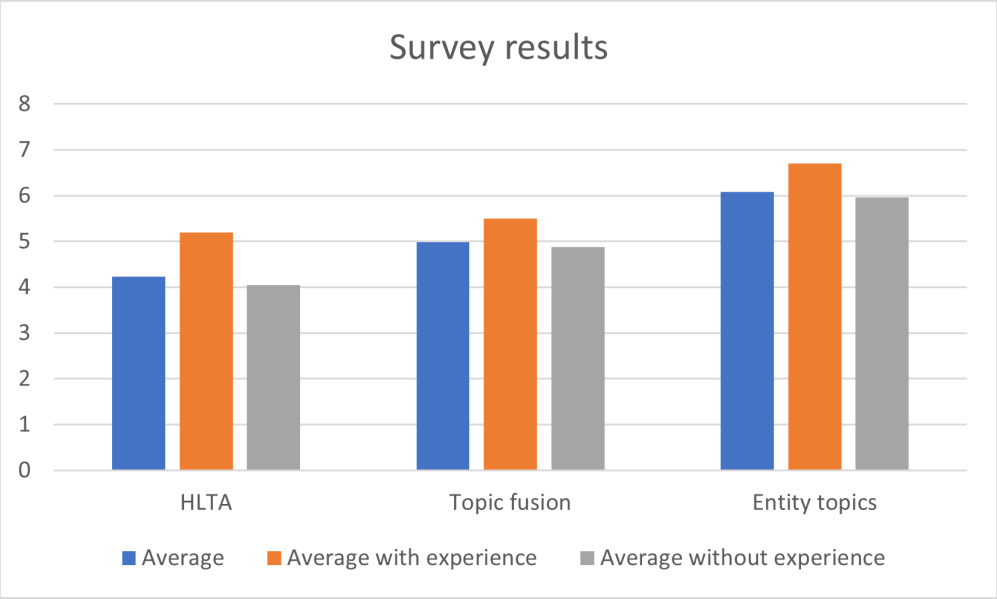

Figure 5.8.: Survey results performed by 30 people where in a scale of 1 to 10, they had to choose how meaningful some topics were. 5 people had previous experience analyzing topic models and 25 did not have any experience.

[1] https://docs.google.com/forms/d/1_gKI5GC_2wcPJqkXwpJds3QwhH58N3gKP--luM1bWcE/viewanalytics

6. Conclusions and recommendations

6.1. Conclusions

In this work, we presented a methodology where we can obtain more meaningful topics by splitting the dataset into more specific bag of words; in this case, named entities and POS tags. Additional connections between the entity topics can be found using clustering. Furthermore, a topic contains some information about words and documents, so, a multi-view representation of topics is possible. Although different topic detection algorithms might have different views from the ones presented here, it is highly probable that a multi-view approach has better results, as happened in this work with the experiments carried out on three different datasets. On top of that, with the fast improvements in transformer neural networks [48] with pre-trained models such as BERT [51] and GPT-2 [53], the quality of POS tags and named entities recognition can improve as well, thus allowing the detection of even more specific and meaningful entity topics.

6.2. Recommendations

The fusion of topics presented in this work has mainly a purpose of summarization and visualization, using this as the main topic model still has some aspects to be considered as the topic space is modified. When merging topics, the words and documents probabilities in them have to be calculated in a way such that topics in the latent space are still coherent. This can be tricky as mentioned in [19], since paying too much attention to the latent space behind the model does not lead to semantically meaningful topics. Using a bag of entity topic models jointly by finding their relations and avoiding the merging of topics avoids this problem.

We presented a methodology for detecting topics on a corpus of documents, but processing a stream of documents, where topics can be modified or added as new data arrives, has yet to be considered, as in most scenarios, data is not static.

In the methodology presented in this work, it is possible to easily change the algorithms on each step for another implementations, for instance, if hierarchical topics are not necessary, a traditional implementation of LDA can be used instead, or if there is a large amount of

topics, a different clustering method that has a usecase for many clusters might be better suited instead of spectral clustering.

In general, automatically created topics consist of a set of words, reading these words can be confusing for users who cannot easily infer a context or concept based on such words, using a technique that can infer a concept for a set of words could make it easier to visualize and analyze topic trees. Few-shot learners such as GPT-3 [54] could perform well on a task like this, allowing for a better visualization of topics by just displaying one or very few words per topic.

A. Appendix: Topic tree examples

- 0.060 genetico molecular relacionar encontrar dirigir comparativo
  - 0.160 relacionar encontrar dirigir comparativo
  - 0.000 genetico molecular
- 0.083 positivo negativo agudo medico adverso secundario quirurgico
  - 0.000 positivo negativo
  - 0.016 medico quirurgico someter llamar
  - 0.125 secundario primario terapeutico celular
  - 0.026 agudo adverso independiente derivar
- 0.154 social publico politico territorial urbano economico cultural
  - 0.086 economico ambiental sostenible natural
  - 0.084 urbano territorial cultural rural
  - 0.000 social politico
  - 0.000 privar publico
- 0.141 nacional local internacional juridico regional quimico normativo
  - 0.000 regional local
  - 0.046 quimico organico fisico biologico
  - 0.116 juridico normativo colombiano experimental
  - 0.000 internacional nacional
- 0.052 electrico magnetico optico solar electronico termico estructural
  - 0.037 electrico magnetico optico solar
  - 0.021 electronico estructural termico morfologico

Figure A.1.: Some adjective topics from the abstracts dataset.

- 0.252 consentimiento patologia medico ansiedad enfermo clinico sensacion
  - 0.236 patologia medico clinico sintoma terapia sexo prevalencia
    - 0.243 sangrar infeccion medicamento lesion cancer farmaco tumor
      - 0.232 infeccion medicamento glandula estre citoquina conjunto hembra
        - 0.127 glandula citoquina hembra buffer
        - 0.301 conjunto convergencia notacion instante
        - 0.000 estre afrontamiento
        - 0.000 medicamento infeccion
      - 0.109 musculo cerebro neurona corteza
        - 0.000 cerebro musculo
        - 0.000 corteza neurona
      - 0.137 sangrar inflamacion tracto apoptosis
        - 0.000 sangrar tracto
        - 0.000 apoptosis inflamacion
      - 0.133 lesion cancer farmaco tumor
        - 0.000 lesion cancer
        - 0.000 tumor farmaco
    - 0.197 patologia clinico terapia prevalencia trastorno complicacion hospitalizacion
      - 0.106 hospitalizacion morbilidad diabetes adherencia
        - 0.000 morbilidad adherencia
        - 0.000 hospitalizacion diabetes
      - 0.201 trastorno complicacion lesionar desenlazar
        - 0.000 trastorno lesionar
        - 0.000 desenlazar complicacion
      - 0.173 patologia cirugia hipertension progresion colesterol envejecimiento
        - 0.000 cirugia patologia
        - 0.096 hipertension progresion colesterol envejecimiento
      - 0.178 clinico terapia prevalencia pacientes hospital criterios
        - 0.174 clinico terapia pacientes hospital
        - 0.000 criterios prevalencia
    - 0.187 sintoma disfuncion severidad puntaje maduracion obesidad pulmon
      - 0.149 sintoma disfuncion severidad puntaje
        - 0.000 severidad puntaje
        - 0.000 disfuncion sintoma
      - 0.177 maduracion obesidad pulmon etileno hormona quimioterapia susceptibilidad
        - 0.000 maduracion etileno
        - 0.000 susceptibilidad conteo
        - 0.104 pulmon quimioterapia hipoxia mejoria
        - 0.000 obesidad hormona

Figure A.2.: Some noun topics from the thesis dataset.

- 0.157 dane cepal departamento_administrativo_nacional_de_estadistica banco_mundial icbf instituto_colombiano_de_bienestar_familiar conpes
  - 0.132 dane cepal departamento_administrativo_nacional_de_estadistica instituto_colombiano_de_bienestar_familiar sdp ideca
    - 0.000 departamento_administrativo_nacional_de_estadistica cepal
    - 0.056 dane instituto_colombiano_de_bienestar_familiar sdp ideca
  - 0.061 banco_mundial icbf conpes unicef seguridad_alimentaria seguridad_alimentaria_y_nutricional educacion
    - 0.000 seguridad_alimentaria_y_nutricional icbf
    - 0.032 unicef educacion nutricional suma
    - 0.000 banco_mundial conpes
    - 0.037 seguridad_alimentaria precio rae director
- 0.033 ecopetrol epm creg energia unidad_de_planeacion_minero_energetica isa iea
  - 0.012 energia gas bolsa ahorro compania mem sec
    - 0.011 gas bolsa sec ctcp
    - 0.040 energia ahorro compania mem
  - 0.016 ecopetrol anh ecopetrol_sa shell electroforesis barrera rpmi
    - 0.352 ecopetrol_sa barrera rpmi prm smith_channel autopista
    - 0.013 ecopetrol anh shell electroforesis
  - 0.032 isa anla isagen sspd ani mineria invias
    - 0.004 isagen cvc construccion urss
    - 0.020 anla mineria invias departamento_administrativo
    - 0.005 ani bbva corantioquia pmi
    - 0.005 isa sspd _sons eta
  - 0.021 epm creg unidad_de_planeacion_minero_energetica iea sina cnd ocde
    - 0.023 epm creg iea ocde
    - 0.034 unidad_de_planeacion_minero_energetica sina cnd autoridad

Figure A.3.: Some organization topics from the thesis dataset.

- 0.166 incubar centrifugar secar mezclar agitar disolver calentar
  - 0.144 remover lavar catalizar proteina activar amplificar aislar
    - 0.234 remover catalizar estabilizar acelerar observandose modular
      - 0.209 acelerar intensificar constituyendose referenciar
      - 0.278 estabilizar observandose modular transcurrir
      - 0.119 catalizar reclutar mostro ahorrar
      - 0.215 remover atribuirse diligenciar nota
    - 0.260 decrecer simular exhibir descomponer conocido graficar
      - 0.000 graficar simular
      - 0.280 decrecer exhibir descomponer conocido
    - 0.061 proteina activar amplificar bloquear
      - 0.000 bloquear activar
      - 0.000 proteina amplificar
    - 0.110 lavar aislar express reverse report codificar significant
      - 0.000 determine significant
      - 0.157 lavar aislar codificar revertir
      - 0.049 report donar partiendo cualificar
      - 0.034 express reverse linear restituir
  - 0.195 secar mezclar agitar disolver calentar enfriar purificar
    - 0.188 secar mezclar agitar disolver purificar precipitar degradar
      - 0.000 solubilizar disolver
      - 0.185 secar mezclar purificar degradar
      - 0.000 precipitar hidrolizar
      - 0.000 verter agitar
    - 0.139 enfriar calentar evaporar depositar obteniendose persuadir
      - 0.130 enfriar depositar obteniendose persuadir
      - 0.000 calentar evaporar

Figure A.4.: Some verb topics from the thesis dataset.

- 0.002 the_great_famine_ the_great_plague_of_london the_great_northern_war the_great_plague_of_vienna the_great_famine the_black_death ireland
  - 0.000 the_falklands_war vietnam_war_vietnam christmas great_war the_world_war_i the_venezuelan_coastal_range the_wild_bird_conservation_act
    - 0.000 vietnam_war_vietnam the_falklands_war the_venezuelan_coastal_range bromeliaceaebromeliaceae
    - 0.392 the_wild_bird_conservation_act edgew world_war_2 super_bowl_50
    - 0.000 great_war the_world_war_i argentina_la_rioja the_trojan_war
    - 0.391 christmas us_fish_and_wildlife_service the_spr the_pleistocene_epoch
  - 0.001 the_great_plague_of_london the_great_northern_war the_black_death the_great_plague_of_vienna third_pandemic the_great_plague cannes_film_festival
    - 0.385 the_sundance_film_festival stal deathdate mantissa_plantarum_2
    - 0.002 cannes_film_festival cannes_film_festival_cannes the_palme_d spike_video_game_awards
    - 0.001 the_great_northern_war the_great_plague_of_vienna the_great_plague great_northern_war
    - 0.001 the_great_plague_of_london the_black_death third_pandemic great_kanto
  - 0.001 the_great_famine_ the_great_famine ireland wwii the_us_civil_war the_english_civil_war world_war_i_and_world_war_ii
    - 0.001 the_great_famine_ the_great_famine the_us_civil_war the_us_fish_and_wildlife_service
    - 0.001 ireland wwii the_english_civil_war world_war_i_and_world_war_ii
  - 0.001 the_seminole_war the_black_hawk_war the_war_of_1812 the_second_seminole_war revolutionary_war the_battle_of_new_orlean battle_of_horseshoe_bend
    - 0.000 the_black_hawk_war battle_of_horseshoe_bend the_creek_war the_revolutionary_war
    - 0.000 black_hawk_war hurricane_ivan place ameghino_1902
    - 0.000 the_seminole_war the_battle_of_new_orlean war_dog great_britain
    - 0.000 the_war_of_1812 the_second_seminole_war revolutionary_war the_texas_revolution

Figure A.5.: Some event topics from the organisms dataset.

- 0.039 colombia bolivia venezuela peru brazil south_america ecuador
  - 0.008 nicaragua uruguay chile sao_paulo rio_de_janeiro santa_catarina parana
    - 0.003 santa_catarina willd spreng c_africa
    - 0.005 uruguay chile sao_paulo havana
    - 0.001 rio_de_janeiro parana espirito_santo __brazil
    - 0.002 nicaragua antilles lesser_antilles the_yucatan_peninsula
  - 0.038 brazil mexico costa_rica central_america paraguay honduras veracruz
    - 0.017 brazil mexico central_america schaus
    - 0.009 puerto_rico bermuda patagonia humboldt
    - 0.009 honduras veracruz the_west_indies haiti
    - 0.007 costa_rica paraguay el_salvador jamaica
  - 0.038 colombia bolivia peru venezuela south_america ecuador argentina
    - 0.012 south_america panama belize nw_brazil
    - 0.023 peru ecuador tobago caribbean
    - 0.014 venezuela argentina trinidad west_indies
    - 0.000 bolivia colombia
  - 0.001 south_georgia falkland_islands gondwana the_aleutian_islands central_park island north_and_south_america
    - 0.005 the_aleutian_islands north_and_south_america the_southern_ocean san_diego_county
    - 0.427 campbell_island hackenheim stenaliini aeridinae
    - 0.000 south_georgia falkland_islands island reunion_island
    - 0.003 gondwana central_park antarctica brooklyn

- 0.022 gabon ethiopia somalia the_central_african_republic central_african_republic rwanda holland

- 0.007 kazakhstan uzbekistan turkmenistan kyrgyzstan tajikistan morocco iran

Figure A.6.: Some location topics from the organisms dataset.

- 0.002 acraea afrotropical_butterflies banksia_ser rebel banksia subg nymphalidae
  - 0.000 roding strombus crc_handbook_of_avian_body_mass cell _bates pterostylis ceanothus
    - 0.390 ceanothus hindi alpine the_black_death
    - 0.000 roding strombus salticidae ptinidae
    - 0.071 _bates erebidae snellen galileo
    - 0.047 cell crc_handbook_of_avian_body_mass pterostylis cambaridae
  - 0.002 acraea afrotropical_butterflies nymphalidae banksia_ser banksia subg le_cerf
    - 0.003 melaleuca woody_capsule subdivisionranks euthyneura
    - 0.003 acraea le_cerf breuning macleay
    - 0.136 subg asilidae _rogenhofer familiaauthority_
    - 0.003 afrotropical_butterflies nymphalidae banksia_ser banksia
  - 0.003 rebel cricetidae gelechiidae le_doux _ortea liebm rubus
    - 0.799 world_checklist_of_selected_plant_families_publisher malvaceae marathi coenagrionidae
    - 0.012 rebel cricetidae gelechiidae rubus
    - 0.445 lecithoceridae nextyear amur critically_endangered
    - 0.166 le_doux _ortea liebm pleurotoma
  - 0.552 gracillariidae dytiscidae_species_list_at_joel_hallan _park aspergillus red_list_of_threatened_specy thamnophilidae iris
    - 0.359 gracillariidae dytiscidae_species_list_at_joel_hallan curculionidae fabaceae
    - 0.371 aspergillus iris familiaauthority streptomyce
    - 0.262 virusgroup charaxes _lane cerithiidae
    - 0.273 _park red_list_of_threatened_specy thamnophilidae mordellidae
- 0.000 strand coptops sepal drillia_ clavus morch mull
  - 0.001 turbonilla_ gastropoda dipteres radix stenoma taxon crambus
    - 0.007 turbonilla_ gastropoda radix _breuning
    - 0.337 geastrum billbergia underworld chaetodon the_care_bears avengers
    - 0.430 dipteres stenoma crambus taxon
    - 0.016 linecolor calotes deelemanreinhold pack sedum aloeides
  - 0.067 cyprinidae the_iucn_red_list_of_threatened_species ellis_ tagetes expand_section_date amalda peristome

Figure A.7.: Some object topics from the organisms dataset.

- 0.004 james bruce_wayne joker james_gordon tom jerry wayne
  - 0.003 tom thomas jerry wayne penguin scott_bradley harley_quinn
    - 0.001 thomas scott_bradley hanna brown
    - 0.000 tom jerry
    - 0.001 wayne damian oswald herbert_druce_druce
    - 0.001 penguin harley_quinn barbara jerome
  - 0.008 james charles henry mary charlie hamilton shortsummary
    - 0.005 henry bob shortsummary russell
    - 0.281 griseb boletaceae eugene_g_munroe_munroe gill
    - 0.009 james hamilton parker victoria
    - 0.007 charles mary charlie phil
  - 0.002 bruce_wayne joker james_gordon bruce robin alfred gotham
    - 0.001 gotham riddler bane hush
    - 0.001 alfred catwoman al_ghul jason
    - 0.002 bruce robin superman scarecrow
    - 0.002 bruce_wayne joker james_gordon dick_grayson
  - 0.001 tweety earl sylvester ralph charaxes robert fisher
    - 0.013 earl charaxes ralph butler
    - 0.014 robert williams garfield turner
    - 0.008 fisher diana rupert weber
    - 0.001 tweety sylvester dorothy robinson

Figure A.8.: Some people topics from the organisms dataset.

B. Appendix: Topic clustering results

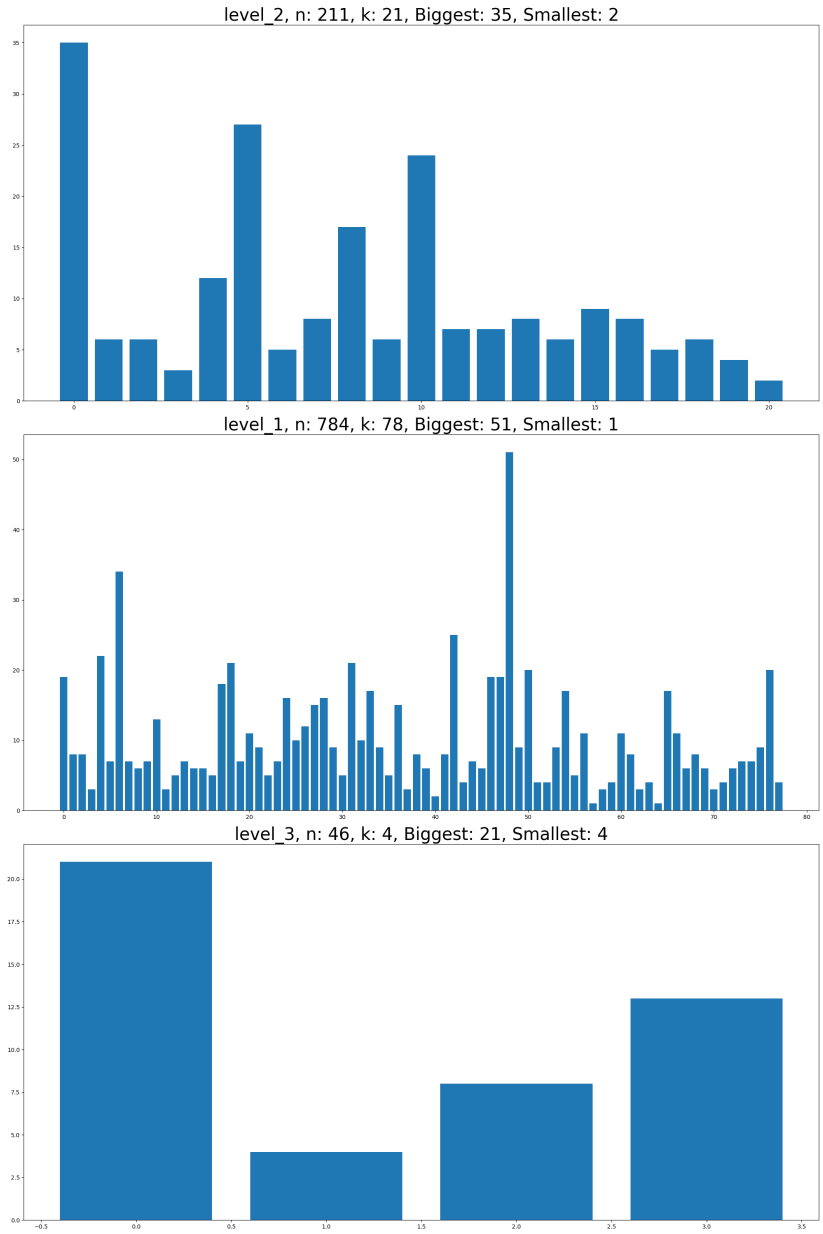

Figure B.1.: Topic clustering per tree level from the abstracts entity topic models.

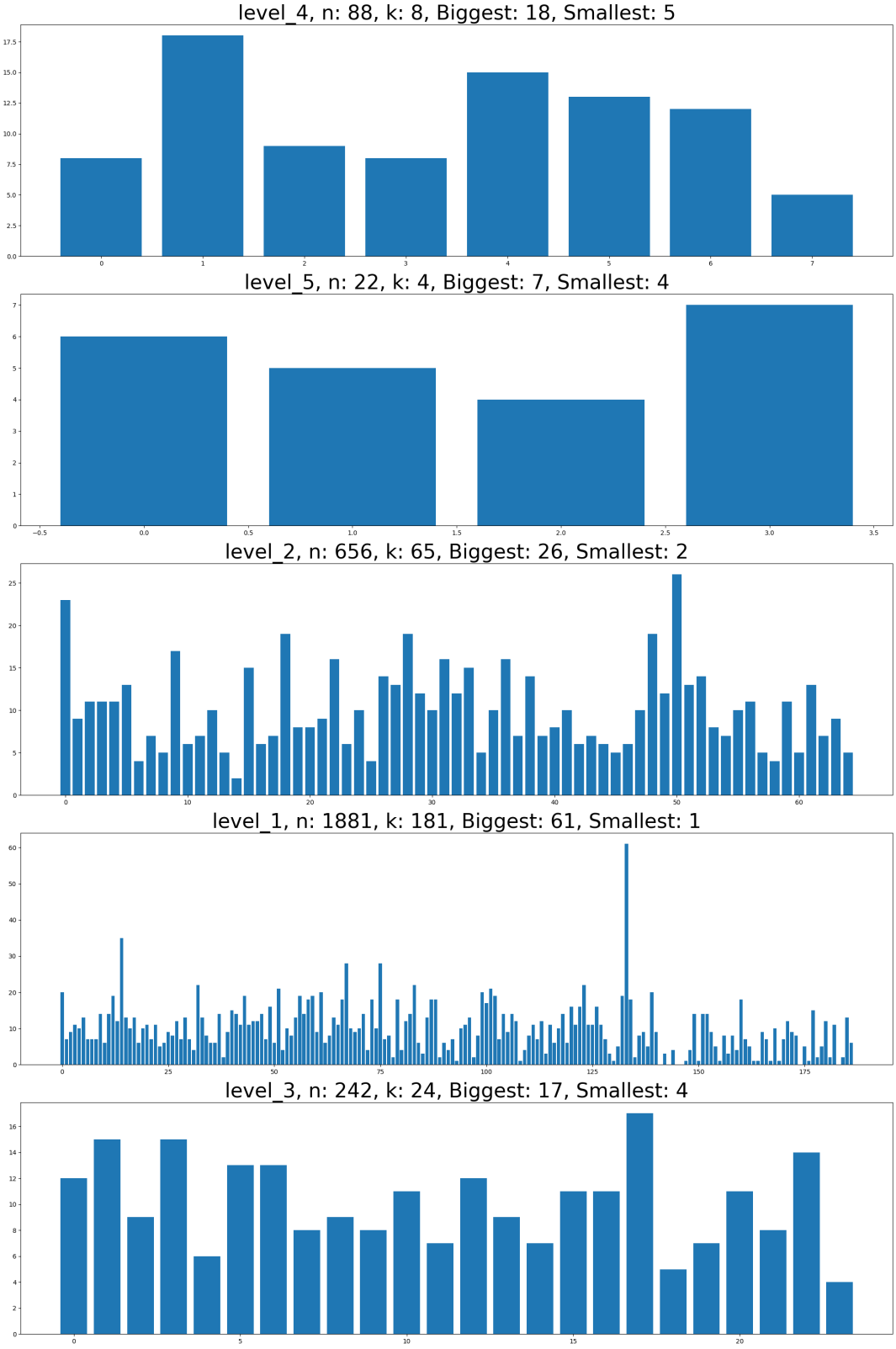

Figure B.2.: Topic clustering per tree level from the thesis entity topic models.

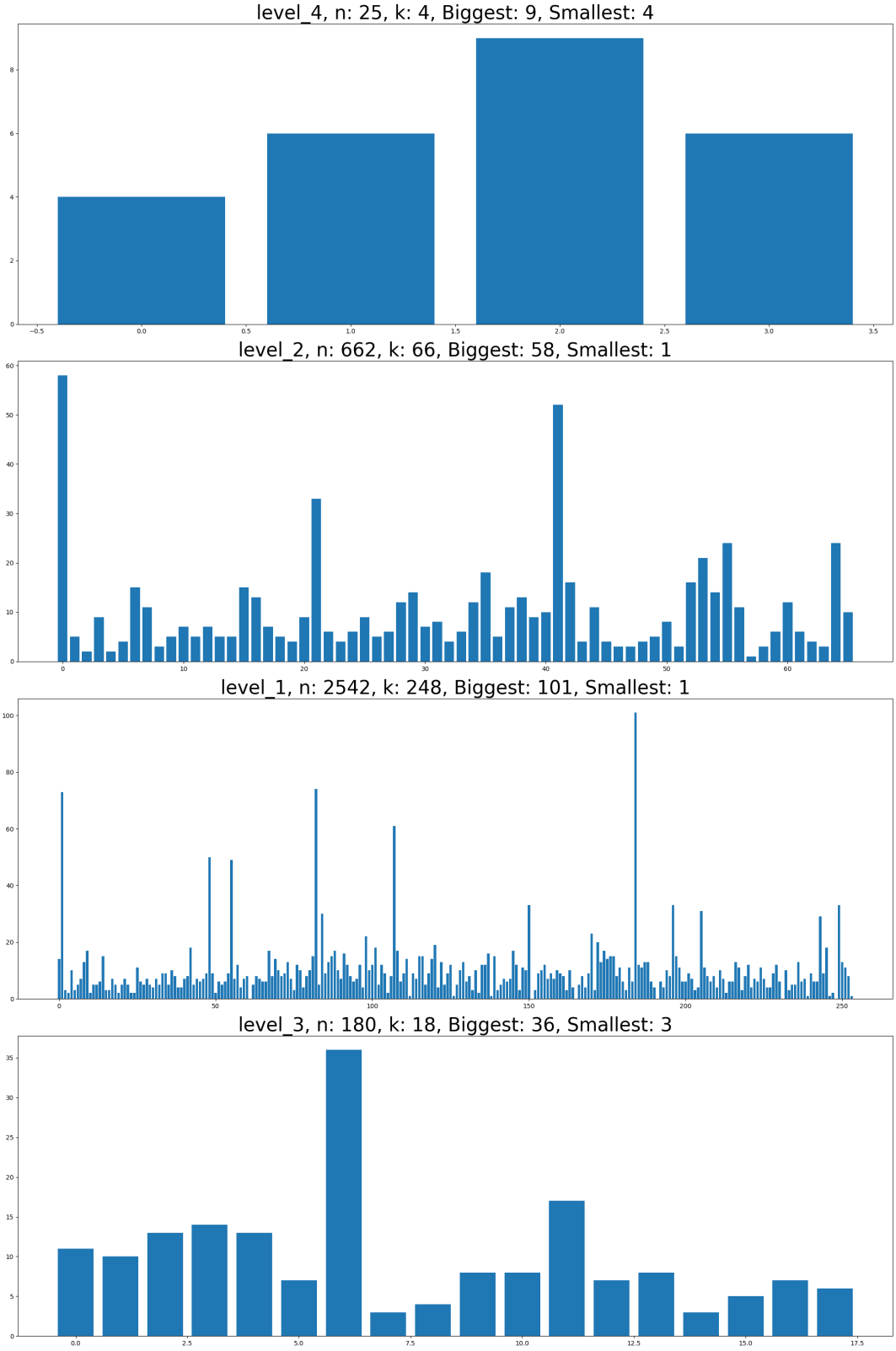

Figure B.3.: Topic clustering per tree level from the organisms entity topic models.

C. Appendix: Topic hierarchies

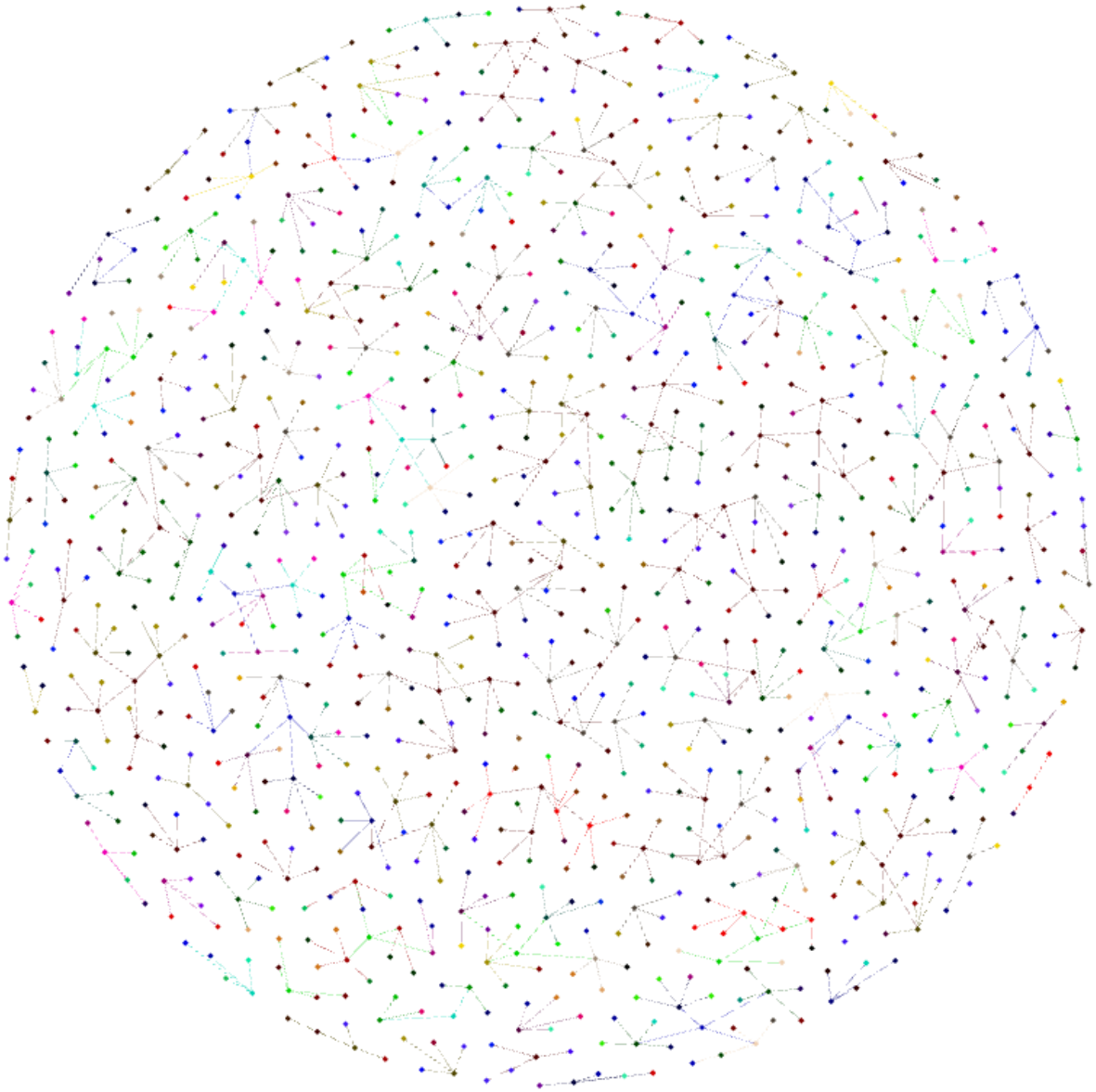

Figure C.1.: Topic hierarchies of the abstracts entities, the color represents the cluster topics belong to.

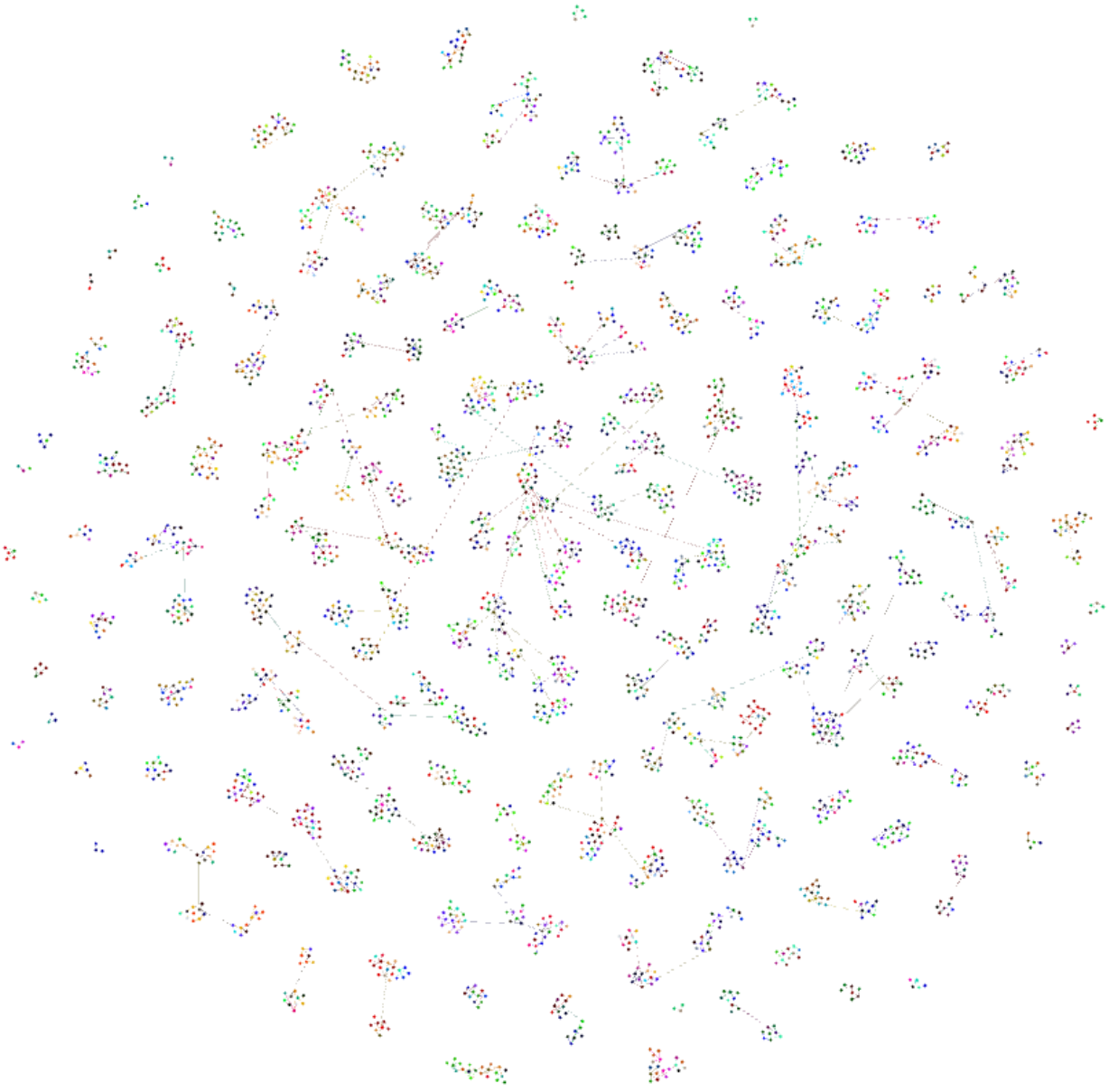

Figure C.2.: Topic hierarchies of the theses entities, the color represents the cluster topics belong to.

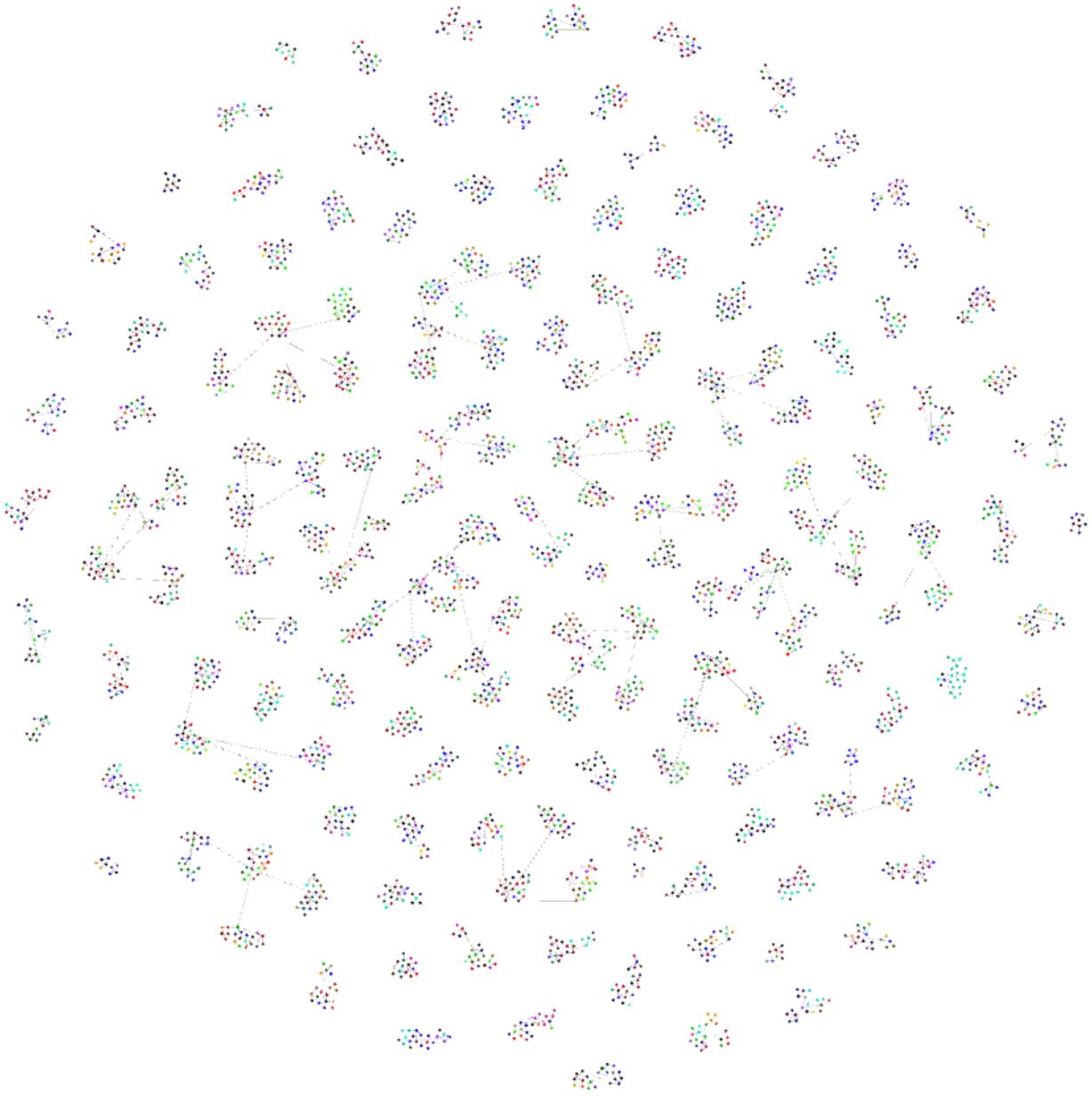

Figure C.3.: Topic hierarchies of the Organisms entities, the color represents the cluster topics belong to.

D. Appendix: Topic fusion

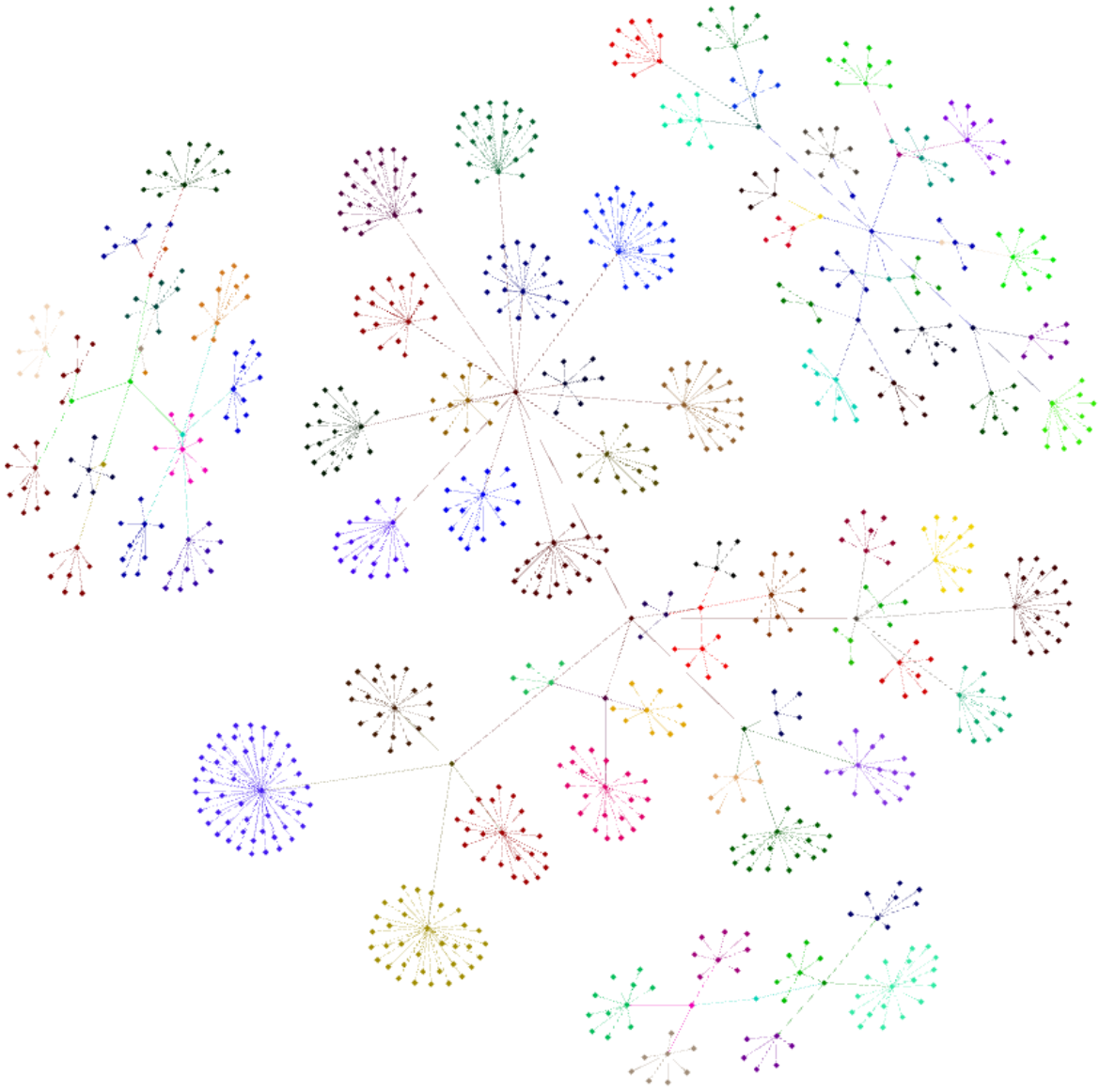

Figure D.1.: Topic hierarchies after fusion from the abstracts entities, the color represents the cluster topics belong to.

- U_3_2: evaluar temperatura concentracion cultivar acido tratamiento ensayo
  - U_2_8: disminucion concentracion diferenciar afecto tratamiento disminuir mayor
    - NOUN: concentracion mezclar
    - U_1_73: carbonar reduccion tejer gas modificacion adaptacion nitrogeno
    - U_1_45: geometrico alto menor reducir disminuir mostrar favorecer
  - U_2_7: temperatura espesor superficie variar electrico incentivar magnetico
    - U_1_40: cubrir forzar espesor superficie pelicula profundidad expresar
    - U_1_62: material estructural electrico electronico magnetico fase caracterizacion
      - NOUN: fase transicion
      - NOUN: material propiedad
      - ADJ: electronico estructural termico morfologico
      - NOUN: caracterizacion caracteristico identificacion validacion
      - ADJ: electrico magnetico optico solar
  - U_2_9: modificar hoja sustanciar aislar inducir proteina reportar
    - U_1_35: identificar control actividad restringir aislar adaptarse bacterium
      - ADJ: aislar reportar bacteriano similar
      - NOUN: actividad cepa secuenciar peptido
      - VERB: aislar identificar ensayar persistir
      - VERB: adaptarse restringir escoger programar
      - NOUN: control perdida capacidad agente
      - NOUN: bacterium microorganismo infeccion mortalidad
    - VERB: medir calcular evitar suministrar
  - U_2_20: area especie centimetro espacial temporal zona tamano
  - U_2_18: tecnologia utilizar producto acido fermentacion disponible nutricional

- U_3_3: cuidar periodo salud entrevistar resultados mesar paciente
  - U_2_13: cuidar confiabilidad brindar nacer entrevistar interno muerte
  - U_2_17: periodo positivo normal descriptivo grupo resultados paciente

Figure D.2.: Some entity topics after fusion from the abstracts dataset.

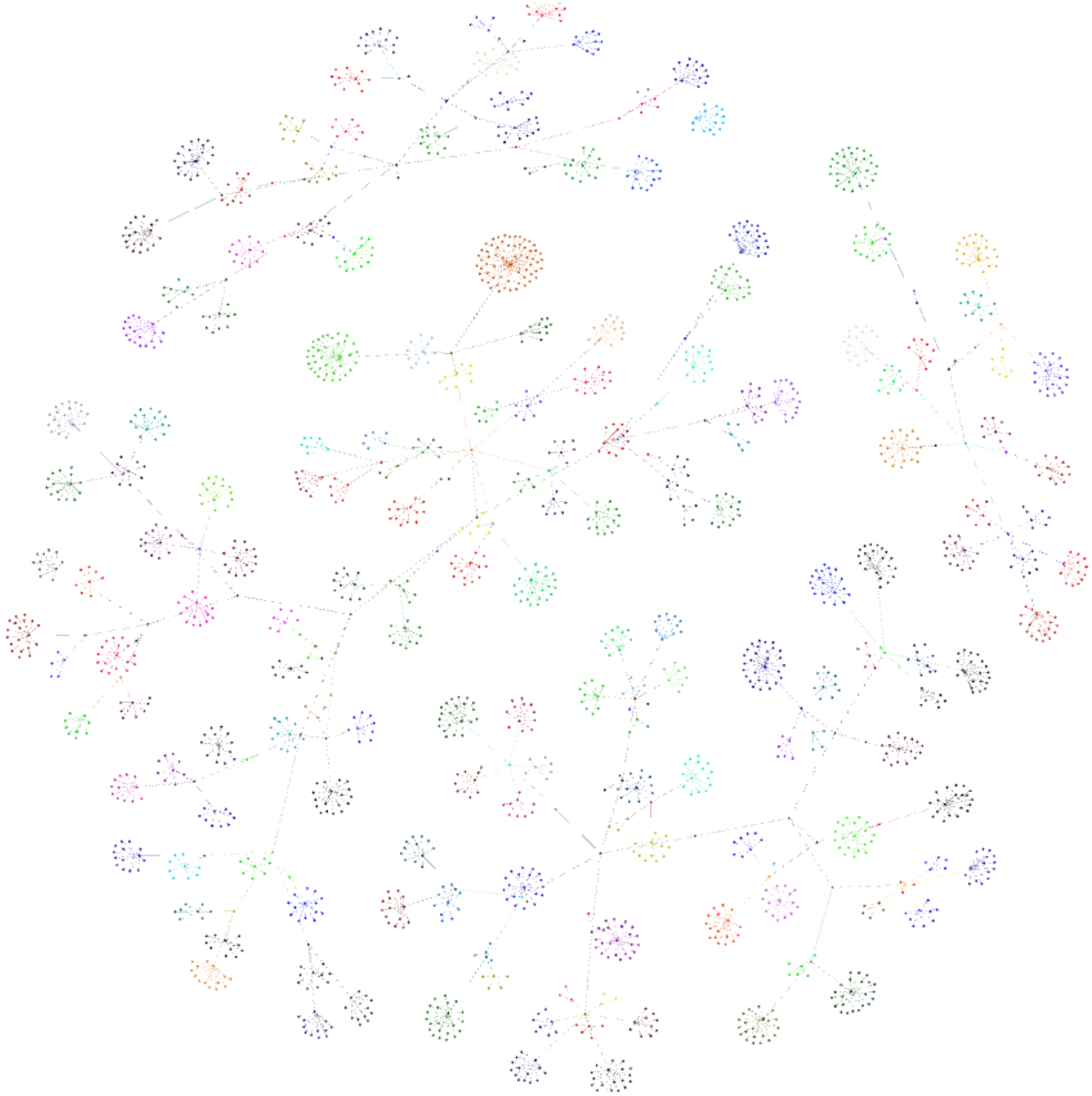

Figure D.3.: Topic hierarchies after fusion from the thesis entities, the color represents the cluster topics belong to.

- U_5_1: acuerdo garantir ciudadania discursar pobreza inteligencia prestacion
  - U_4_3: nacion pagar juridico constitucional ley negar denunciar
    - U_3_6: garantir nacion expedir juridico territorial ley estado
    - U_3_15: disputar tercer_mundo armar el_espectador sufrimiento policia el_tiempo
    - U_3_5: demandar reclamar prevalecer constitucion_politica_de_colombia vislumbrar corte_constitucional corte_suprema_de_justicia
    - U_3_4: espiritual cuestionar entendido siguiendo materializar negar denunciar
  - U_4_2: demandar francia psicologico reflexionar latinoamericano inteligencia pedagogico
    - U_3_0: actualizar comprende inteligencia ieee entorno gestionar horario
    - U_3_9: la_tierra men reflexionar saber dibujar alumno pedagogico
    - U_3_7: banco_de_la_republica capital negociacion beneficiar pagar empresarial cliente
      - U_2_8: fuente turismo beneficiar difundir harvard_business_review mcgraw_hill tecnologia
        - U_1_166: captacion industrializacion escasez desaparicion impulsar especializacion auge
        - U_1_95: liberal apuesto dimensionar reivindicacion tenida cabe
        - U_1_18: stakeholders mintzberg innovar competir sanin harvard_business_review rse
        - U_1_73: industria turismo representantes emprendimiento velar tratar salud
        - U_1_19: neto fuente sectorial estrategias maximizar porcentual
      - U_2_27: contratar banco_de_la_republica monetario administrador gasto pagar financiera
        - U_1_87: interpretarse banco_de_la_republica icp tes superintendencia_financiera_de_colombia financiera junta
          - NOUN: patrimonio colombiano
          - NOUN: montar imponer prestamo pescar
          - ORG: banco_de_la_republica banco tercer_mundo_editores la_violencia
          - ORG: icp mud aaa finagro
          - VERB: interpretarse liquidar empeorar acotar
          - ADJ: contractual patrimonial
          - LOC: mercado administracion industrial base
          - ORG: tes bvc comision_de_regulacion_de_energia_y_gas ens
          - ORG: superintendencia_financiera_de_colombia inversiones iasb dian
          - NOUN: remuneracion transparencia portafolio diversificacion
          - NOUN: administrador agenciar financiamiento respaldar
          - NOUN: banco cooperativo dolar gobernanza
          - ORG: financiera capm meci caja
          - NOUN: inflacion bono
          - NOUN: contabilidad volatilidad sustentabilidad caer

Figure D.4.: Some entity topics after fusion from the thesis dataset.

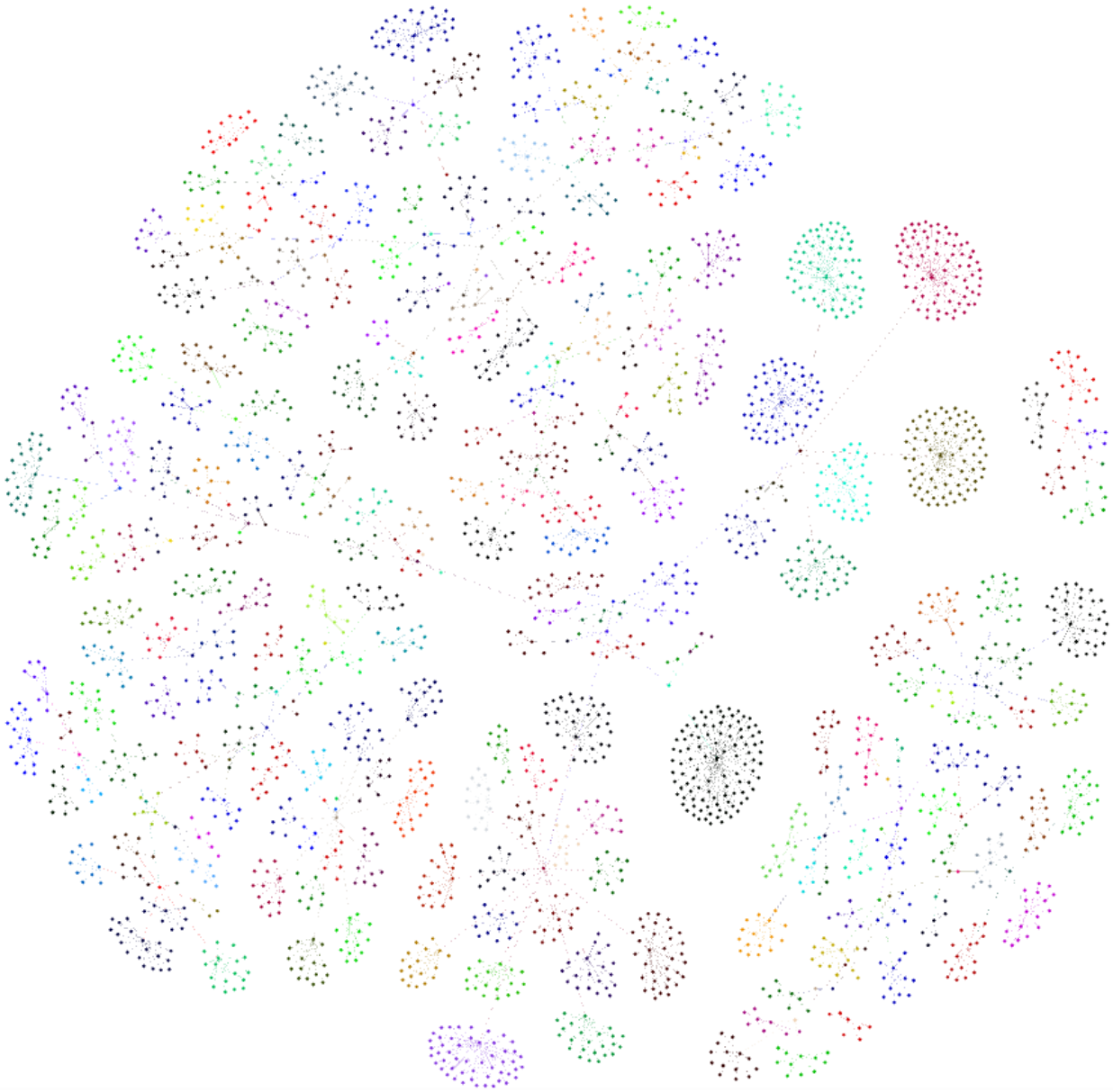

Figure D.5.: Topic hierarchies after fusion from the organisms entities, the color represents the cluster topics belong to.

- U_4_2: snail flower thailand date range biology moth
  - U_3_9: botanical flower herb similar dorsal albany grow
    - U_2_20: flower mountain abbr diameter area petiole plant
      - U_1_114: wheatbelt hering region western_australia margin flowering hill
      - U_1_107: sepal autumn mountain soil spring seed plant
        - ADJ: present mature warm cold
        - VERB: arise stem bloom update
        - NOUN: mountain bloom plate limestone
        - VERB: cultivate propagate pollinate naturalize
        - ADJ: evergreen ornamental
        - ORG: greek university_of_california_press s_award_of_garden_merit lassie
        - ADJ: tall perennial native herbaceous
        - NOUN: spring land mating damage
        - ORG: flower smilax sri_lankan drc
        - VERB: grow flower gain hold
        - NOUN: autumn confusion hibernation shark
        - NOUN: seed growth cluster foliage
        - NOUN: soil cultivation blade moisture
        - NOUN: cultivar frost grin tropic
        - VERB: plant thrive germinate branch
        - NOUN: hybrid west purple germination
        - NOUN: sepal leave
        - NOUN: plant leaf botanist subsp
      - U_1_60: branch botanical bundle diameter stem the_esperance_plain department
      - U_1_108: horticulture publish overlap royal_botanic_gardens authorlink tube feature
      - U_1_10: female spot length pair back week size
  - U_3_7: snail cladobranchia mollusk gastropod aperture wrms bouchet
    - U_2_24: slug south_georgia port caribbean_sea the_falkland_islands mainland guadeloupe
    - U_2_22: lip whorl mollusk the_north_pacific_ocean shell aperture vol
    - U_2_25: ally hyla database ocean bouchet access accept
    - U_2_27: snail marine eulimidae wrms biology displayparents geometridae
    - U_2_23: taxobox gastropod positive superfamilia binomialauthority_ gram binomialauthority
  - U_3_5: hunting density threaten hunt greek new_england ear
  - U_3_1: replication distance dna ability associate genetic research
  - U_3_4: taxonomy habitat india biology the_black_sea canada moth

Figure D.6.: Some entity topics after fusion from the organisms dataset.

E. Appendix: Abstracts single-view topic clustering results

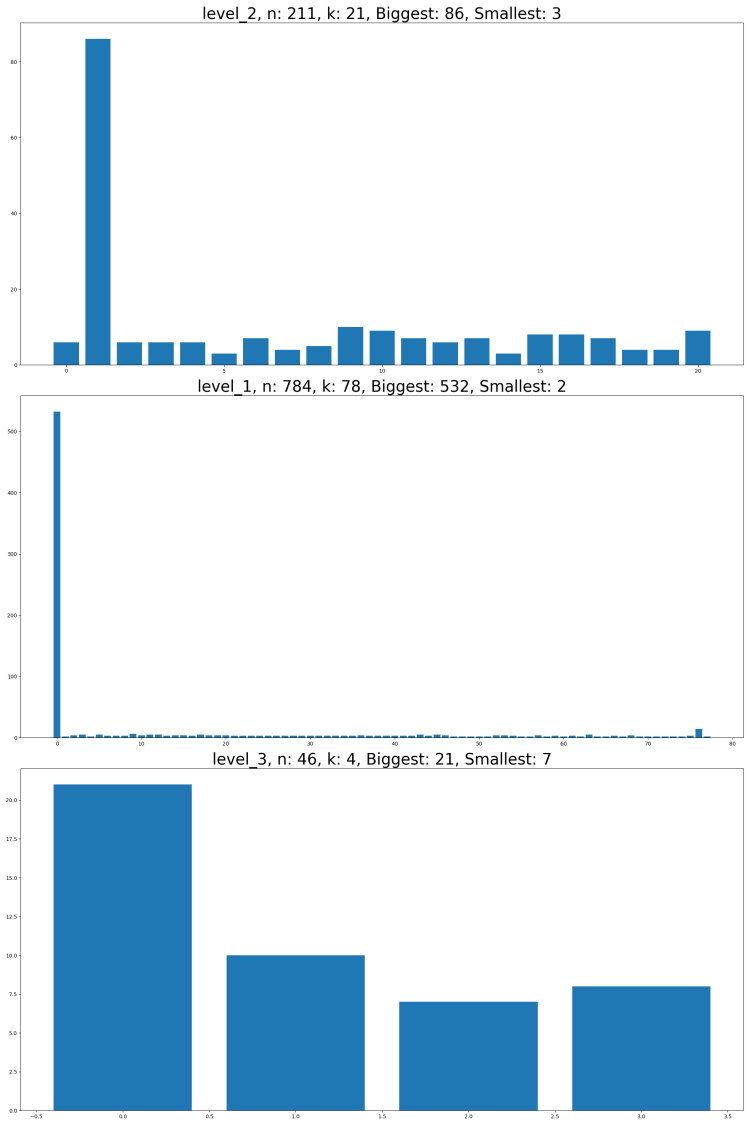

Figure E.1.: Abstracts topic clustering per tree level using only each topic words probability information.

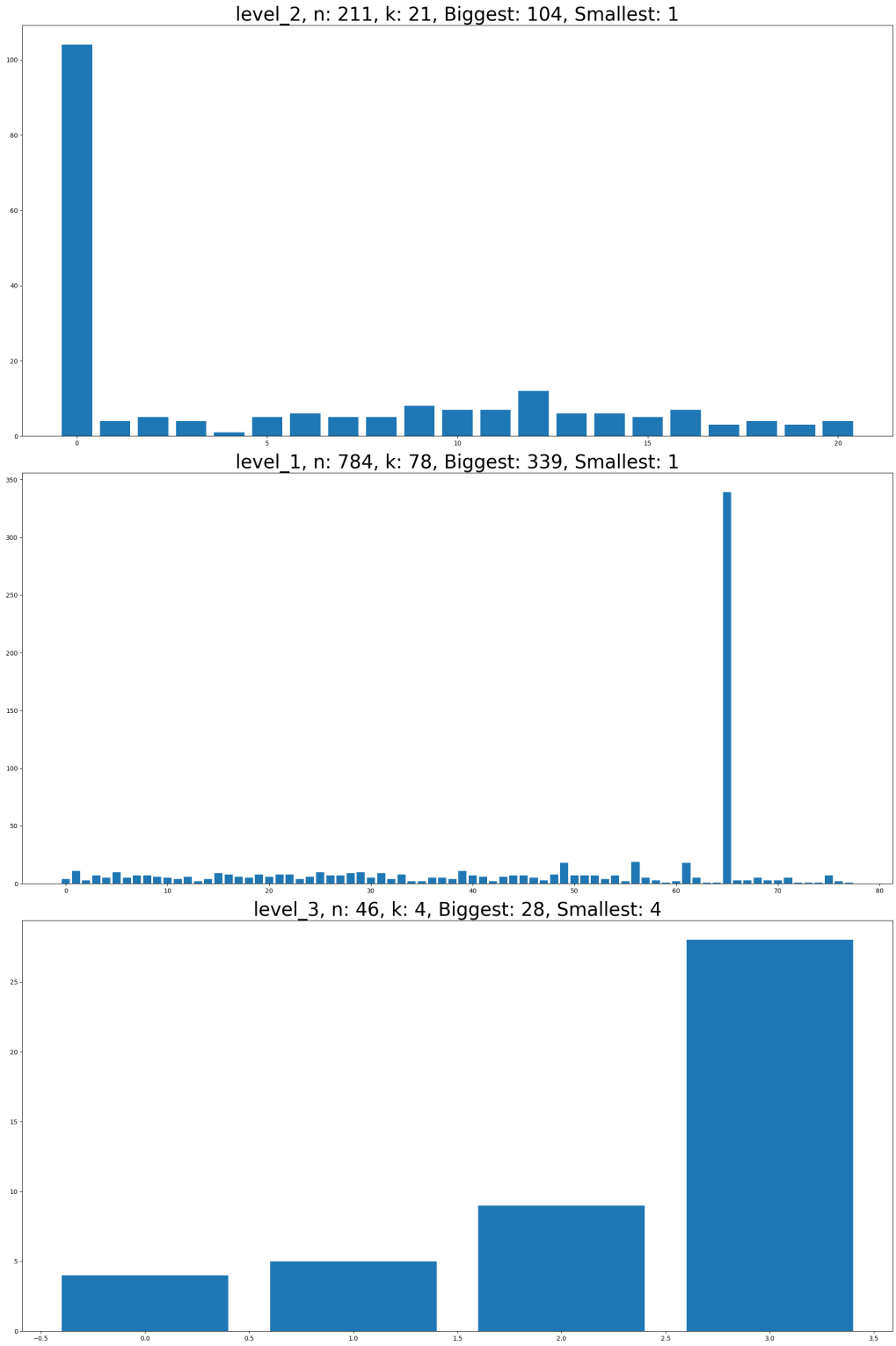

Figure E.2.: Abstracts topic clustering per tree level using only each topic documents probability information.

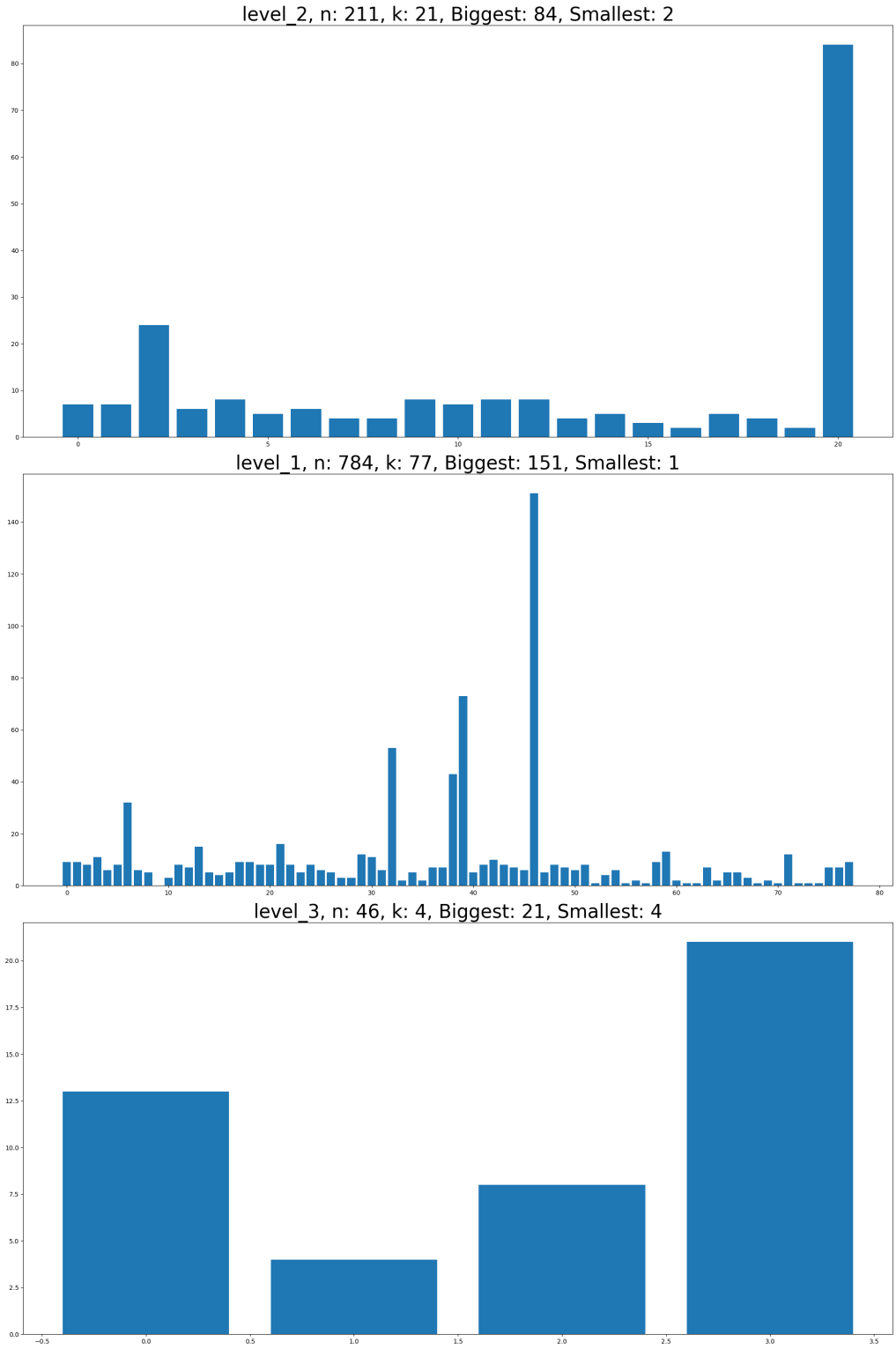

Figure E.3.: Abstracts topic clustering per tree level using only the TF-IDF score of the words present on each topic documents.

F. Appendix: Theses single-view topic clustering results

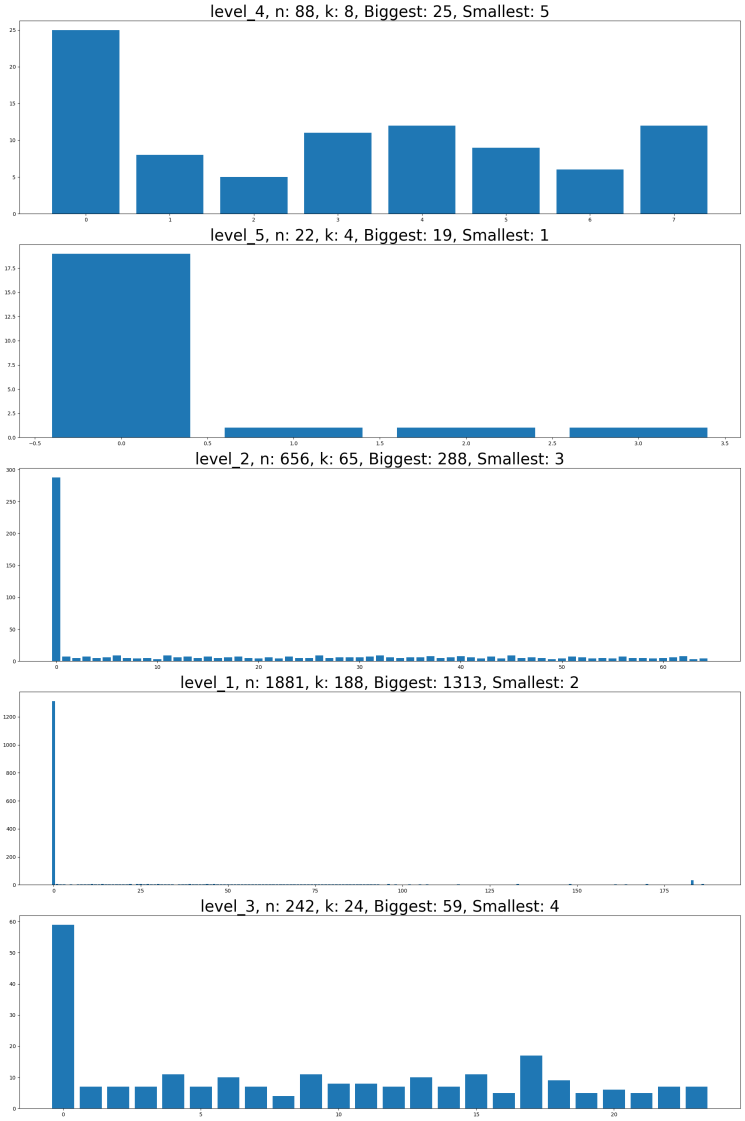

Figure F.1.: Thesis topic clustering per tree level using only each topic words probability information.

Figure F.2.: Thesis topic clustering per tree level using only each topic documents probability information.

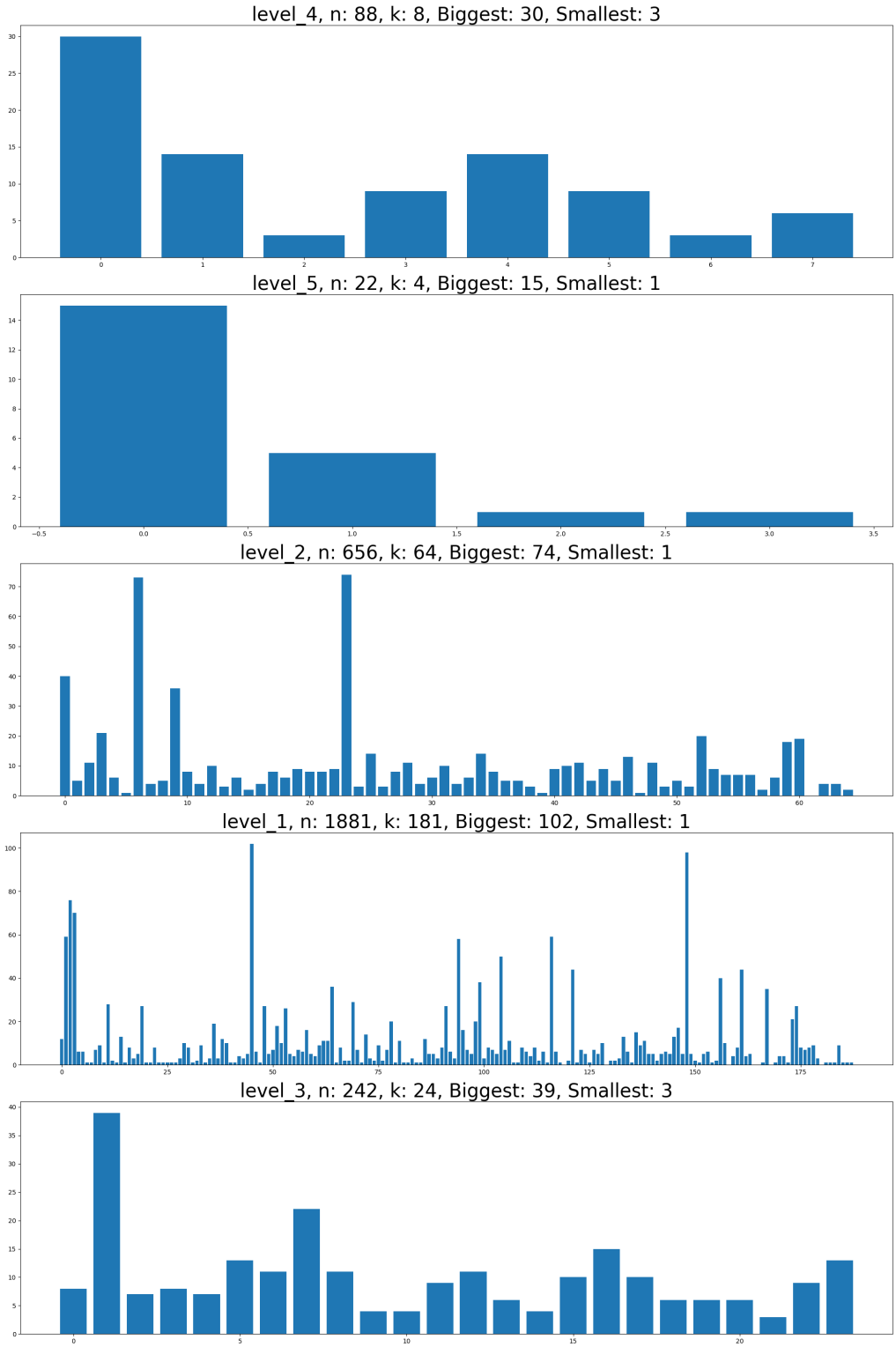

Figure F.3.: Thesis topic clustering per tree level using only the TF-IDF score of the words present on each topic documents.

G. Appendix: Organisms single-view topic clustering results

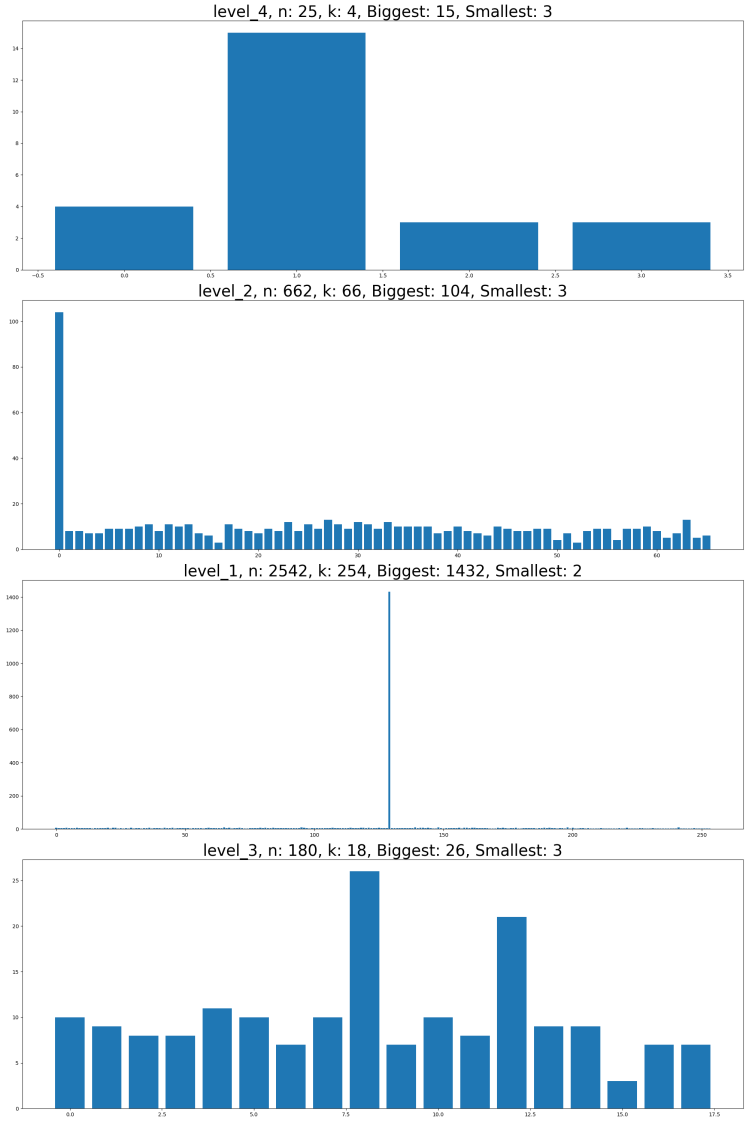

Figure G.1.: Organisms topic clustering per tree level using only each topic words probability information.

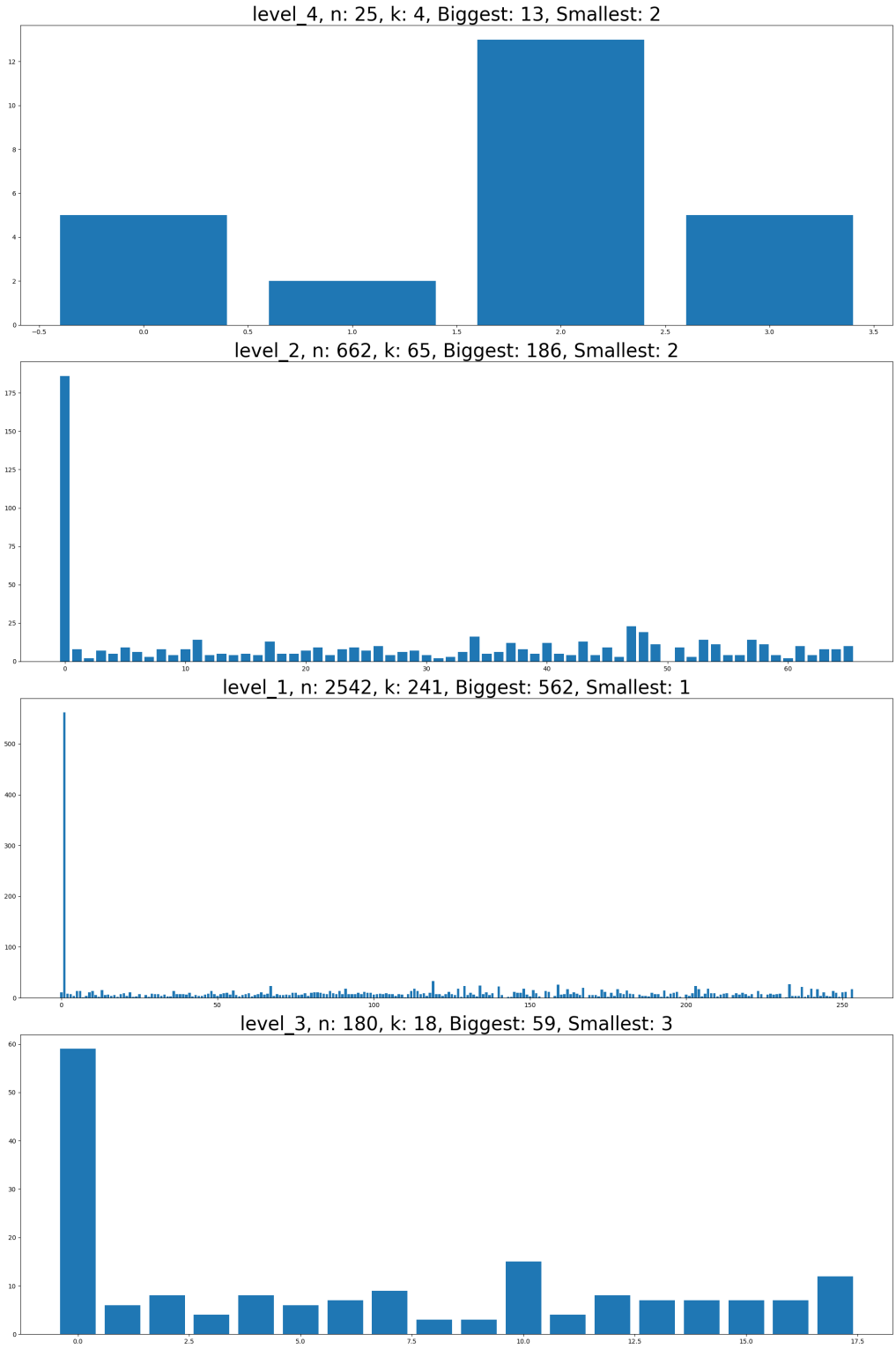

Figure G.2.: Organisms topic clustering per tree level using only each topic documents probability information.

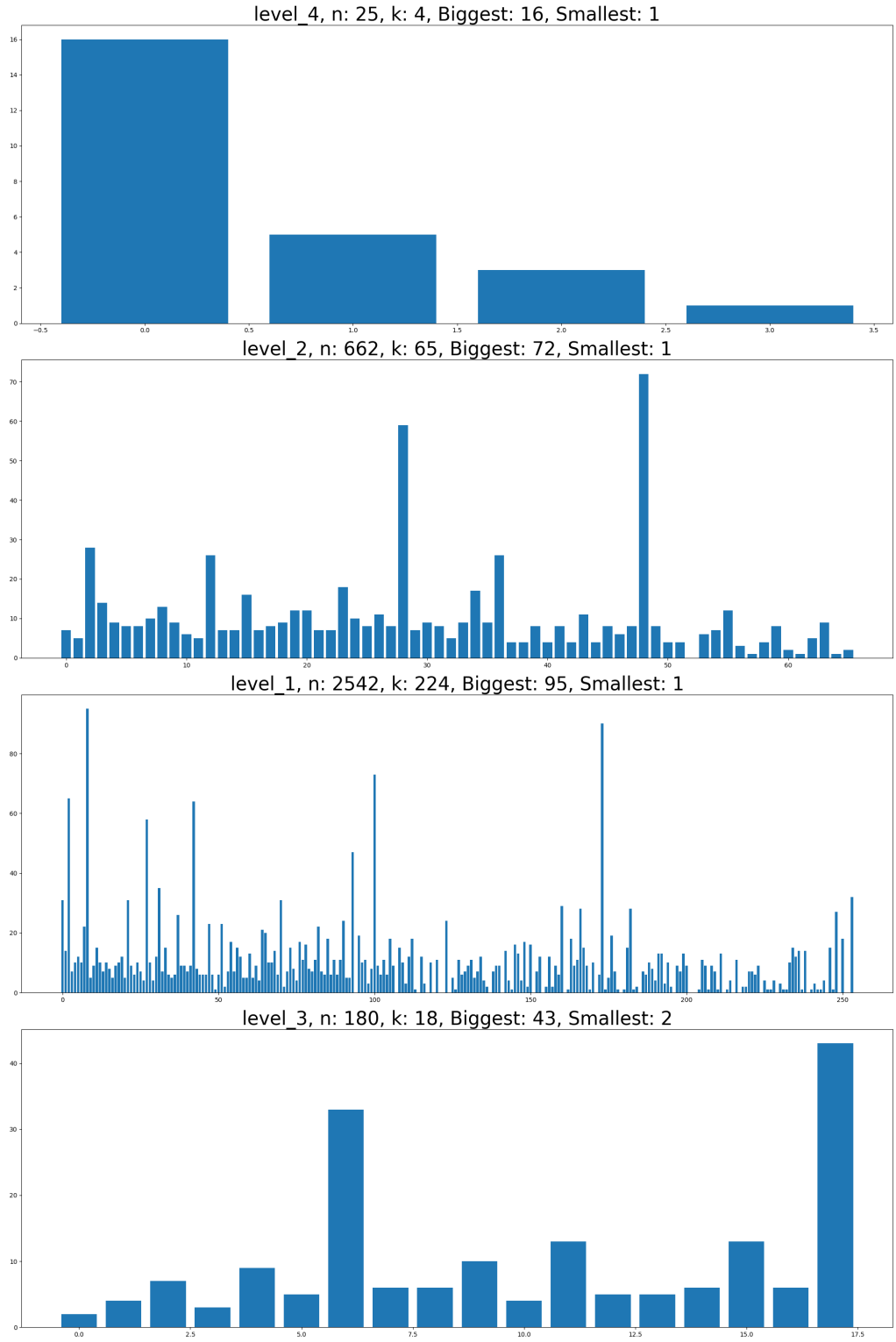

Figure G.3.: Organisms topic clustering per tree level using only the TF-IDF score of the words present on each topic documents.

H. Appendix: Feature concatenation topic clustering results

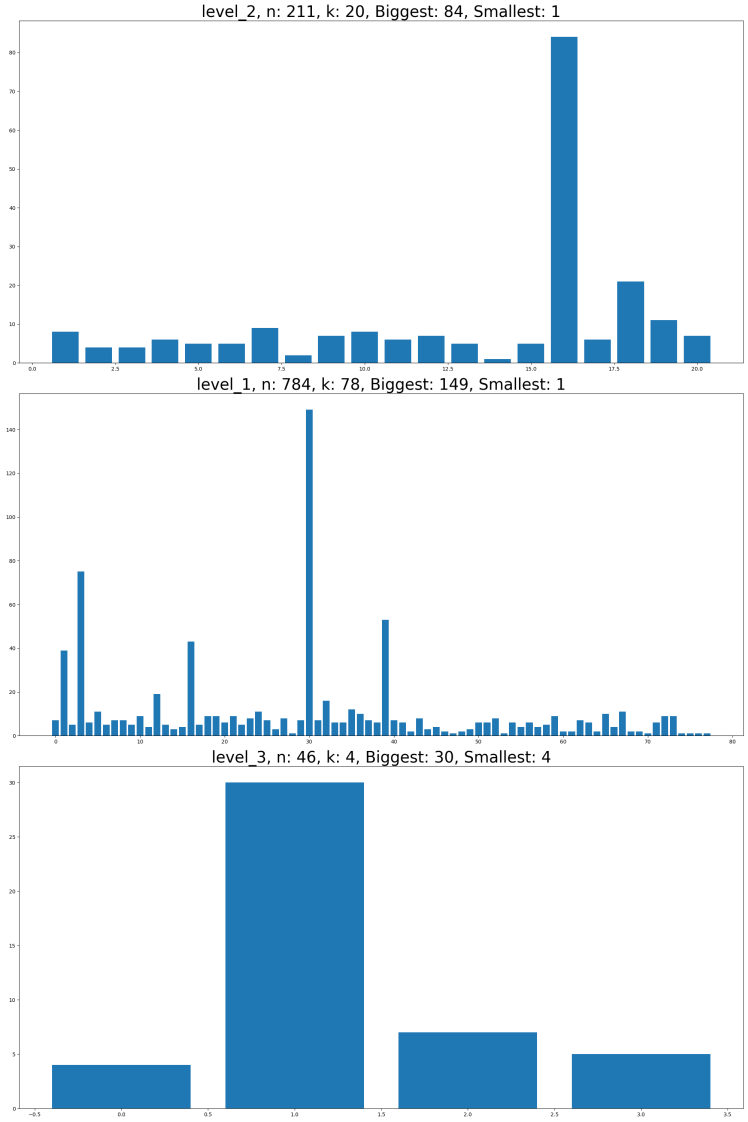

Figure H.1.: Abstracts topic clustering per tree level with all views concatenated as a single one.

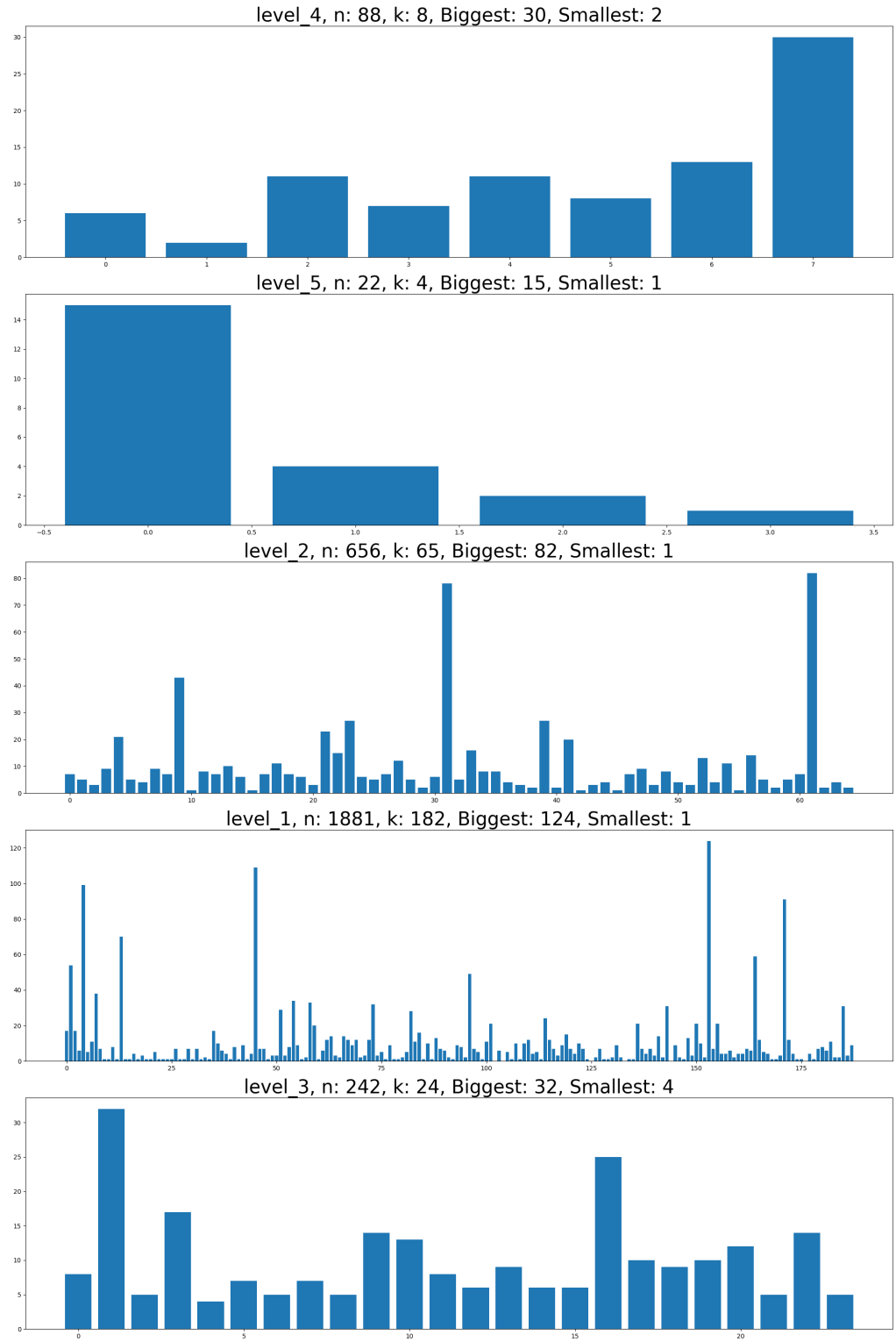

Figure H.2.: Thesis topic clustering per tree level with all views concatenated as a single one.

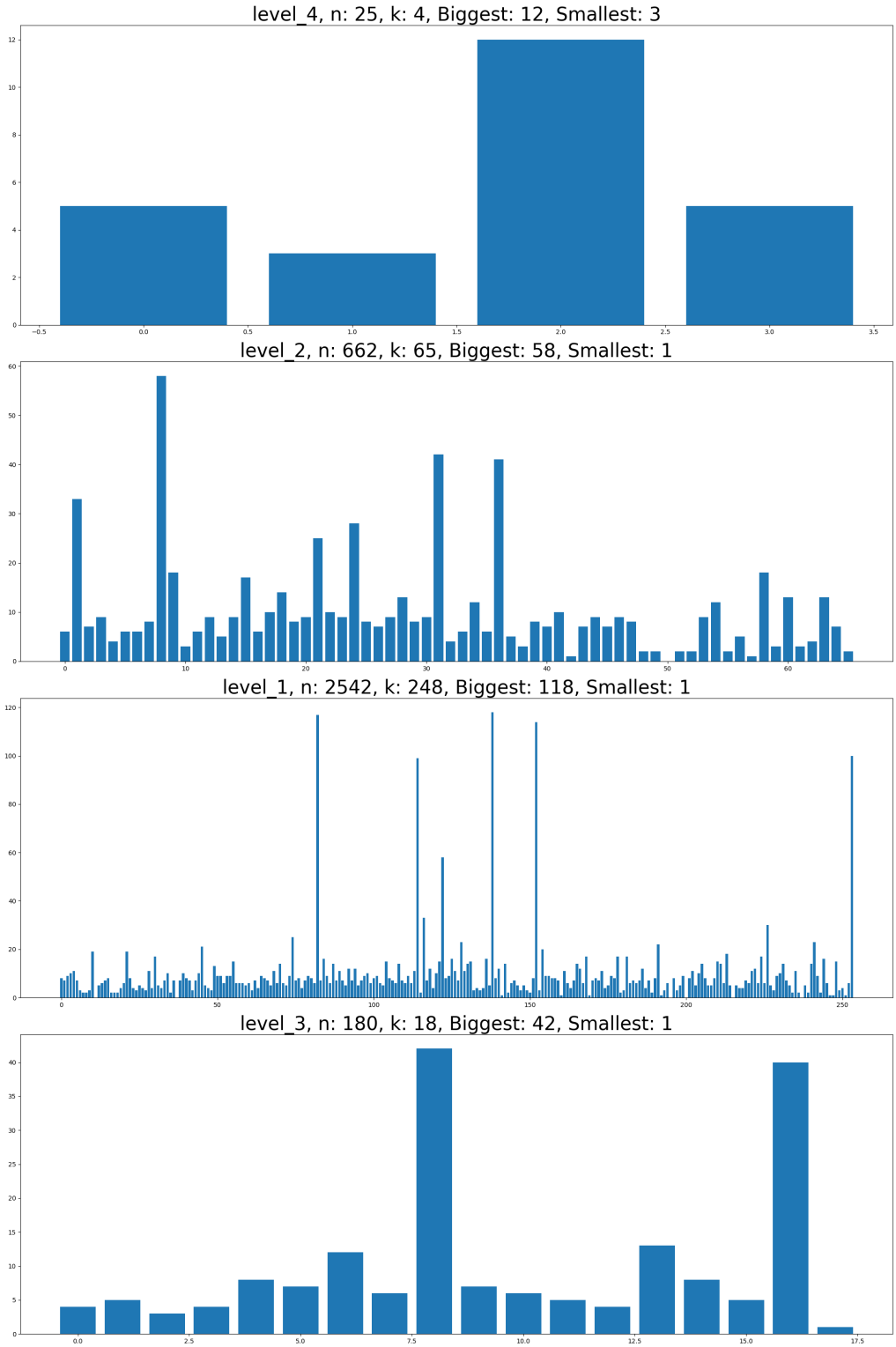

Figure H.3.: Organisms topic clustering per tree level with all views concatenated as a single one.

I. Appendix: Organisms subcategories

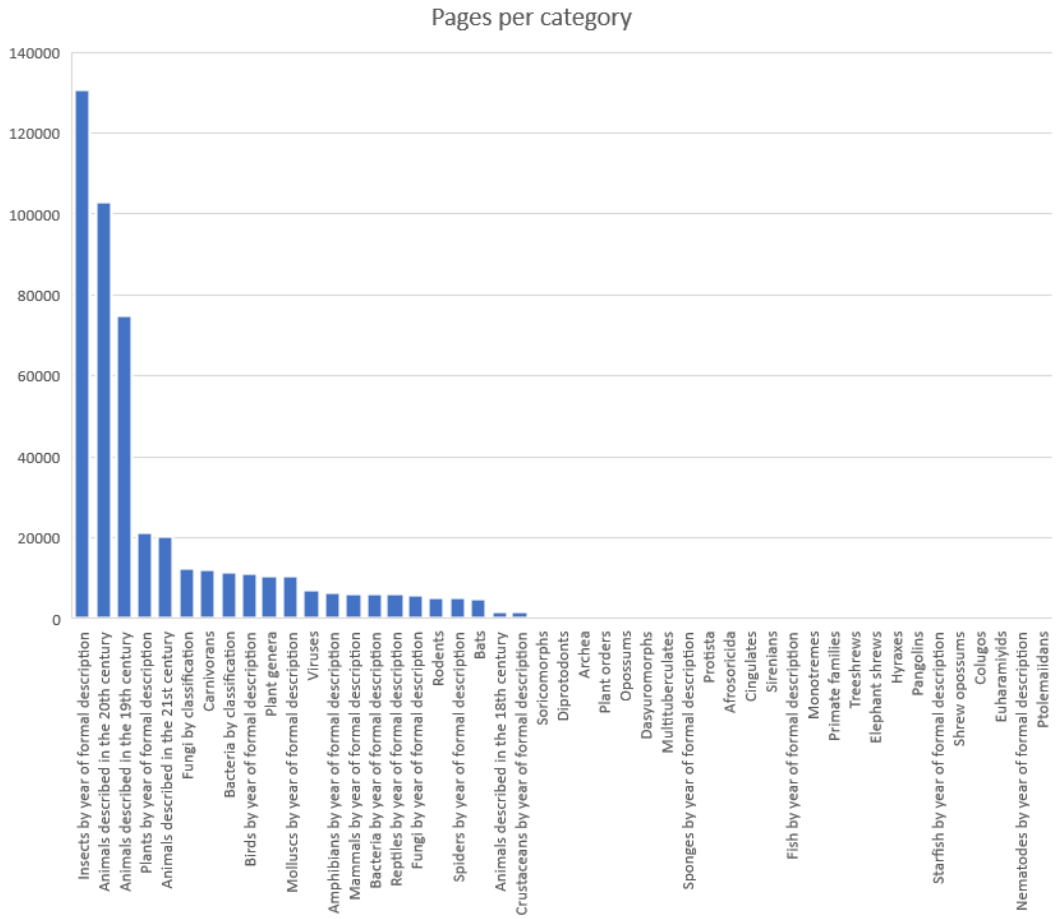

Figure I.1.: Histogram for the number of pages per Wikipedia category.

The organisms subcategories are:

- Animals (https://en.wikipedia.org/wiki/Category:Animals)
- Archaea (https://en.wikipedia.org/wiki/Category:Archaea)
- Bacteria (https://en.wikipedia.org/wiki/Category:Bacteria)
- Fungi (including yeasts) (https://en.wikipedia.org/wiki/Category:Fungi)
- Plants (including most algae) (https://en.wikipedia.org/wiki/Category:Plants)
- Protista (https://en.wikipedia.org/wiki/Category:Protista)
- Viruses (https://en.wikipedia.org/wiki/Category:Viruses)

The subcategories downloaded from the **Animalas** category are:

- Afrosoricida (https://en.wikipedia.org/wiki/Category:Afrosoricida)
- Amphibians by year of formal description (https://en.wikipedia.org/wiki/Category:Amphibians_by_year_of_formal_description)
- Animals described in the 18th century (https://en.wikipedia.org/wiki/Category:Animals_described_in_the_18th_century)
- Animals described in the 19th century (https://en.wikipedia.org/wiki/Category:Animals_described_in_the_19th_century)
- Animals described in the 20th century (https://en.wikipedia.org/wiki/Category:Animals_described_in_the_20th_century)
- Animals described in the 21st century (https://en.wikipedia.org/wiki/Category:Animals_described_in_the_21st_century)
- Bats (https://en.wikipedia.org/wiki/Category:Bats)
- Birds by year of formal description (https://en.wikipedia.org/wiki/Category:Birds_by_year_of_formal_description)
- Carnivorans (https://en.wikipedia.org/wiki/Category:Carnivorans)
- Cingulates (https://en.wikipedia.org/wiki/Category:Cingulates)
- Colugos (https://en.wikipedia.org/wiki/Category:Colugos)

- Crustaceans by year of formal description (https://en.wikipedia.org/wiki/Category:Crustaceans_by_year_of_formal_description)
- Dasyuromorphs (https://en.wikipedia.org/wiki/Category:Dasyuromorphs)
- Diprotodonts (https://en.wikipedia.org/wiki/Category:Diprotodonts)
- Elephant shrews (https://en.wikipedia.org/wiki/Category:Elephant_shrews)
- Euharamiyids (https://en.wikipedia.org/wiki/Category:Euharamiyids)
- Fish by year of formal description (https://en.wikipedia.org/wiki/Category:Fish_by_year_of_formal_description)
- Hyraxes (https://en.wikipedia.org/wiki/Category:Hyraxes)
- Insects by year of formal description (https://en.wikipedia.org/wiki/Category:Insects_by_year_of_formal_description)
- Mammals by year of formal description (https://en.wikipedia.org/wiki/Category:Mammals_by_year_of_formal_description)
- Molluscs by year of formal description (https://en.wikipedia.org/wiki/Category:Molluscs_by_year_of_formal_description)
- Monotremes (https://en.wikipedia.org/wiki/Category:Monotremes)
- Multituberculates (https://en.wikipedia.org/wiki/Category:Multituberculates)
- Nematodes by year of formal description (https://en.wikipedia.org/wiki/Category:Nematodes_by_year_of_formal_description)
- Opossums (https://en.wikipedia.org/wiki/Category:Opossums)
- Pangolins (https://en.wikipedia.org/wiki/Category:Pangolins)
- Primate families (https://en.wikipedia.org/wiki/Category:Primate_families)
- Ptolemaiidans (https://en.wikipedia.org/wiki/Category:Ptolemaiidans)
- Reptiles by year of formal description (https://en.wikipedia.org/wiki/Category:Reptiles_by_year_of_formal_description)
- Rodents (https://en.wikipedia.org/wiki/Category:Rodents)
- Shrew opossums (https://en.wikipedia.org/wiki/Category:Shrew_opossums)
- Sirenians (https://en.wikipedia.org/wiki/Category:Sirenians)

- Soricomorphs (https://en.wikipedia.org/wiki/Category:Soricomorphs)
- Spiders by year of formal description (https://en.wikipedia.org/wiki/Category:Spiders_by_year_of_formal_description)
- Sponges by year of formal description (https://en.wikipedia.org/wiki/Category:Sponges_by_year_of_formal_description)
- Starfish by year of formal description (https://en.wikipedia.org/wiki/Category:Starfish_by_year_of_formal_description)
- Treeshrews (https://en.wikipedia.org/wiki/Category:Treeshrews)

The subcategories downloaded from the **Bacteria** category are:

- Bacteria by year of formal description (https://en.wikipedia.org/wiki/Category:Bacteria_by_year_of_formal_description)
- Bacteria by classification (https://en.wikipedia.org/wiki/Category:Bacteria_by_classification)

The subcategories downloaded from the **Fungi** category are:

- Fungi by year of formal description (https://en.wikipedia.org/wiki/Category:Fungi_by_year_of_formal_description)
- Fungi by classification (https://en.wikipedia.org/wiki/Category:Fungi_by_classification)

The subcategories downloaded from the **Plants** category are:

- Plants by year of formal description (https://en.wikipedia.org/wiki/Category:Fungi_by_year_of_formal_description)
- Plant genera (https://en.wikipedia.org/wiki/Category:Plant_genera)
- Plant orders(https://en.wikipedia.org/wiki/Category:Plant_orders)

The **Protista**, **Viruses** and **Archea** categories were fully downloaded.

Bibliography

[1] Stephen E. Palmer. "Hierarchical structure in perceptual representation". In: *Cognitive Psychology* 9.4 (Oct. 1977), pp. 441–474. ISSN: 0010-0285. DOI: 10.1016/0010-0285(77)90016-0. URL: https://www.sciencedirect.com/science/article/pii/0010028577900160.

[2] E. Wachsmuth, M. W. Oram, and D. I. Perrett. "Recognition of Objects and Their Component Parts: Responses of Single Units in the Temporal Cortex of the Macaque". In: *Cerebral Cortex* 4.5 (Sept. 1994), pp. 509–522. ISSN: 1047-3211. DOI: 10.1093/cercor/4.5.509. URL: https://academic.oup.com/cercor/article-lookup/doi/10.1093/cercor/4.5.509.

[3] N K Logothetis and D L Sheinberg. "Visual Object Recognition". In: *Annual Review of Neuroscience* 19.1 (Mar. 1996), pp. 577–621. ISSN: 0147-006X. DOI: 10.1146/annurev.ne.19.030196.003045. URL: http://www.annualreviews.org/doi/10.1146/annurev.ne.19.030196.003045.

[4] Daniel D. Lee and H. Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization". In: *Nature* 401.6755 (Oct. 1999), pp. 788–791. ISSN: 00280836. DOI: 10.1038/44565. URL: http://www.nature.com/articles/44565.

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3.Jan (2003), pp. 993–1022. ISSN: ISSN 1533-7928. URL: http://www.jmlr.org/papers/v3/blei03a.html.

[6] Thomas L. Griffiths et al. "Hierarchical Topic Models and the Nested Chinese Restaurant Process". In: *Advances in Neural Information Processing Systems* (2003), pp. 17–24. URL: https://papers.nips.cc/paper/2466-hierarchical-topic-models-and-the-nested-chinese%20-restaurant-process.pdf.

[7] Stella X. Yu and Jianbo Shi. "Multiclass spectral clustering". In: *Proceedings of the IEEE International Conference on Computer Vision*. Vol. 1. Institute of Electrical and Electronics Engineers Inc., 2003, pp. 313–319. DOI: 10.1109/iccv.2003.1238361. URL: https://ieeexplore.ieee.org/abstract/document/1238361.

[8] S. Bickel and T. Scheffer. "Multi-View Clustering". In: *Fourth IEEE International Conference on Data Mining (ICDM'04)*. IEEE, 2004, pp. 19–26. ISBN: 0-7695-2142-8. DOI: 10.1109/ICDM.2004.10095. URL: http://ieeexplore.ieee.org/document/1410262/.

[9] Nevin L Zhang and Lzhang@cs Ust Hk. *Hierarchical Latent Class Models for Cluster Analysis*. Tech. rep. 2004, pp. 697–723. URL: https://www.jmlr.org/papers/volume5/zhang04a/zhang04a.pdf.

[10] David Newman, Chaitanya Chemudugunta, and Padhraic Smyth. "Statistical entity-topic models". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Vol. 2006. Association for Computing Machinery, 2006, pp. 680–686. ISBN: 1595933395. DOI: 10.1145/1150402.1150487.

[11] Li Wei and Andrew McCallum. "Pachinko allocation: DAG-structured mixture models of topic correlations". In: *ACM International Conference Proceeding Series*. Vol. 148. 2006, pp. 577–584. ISBN: 1595933832. DOI: 10.1145/1143844.1143917. URL: https://dl.acm.org/doi/abs/10.1145/1143844.1143917.

[12] David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. "The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies". In: (Oct. 2007). URL: https://arxiv.org/abs/0710.0845.

[13] David Mimno, Wei Li, and Andrew McCallum. "Mixtures of hierarchical topics with Pachinko allocation". In: *ACM International Conference Proceeding Series*. Vol. 227. 2007, pp. 633–640. DOI: 10.1145/1273496.1273576. URL: https://dl.acm.org/doi/abs/10.1145/1273496.1273576.

[14] Tomoaki Nakamura, Takayuki Nagai, and Naoto Iwahashi. "Multimodal object categorization by a robot". In: *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, Oct. 2007, pp. 2415–2420. ISBN: 978-1-4244-0911-2. DOI: 10.1109/IROS.2007.4399634. URL: http://ieeexplore.ieee.org/document/4399634/.

[15] Chaitanya Chemudugunta et al. "Modeling Documents by Combining Semantic Concepts with Unsupervised Statistical Learning". In: 2008, pp. 229–244. DOI: 10.1007/978-3-540-88564-1{_}15.

[16] Yi Wang, Nevin L. Zhang, and Tao Chen. "Latent tree models and approximate inference in Bayesian networks". In: *Journal of Artificial Intelligence Research* 32 (Aug. 2008), pp. 879–900. ISSN: 10769757. DOI: 10.1613/jair.2530. URL: https://www.jair.org/index.php/jair/article/view/10564.

[17] Nevin L. Zhang et al. "Latent tree models and diagnosis in traditional Chinese medicine". In: *Artificial Intelligence in Medicine* 42.3 (Mar. 2008), pp. 229–245. ISSN: 09333657. DOI: 10.1016/j.artmed.2007.10.004. URL: https://www.sciencedirect.com/science/article/pii/S0933365707001443.

[18] David Andrzejewski, Xiaojin Zhu, and Mark Craven. "Incorporating domain knowledge into topic modeling via Dirichlet forest priors". In: *ACM International Conference Proceeding Series*. Vol. 382. 2009. ISBN: 9781605585161. DOI: 10.1145/1553374.1553378.

[19] Jonathan Chang et al. *Reading Tea Leaves: How Humans Interpret Topic Models*. Tech. rep. 2009. URL: http://rexa.info.

[20] Tomoaki Nakamura, Takayuki Nagai, and Naoto Iwahashi. "Grounding of word meanings in multimodal concepts using LDA". In: *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, Oct. 2009, pp. 3943–3948. ISBN: 978-1-4244-3803-7. DOI: 10.1109/IROS.2009.5354736. URL: http://ieeexplore.ieee.org/document/5354736/.

[21] Guangcan Liu et al. "Robust Recovery of Subspace Structures by Low-Rank Representation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.1 (Oct. 2010), pp. 171–184. DOI: 10.1109/TPAMI.2012.88. URL: http://arxiv.org/abs/1010.2955%20http://dx.doi.org/10.1109/TPAMI.2012.88.

[22] James Petterson et al. *Word Features for Latent Dirichlet Allocation*. Tech. rep. 2010, pp. 1921–1929.

[23] Nakatani Shuyo. *Language Detection Library for Java*. 2010. URL: http://code.google.com/p/language-detection/.

[24] Abhishek Kumar and Hal Daumé III. *A Co-training Approach for Multi-view Spectral Clustering*. Tech. rep. 2011. URL: http://legacydirs.umiacs.umd.edu/~abhishek/cospectral.icml11.pdf.

[25] Abhishek Kumar, Piyush Rai, and Hal Daumé III. *Co-regularized Multi-view Spectral Clustering*. Tech. rep. 2011.

[26] David Mimno et al. *Optimizing Semantic Coherence in Topic Models*. Tech. rep. 2011, pp. 262–272. URL: https://www.aclweb.org/anthology/D11-1024.pdf.

[27] Tomoaki Nakamura, Takayuki Nagai, and Naoto Iwahashi. "Bag of multimodal LDA models for concept formation". In: *2011 IEEE International Conference on Robotics and Automation*. IEEE, May 2011, pp. 6233–6238. ISBN: 978-1-61284-386-5. DOI: 10.1109/ICRA.2011.5980324. URL: http://ieeexplore.ieee.org/document/5980324/.

[28] Ehsan Elhamifar and Rene Vidal. "Sparse Subspace Clustering: Algorithm, Theory, and Applications". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.11 (Mar. 2012), pp. 2765–2781. URL: http://arxiv.org/abs/1203.1005.

[29] Jagadeesh Jagarlamudi, Hal Daumé Iii, and Raghavendra Udupa. *Incorporating Lexical Priors into Topic Models*. Tech. rep. 2012, pp. 204–213. URL: https://www.aclweb.org/anthology/E12-1021.pdf.

[30] Xiao Cai, Feiping Nie, and Heng Huang. *Multi-View K-Means Clustering on Big Data*. Tech. rep. 2013. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.415.8610&rep=rep1&type=pdf.

[31] Zhiyuan Chen et al. "Discovering Coherent Topics Using General Knowledge Data Mining View project Web-KDD-KDD Workshop Series on Web Mining and Web Usage Analysis View project Discovering Coherent Topics Using General Knowledge". In: *dl.acm.org* (2013), pp. 209–218. DOI: 10.1145/2505515.2505519. URL: http://dx.doi.org/10.1145/2505515.2505519.

[32] Zhiyuan Chen et al. "Leveraging Multi-Domain Prior Knowledge in Topic Models". In: *IJCAI International Joint Conference on Artificial Intelligence*. Nov. 2013, pp. 2071–2077.

[33] Linmei Hu et al. "Incorporating entities in news topic modeling". In: *Communications in Computer and Information Science*. Vol. 400. Springer Verlag, Nov. 2013, pp. 139–150. ISBN: 9783642416439. DOI: 10.1007/978-3-642-41644-6{_}14. URL: https://link.springer.com/chapter/10.1007/978-3-642-41644-6_14.

[34] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. "Linguistic Regularities in Continuous Space Word Representations". In: June (2013), pp. 746–751.

[35] Tomas Mikolov et al. *Distributed Representations of Words and Phrases and their Compositionality*. Tech. rep. 2013. URL: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and.

[36] Tomas Mikolov et al. "Efficient estimation of word representations in vector space". In: *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*. International Conference on Learning Representations, ICLR, Jan. 2013.

[37] Konstantinos N. Vavliakis, Andreas L. Symeonidis, and Pericles A. Mitkas. "Event identification in web social media through named entity recognition and topic modeling". In: *Data and Knowledge Engineering* 88 (Nov. 2013), pp. 1–24. ISSN: 0169023X. DOI: 10.1016/j.datak.2013.08.006.

[38] Yuening Hu et al. "Interactive topic modeling". In: *Mach Learn* 95 (2014), pp. 423–469. DOI: 10.1007/s10994-013-5413-0. URL: http://www.policyagendas.org/page/topic-codebook..

[39] Yeqing Li et al. *Large-Scale Multi-View Spectral Clustering with Bipartite Graph*. Tech. rep. 2015. URL: https://dl.acm.org/doi/10.5555/2886521.2886704.

[40] Zechao Li et al. "Robust structured subspace learning for data representation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.10 (Oct. 2015), pp. 2085–2098. ISSN: 01628828. DOI: 10.1109/TPAMI.2015.2400461. URL: https://ieeexplore.ieee.org/document/7031960.

[41] Andrew J. McMinn and Joemon M. Jose. "Real-time entity-based event detection for twitter". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 9283. Springer Verlag, 2015, pp. 65–77. ISBN: 9783319240268. DOI: 10.1007/978-3-319-24027-5{_}6. URL: https://link.springer.com/chapter/10.1007/978-3-319-24027-5_6.

[42] John Paisley et al. "Nested hierarchical dirichlet processes". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.2 (Feb. 2015), pp. 256–270. ISSN: 01628828. DOI: 10.1109/TPAMI.2014.2318728. URL: https://ieeexplore.ieee.org/abstract/document/6802355.

[43] Zhao Zhang et al. "Joint low-rank and sparse principal feature coding for enhanced robust representation and visual classification". In: *IEEE Transactions on Image Processing* 25.6 (June 2016), pp. 2429–2443. ISSN: 10577149. DOI: 10.1109/TIP.2016.2547180. URL: https://ieeexplore.ieee.org/document/7442126.

[44] Mehdi Allahyari and Krys Kochut. "Discovering Coherent Topics with Entity Topic Models". In: *Proceedings - 2016 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2016*. Institute of Electrical and Electronics Engineers Inc., Jan. 2017, pp. 26–33. ISBN: 9781509044702. DOI: 10.1109/WI.2016.0015.

[45] Peixian Chen et al. "Latent Tree Models for Hierarchical Topic Detection". In: *Artificial Intelligence* 250 (May 2017), pp. 105–124. URL: http://arxiv.org/abs/1605.06650.

[46] Zhourong Chen et al. *Sparse Boltzmann Machines with Structure Learning as Applied to Text Analysis*. Tech. rep. 2017. URL: www.aaai.org.

[47] Matthew Honnibal and Ines Montani. "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing". 2017.

[48] Ashish Vaswani et al. "Transformer: Attention is all you need". In: *Advances in Neural Information Processing Systems 30* (2017), pp. 5998–6008. ISSN: 10495258. URL: https://arxiv.org/abs/1706.03762.

[49] Jing Zhao et al. "Multi-view learning overview: Recent progress and new challenges". In: *Information Fusion* 38 (2017), pp. 43–54. ISSN: 15662535. DOI: 10.1016/j.inffus.2017.02.007. URL: http://dx.doi.org/10.1016/j.inffus.2017.02.007.

[50] Xiaojun Chen et al. "Spectral clustering of large-scale data by directly solving normalized cut". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, July 2018, pp. 1206–1215. ISBN: 9781450355520. DOI: 10.1145/3219819.3220039. URL: https://dl.acm.org/doi/10.1145/3219819.3220039.

[51] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: (Oct. 2018). URL: http://arxiv.org/abs/1810.04805.

[52] Zhao Kang et al. "Multi-graph Fusion for Multi-view Spectral Clustering". In: *Knowledge-Based Systems* 189 (Sept. 2019). URL: http://arxiv.org/abs/1909.06940.

[53] Alec Radford et al. "Language Models are Unsupervised Multitask Learners". In: (2019). URL: http://www.persagen.com/files/misc/radford2019language.pdf.

[54] Tom B. Brown et al. "Language Models are Few-Shot Learners". In: *arXiv* (May 2020). URL: http://arxiv.org/abs/2005.14165.