Bachelorarbeit am Institut für Informatik der Freien Universität Berlin

Human-Centered Computing (HCC)

Comparing interpretability techniques for unsupervised topic extraction

Tim Korjakow

Matrikelnummer: 372862

tim.korjakow@gmx.de

Betreuerin und Erstgutachterin: Prof. Dr. C. Müller-Birn

Zweitgutachter: Prof. Dr. K. Müller

Berlin, 31.07.2019

Eidesstattliche Erklärung

Ich versichere hiermit an Eides Statt, dass diese Arbeit von niemand anderem als meiner Person verfasst worden ist. Alle verwendeten Hilfsmittel wie Berichte, Bücher, Internetseiten oder ähnliches sind im Literaturverzeichnis angegeben, Zitate aus fremden Arbeiten sind als solche kenntlich gemacht. Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungskommission vorgelegt und auch nicht veröffentlicht.

Berlin, den July 16, 2019

<Name>

Abstract

<Please summarize your thesis in a brief but meaningful way (about one page). Include in your abstract the topic of this thesis, important contents, results of your research and an evaluation of your results.>

Zusammenfassung

<Hier sollten Sie eine kurze, aussagekräftige Zusammenfassung (ca. eine Seite) Ihrer Arbeit geben, welche das Thema der Arbeit, die wichtigsten Inhalte, die Arbeitsergebnisse und die Bewertung der Ergebnisse umfasst.>

Contents

List of Figures

List of Tables

Vorwort

Allgemeine Hinweise zur Erstellung einer Abschlussarbeit

- Beachten Sie, dass diese Vorlage für einen zweiseitigen Ausdruck angelegt wurde.
- Über die Papierqualität können Sie entscheiden, aber wir empfehlen aber Seiten mit wichtigen, farbigen Grafiken auch in Farbe auszudrucken und dabei ein höherwertiges Papier zu verwenden.
- Bitte stimmen Sie mit dem Betreuer Ihrer Arbeit auch den Zweitgutachter ab. Die Anfrage des Zweitgutachters erfolgt von Ihnen. Es ist an dieser Stelle sinnvoll, die Anfrage mit einer kurzen Zusammenfassung der Arbeit zu stellen.
- Bitte beachten Sie, dass Sie Ihre Abschlussarbeit mit einer Klebebindung versehen, eine Ringbindung ist nicht erwünscht.

1 Introduction

1.1 Project IKON

This thesis has a direct application in a project which tries to explore potentials for knowledge transfer activities at a research museum. Project *IKON* was started in cooperation with the German Natural History Museum in Berlin which houses more than 600 [**TK:** Right number?] scientists, PhD students and other staff. With that size of scientific staff the institution is a global player in research on evolution and biodiversity [Int]. Despite its importance in the research landscape, the museum is challenged with a lack of shared knowledge across working groups and organizational structures such as departments. In interviews researchers from the project were able to trace these problems back to the very intricate and complex layout of rooms and halls in the building which was originally constructed in 1810. In order to mitigate this problem Figure 1.1 shows one of the main deliverables of *IKON* - a ML-driven data visualization which follows the path of knowledge at this research museum from its creation in projects over knowledge transfer activities, where multiple projects exchange their findings and try to generate added value for each other, to the final target group. Knowledge transfer is made explicitly visible by showing projects not in the predefined taxonomy of the museum, but instead in semantic relation to each other. This is accomplished by running all project abstracts through a topic extraction pipeline consisting of four major components, as seen in Figure 1.2, which will be discussed in detail in chapter 3.

First user tests and interviews unveiled that, even though the visualization was specifically tailored to non-technical users [**TK:** neeeds definition], the scientists from the museum had a hard time interpreting and understanding the output generated by the pipeline. Furthermore each component in Figure 1.2 introduces additional parameters which influence the results generated by the pipeline.

In order to lay the groundwork for this thesis and understand the challenges which scientists face while interacting with the visualization I carried out a workshop with the researchers from project *IKON*. In the beginning I asked them which kind of hardships they, based on their past experiences and interviews, observed during the interaction between user and visualization. Followed by an explanation of Figure 1.2 we discussed how these challenges may correlate with goals and questions. Following a description of the key questions each question was categorized according to the pipeline step, , as seen in Table 1.1, which may contribute information in order to support the user in answering his question.

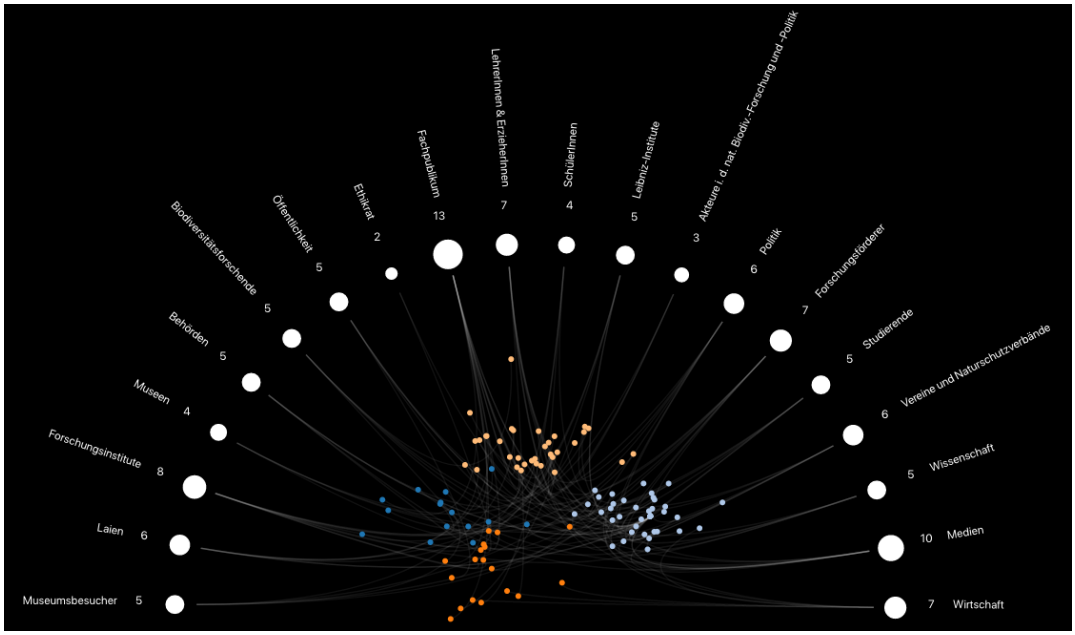

Figure 1.1: Screenshot of the cluster view of the IKON visualization

Question	Applicable pipeline component
How does the research landscape look like and on what kind of topics are prominent?	Topic Extraction
What does a cluster mean?	Clustering
What does the distance between clusters/projects mean?	Topic Extraction / Reduction into 2D
How similar are two projects/clusters?	Topic Extraction

Table 1.1: Table showing the sourced questions and the pipeline step which could provide an answer

1.2 Interpretability

With the surge of the application of machine learning (ML) systems in our daily life there is an increasing demand to make operation and results of these systems interpretable for people with different backgrounds (ML experts, non-technical experts etc.). Contrary to these efforts, interpretability as term has become an ill-defined objective [Lip16] for research and development in ML algorithms since there is no widely agreed upon definition of it. This leads to a very fragmented nature of the field.

Miller et al. [MHS17] support this point by conducting a literature study and uncovering that interpretability research is rarely influenced by insights from the humanities, especially connected fields as explainability or causality research.

[**TK:** Interpretation of machine learning (ML) results is a major challenge for

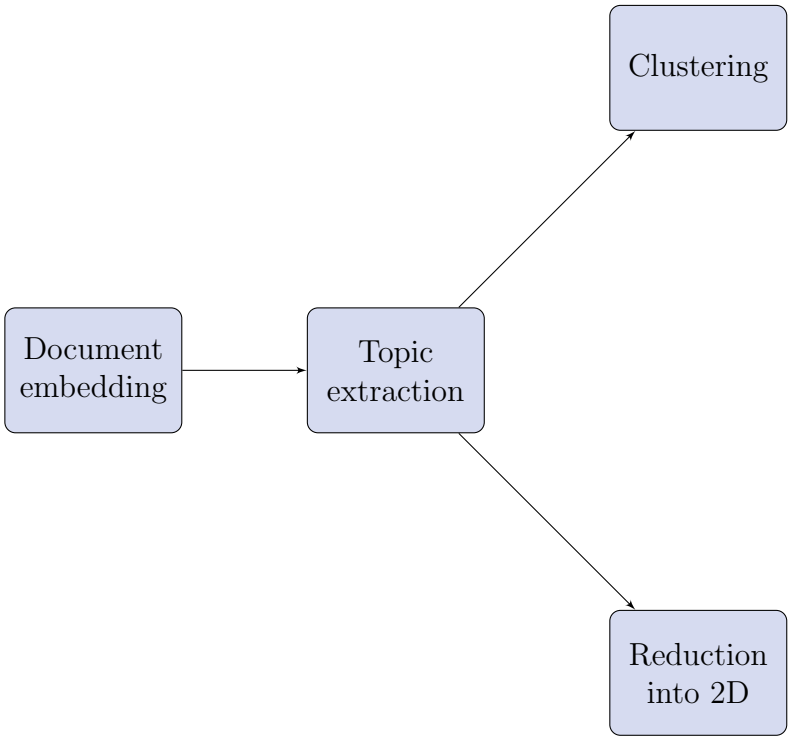

Figure 1.2: Components of a general unsupervised topic extraction pipeline

humans, especially for non-technical experts [ref]. Research on interpretability[1] in the ML community has focused on developing interpretability techniques, i.e. specific technical approaches to generate explanations[2] for ML results. However, applications of these techniques are predominantly concerned with making particular model features understandable, rather than supporting the interpretation of ML-driven systems in a specific context of use. At the same time, research in the HCI domain often remains on a formal, algorithmic level—explanations tend to be technical and tailored to an expert audience, mirroring the technical focus of ML research. Realistic use cases and qualitative, context-aware evaluations to inform the selection and design of interpretability techniques remain rare. While we do not see complete transparency as a prerequisite for interpretability we hypothesize that in general, since interpretation is dependent on context, interpretability techniques cannot be fully context agnostic either. Therefore, our general approach is to research interpretability from a context-aware perspective, i.e. we explore how interpretability can be operationalized in a specified, well-defined domain context.]

[1] Which we position to be a high-level precondition for Explainability from the XAI [**?**] and Fairness, Accountability and Transparency, from the FAT-ML discourse [**?**].

[2] Which we define as instances of interpretability techniques.

2 Literature mapping study

2.1 Motivation

In order to access current methods in the fast-moving field of interpretability research in machine learning in a reproducible and structured fashion I will conduct a literature mapping study according to Petersen et. al [PFMM], which consists of a number of sequential steps which should result in a representative corpus and an analysis using it.

2.2 Methodology

The recommended process is augmented by further steps in order to tailor it to the existing use case and consists of the following seven procedures:

1. Definition of research questions:

   The overall process starts by defining clear questions which should guide the development of the whole literature mapping study and subsequently the result as well. Since I am interested in gaining an overview over the existing interpretability techniques, I chose the following questions:

   a) What kind of explainability techniques are mentioned in the corpus?

   b) What kind of models are enhanced by explainability techniques?

   c) Which techniques are applicable to results produced by the pipeline or the pipeline itself?

2. Construction of a search string:

   Based on the questions one is able to gather a set of key words which are most relevant to the field which is analyzed. Each word is augmented by synonyms which are concatenated with boolean OR operators and several of these synonymous groups are again connected via logical ANDs. Applying this method to the previously found questions yields the following search string:

   ("explainability" OR "explainable" OR "explanation" OR "explaining" OR "interpretability" OR "interpretable" OR "interpretation" OR "interpret" OR "understanding") AND ("machine learning" OR "neural network" OR "neural networks" OR "AI" OR "XAI" OR "artificial intelligence" OR "model") AND ("text" OR "document" OR "NLP" OR "natural language programming" OR "review" OR "method" OR "technique" OR "visualization")

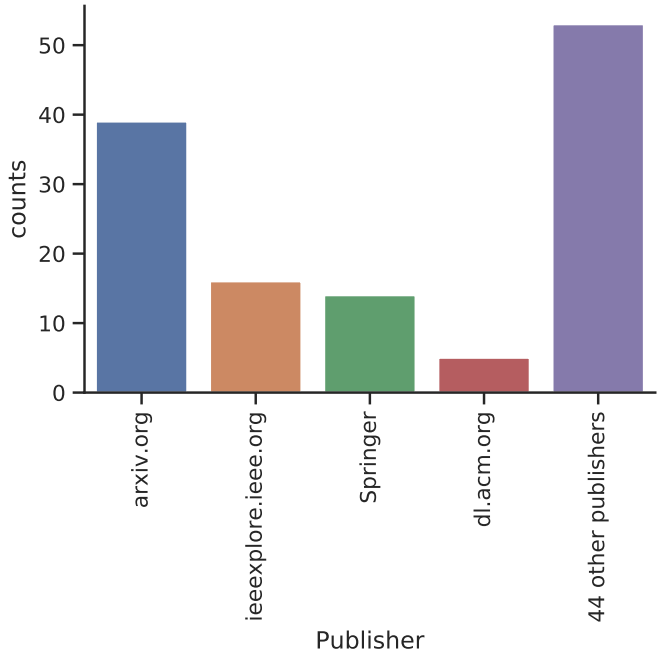

Figure 2.1: Barplot displaying the distribution of publishers occurring in the meta search results

3. Analysis of the main publishers using a meta search and the search string:

   Due to the presumed distributed nature of interpretability research it is not easy to pinpoint the main publishers of scientific articles. In order to mitigate this, a pre-search in the meta-search engine 'Google Scholar' is conducted. It should be noted at this point that any biases which are apparent in the meta search engine therefore apply to this analysis as well. One can see in Figure 2.1 that the main publishers are respectivly Arxiv, IEEE, Springer and ACM. Since all of these publishers are mainly focused on publications in computer science, mathematics and engineering, this speaks in favor of the hypothesis that most of the research is still very technical and research from social sciences rarely influences it. Even though Arxiv is not a credible publisher per se, it seems like the research community uses it as the first place to publish work and therefore it should not be excluded in this analysis.

4. Sourcing of publications in scientific databases:

   Based on the insights from the previous step each of the main publisher's databases is scraped using the search string and their respective 'advanced search' interfaces or their APIs. Since most searches result in more than 1000 publications only the top 100 results ordered by the relevance scoring of the database are taken into account. These publications then form the corpus which is the basis for further analysis.

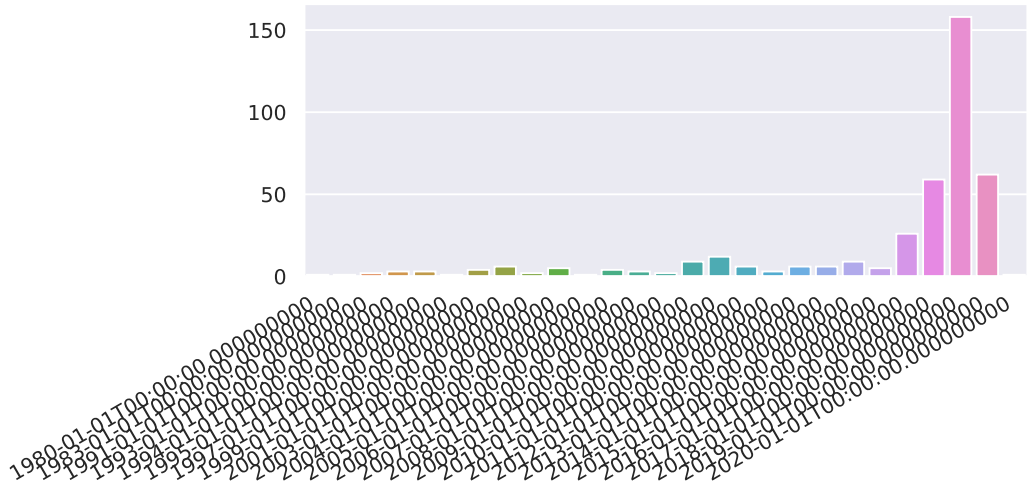

Figure 2.2: Barplot displaying the distribution of publishers occurring in the meta search results

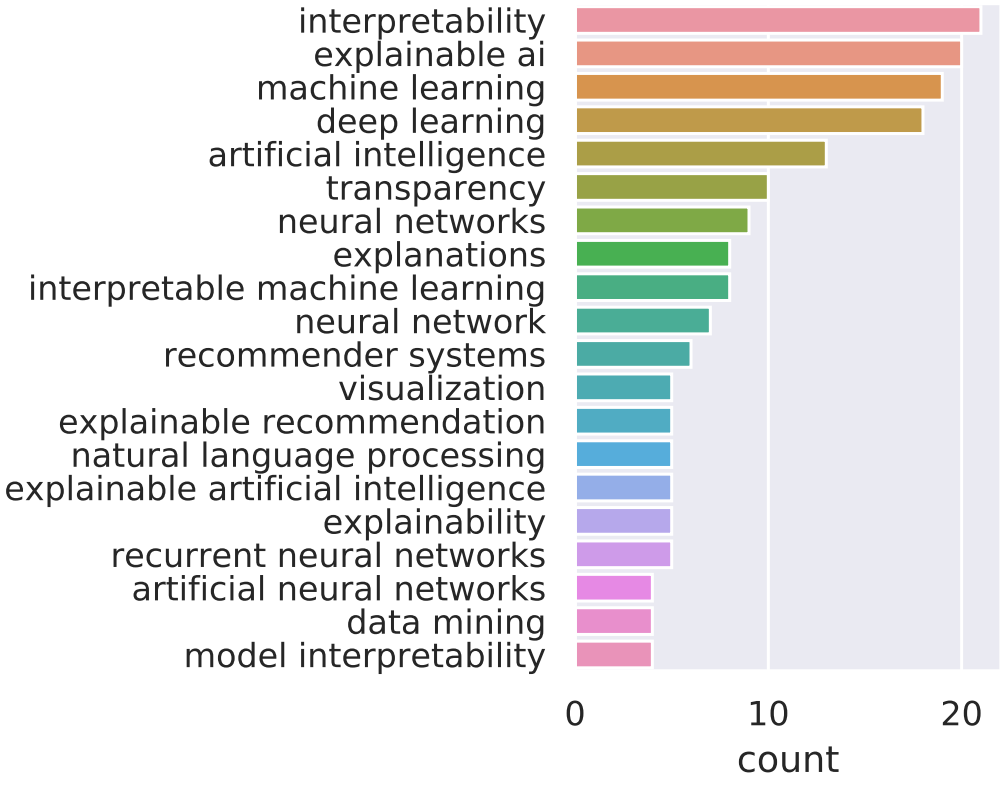

Figure 2.3: List of the 20 most used tags and their absolute frequency

Inclusion criteria	Exclusion criteria
• Reviews the current state of explainability research • Presents a specific method for enhancing explainability for models	• Is not scientific literature • Does not describe the used explainability method • The publication does not focus on explainability • The described method is neither general, nor focused on NLP

Table 2.1: Table showing all used inclusion and exclusion criteria

5. Filtering of these publications by keywording their abstracts:

   Most scientific databases do a full text search on publications and possibly find the supplied keywords from the search string in sections of the paper which are not relevant e.g. the bibliography or in the outlook. Therefore another filtering step is necessary which searches for the search string in the abstracts of the papers of the corpus.

   In order to enhance the quality of the filtering process, the search string is enhances with key words generated by an analysis of the current corpus. As seen in Figure 2.3 the previous search string already contains most of the relevant keywords.

6. Definition and application of inclusion and exclusion criteria to narrow down the pool of publications further:

   The next step serves as another filtering step enhancing the quality of the hitherto automatic selection by using human decision making. A combination of the guiding questions, which were defined in the beginning of the process and a first pass over the whole corpus, in which I skimmed the papers, gave me a clear set of criteria, as seen in Table 2.1, which can be used to filter the corpus further. In a second pass each paper was evaluated and included in the next step if and only if it satisfied at least one inclusion criterion and none of the exclusion criteria.

7. Quantitative assessment of the resulting corpus:

   In the last step the actual mapping is generated. In another pass I first skimmed and then read each paper and based on that classified each

publication and its presented technique according to the Gamuth classification, the type of model to which the technique is applicable and the component where the technique could be applied in the topic extraction pipeline. Additionally each paper was classified as either "Theory", "Method", "Study" or "Report". A "Method" paper presents a single explainability technique and demonstrates its impact in an exemplary use case. A "Theory" paper does so as well, but misses a presented application and evaluation. A "Report" on the other hand summarizes and presents multiple techniques. Finally, a "Study" paper shows the results of an evaluation of an interface which visualizes the results of explainability methods. Publications from the last category are therefore less technical and more concerned with the HCI aspects of explainability techniques and their visualization.

Since most of the overview papers presented a huge amount of techniques which were already covered by the "Method" papers and the corpus was already large, I decided to exclude them from the last mapping step.

2.3 Results

In order to answer my first question concerning the different kinds of researched explainability

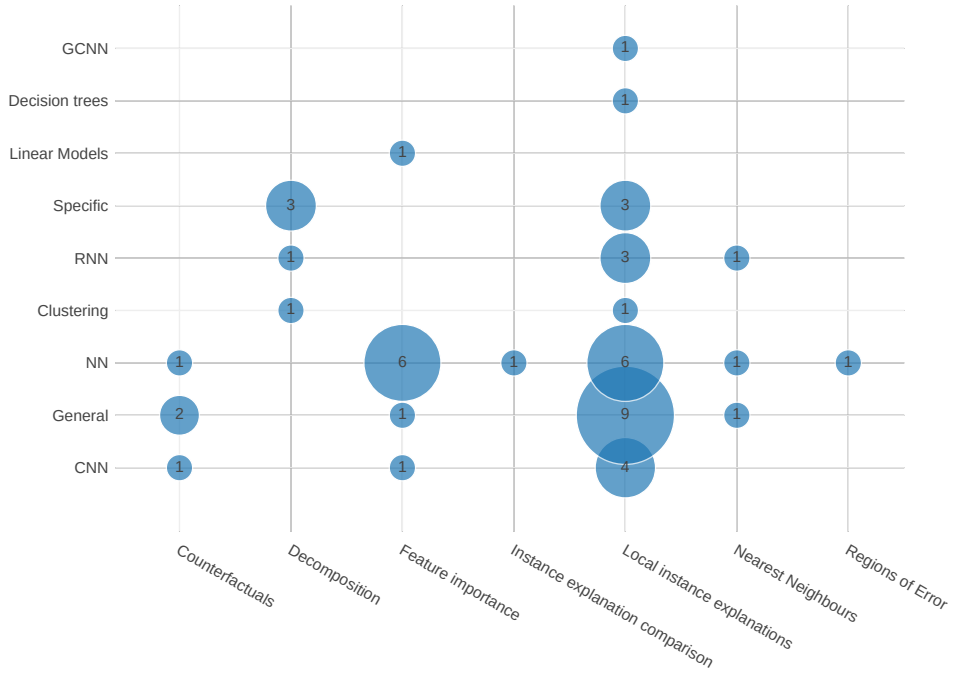

Figure 2.4: Mapping of applicability and gamut classification

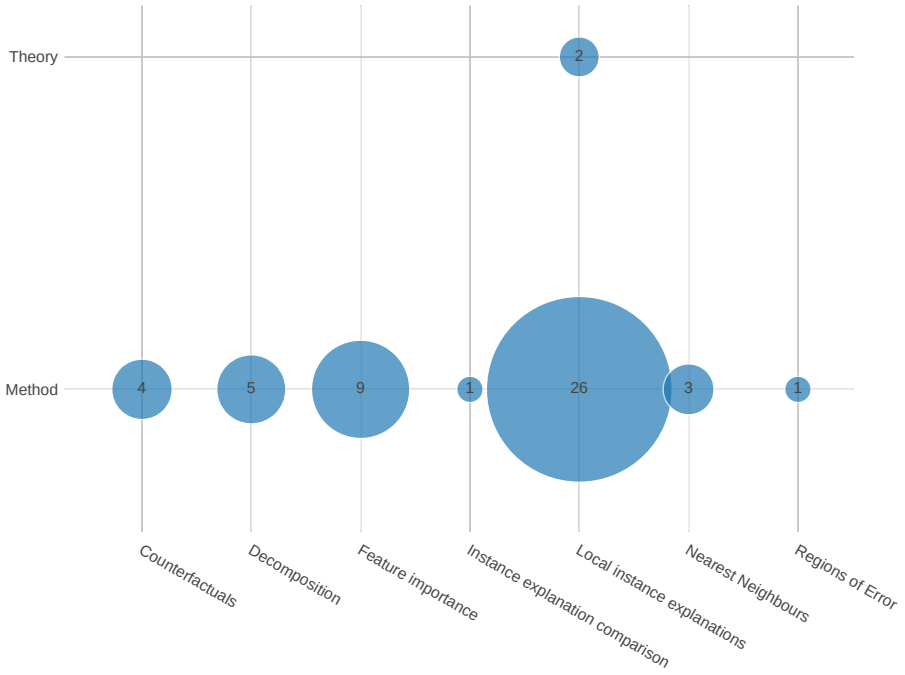

Figure 2.5: Mapping of applicability and gamut classification

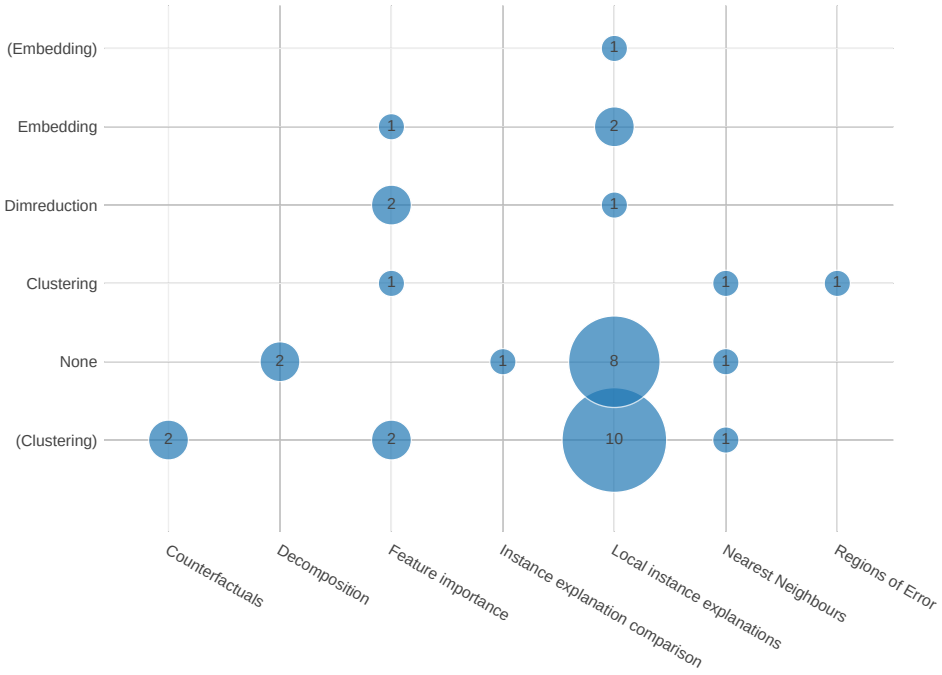

Figure 2.6: Mapping of applicability and gamut classification

3 Implementation

3.1 Data and Preprocessing

In order to connect projects semantically instead of by the rigid taxonomy of the museum, I was able to use the project's self-description which is recorded in the GEPRIS database of the DFG [DFG]. It consists of almost all projects which were supported by the DFG since 2000. Fortunately, another bachelor project before me worked on a scraper which extracted approximately 114.000 projects from the web interface of the database since there is no publicly available API. Each project was characterized by a title, a project abstract in German or English, start and end dates as well as additional meta data like connected institutions or people working in the project.

As one can see in Figure 3.1, there is a peak at word count 3 and one at approximatly 100. The first one corresponds to all projects which do not have descriptions, because they are described with "Keine Zusammenfassung vorhanden". The latter peak on the other hand is produced by projects from a fund which uses the same descriptions for all its projects which are financed through the DFG.

Removing these peaks in Figure 3.2 reveals that most texts have an length of 180 words, while also having smaller peaks at ca. 80 and 360 words. The shortest description has a length of one word and the longest approximately 960 words.

Following the advice of Matthew et al. [Den17] the texts were preprocessed by a P-N-S-W scheme. First punctuation (P) and numbers (N) were removed since sentence boundaries or specific numbers do not bear a lot of information in middle-sized descriptive texts. Following this, according to the categories of Matthew et al., a stemming step (S) is performed, which uses lemmatization to find the lemmas of words by using vocabularies and the context of each word. The last step removes infrequent words without much semantic meaning, commonly known as stopwords (W). Lowercasing and n-gram inclusion were omitted, because casing is an important feature for distinguishing nouns from other word types in the German language, which helps the lemmatization step, and the use of word composition makes most reasonable n-grams in other languages appear as one word in German.

Until the start of this thesis the pipeline did all this preprocessing using regex-based rules and a lemmatization using the SpaCy lemmatizer. This proved to be a viable option until a corpus size of 5000 since after that point the running time was too long to effectively work with it. Therefore I bundled all the preprocessing operations in a new class called *Datapreprocessor*, which should be able to transform any given query into a preprocessed dataset for the

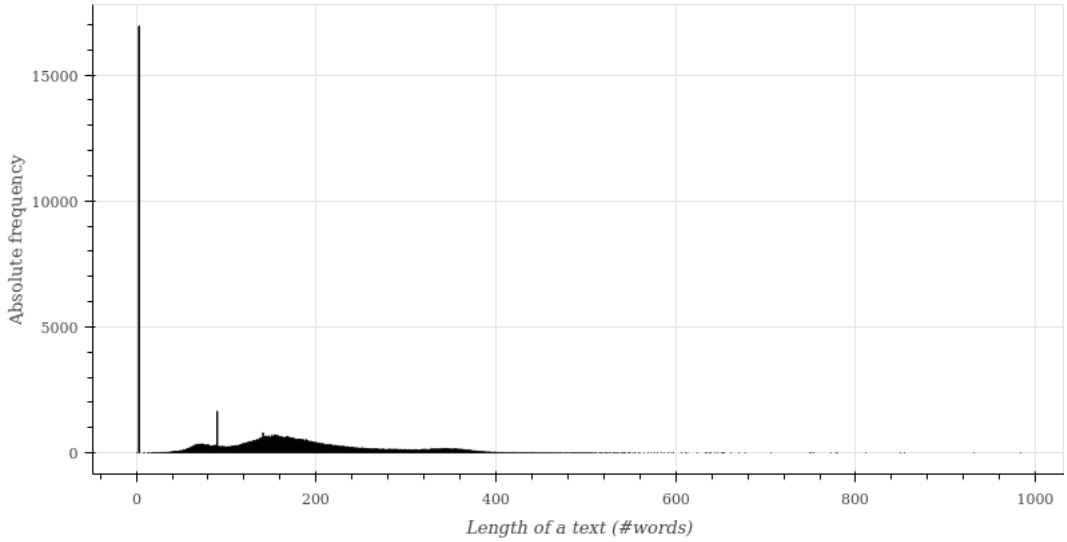

Figure 3.1: Histogram showing the distribution of text lengths in the dataset

following pipeline steps as well as cache its results. In order to do that I rewrote the preprocessing steps and integrated them into the already existing SpaCy pipeline which uses a CNN to apply the previously discussed preprocessing. Additionally it is able to detect the language of a text, which, in turn, makes it possible to filter out all non-German texts. Using this existing framework gave me the opportunity to embed my custom code into the Cython code of the framework accelerating the looping over the corpus. Additionally I was able to fully parallelize the process on n CPUs by splitting the corpus in n chunks and feeding each chunk into a separate sub-process to make use of the batch sizes of the SpaCy neural networks. This accelerated the preprocessing by a factor of 10.

3.2 The existing pipeline

The existing pipeline was implemented by me as a proof-of-concept for project *IKON*. Following the structure of Figure 1.2 the first step is a document vectorization of the given texts in order to embed them in one common vector space. One of the simplest and still effective methods is a Tf-Idf Bag-Of-Words (TfIdf-BOW) embedding. With this procedure each text is represented as a set of terms, the bag of words. Having a whole corpus it is now possible to assign a vector to each document D in corpus $C = \{D_1, ..., D_n\}$ of length $N = |C|$, where each entry i is the number of term occurrences of term t_i in D. That means that each document gets embedded into a vector space of dimensionality |(unique terms in C)| and the corpus becomes a matrix of size $|(\text{unique terms in C})| \times N$. In order to additionally introduce information from the whole corpus into each vectorized document and therefore contextualize it, each entry is replaced by $C_{t,d} = Tf(C_{t,d}) \cdot Idf(C, t, d)$ where $Tf(t, d)$ is

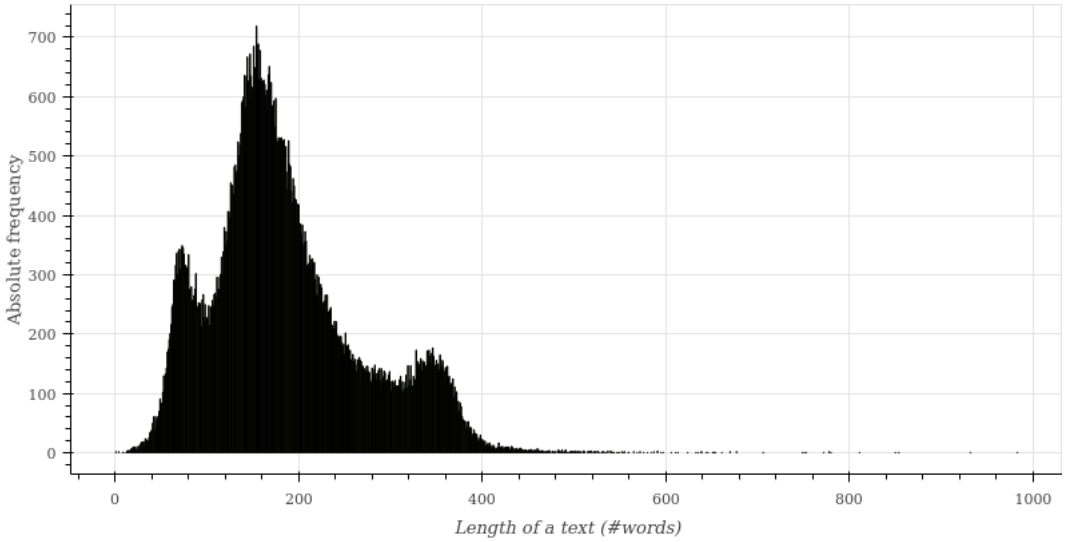

Figure 3.2: Histogram showing the distribution of text lengths in the dataset excluding duplicates and projects without a description

often the identity function and $Idf(C, t, d)$ is $\log \frac{N}{|\{D \in C : t_t \in D\}|}$. [Piv] The notion behind this is intuitive. The higher the term frequency of a term in a document, the more important it is for this specific document and the more a term appears in several documents, the less it caries information to seperate a document from others. [**TK:** Needs maybe rework based on Shannon theory] This ensures that words which are specific to a small group of documents and appear often in them, get a higher weight, while terms which are infrequent or too frequent in many documents, as articles for example, get a small weight.

Now that we have a vector representation of each document, we could work in the existing space and try to cluster our documents in their current form using k-Means, which will be explained later. An exemplary analysis shows that the semantic coherence of the document clusters seems to lack. [**TK:** show proof] That is due to the clustering algorithm failing to perform and facing, what is commonly known as, *the curse of dimensionality*. The curse of dimensionality states for distance based methods that "under certain reasonable assumptions on the data distribution, the ratio of the distances of the nearest and farthest neighbors to a given target in high dimensional space is almost 1 for a wide variety of data distributions and distance functions" [AHK01]. Therefore closeness between points, which is the relevance measure for the k-Means algorithm due to it using the Euclidian distance, becomes effectively meaningless and making it necessary to reduce the dimensionality of the vector space.

One popular method, which is often used in conjunction with Tf-Idf BOW embeddings, is the Latent Semantic Indexing (LSI), also known and henceforth referenced as Latent Semantic Analysis (LSA). A LSA operates on the premise that a vectorized corpus contains latent structures, which my correspond to

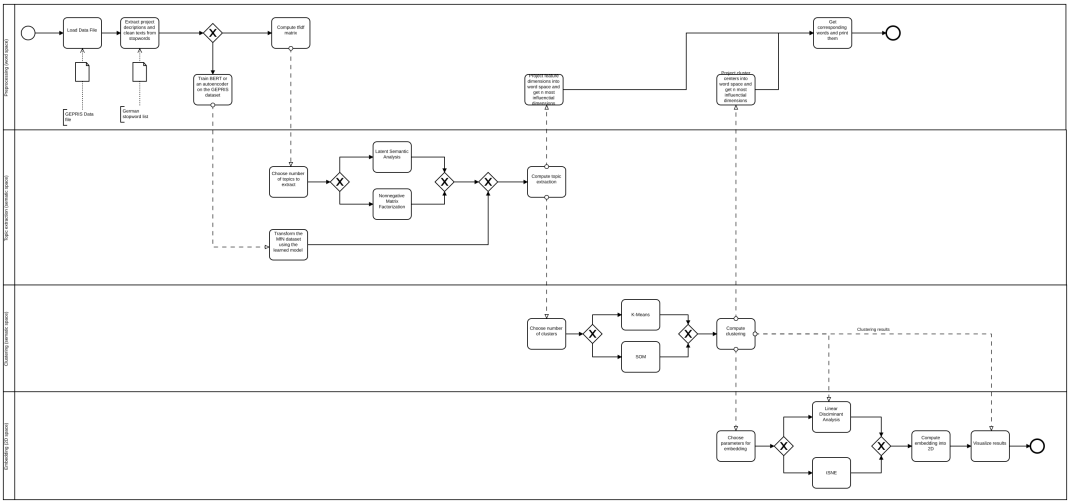

Figure 3.3: BPMN process diagram of the existing topic extraction pipeline

topics for example. Such a topic would consist of several words which are semantically connected and therefore appear together more often than words which are not semantically similar. Adding constraints such as adjustable representational richness, which depicts sufficient parameterisation, explicit representation of both terms and documents and computational tractability for large datasets the authors decided to use a Singular Value Decomposition (SVD) [DDF+]. The SVD is closely related to Principal Component Analysis (PCA) and reduces the dimensionality of a dataset by removing the dimensions with the least variance, effectively projecting the vector space onto the subspace with the highest variance and therefore the most information contained. Applying a SVD on the corpus changes the representation of the document from being a linear combination of words into being a linear combination of latent topics.

This representation is now usable for most other methods such as clustering due to its smaller dimensionality. Now a k-Means algorithm is applied to discover clusters and group the documents.

In order to visualize the high dimensional topic space in 2D a linear discriminant analysis is used using the clustering as labels.

Technique	Parameters	Maximum processable text length	Type
Tf-Idf BOW	5	6	
Doc2Vec	8	9	

Table 3.1: Table summarizing the key features of different document embedding techniques

3.3 Document embedding

3.3.1 A short survey of document embedding techniques

3.4 Topic extraction

3.5 Clustering

3.6 Reduction into 2D

4 Validation

4.1 Setup

4.2 Cognitive Walkthrough

5 Conclusion

5.1 Discussion

5.2 Outlook

- Die Zusammenfassung sollte das Ziel der Arbeit und die zentralen Ergebnisse beschreiben. Des Weiteren sollten auch bestehende Probleme bei der Arbeit aufgezählt werden und Vorschläge herausgearbeitet werden, die helfen, diese Probleme zukünftig zu umgehen. Mögliche Erweiterungen für die umgesetzte Anwendung sollten hier auch beschrieben werden.

Bibliography

[AHK01] Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the Surprising Behavior of Distance Metrics in High Dimensional Space. In Gerhard Goos, Juris Hartmanis, Jan van Leeuwen, Jan Van den Bussche, and Victor Vianu, editors, *Database Theory — ICDT 2001*, volume 1973, pages 420–434. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001.

[DDF+] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, page 17.

[Den17] Arthur (New York University) Denny, Matthew (Penn State University); Spirling. Replication Data for: Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It, 2017.

[DFG] DFG - GEPRIS. https://gepris.dfg.de/gepris/OCTOPUS?task=showAbout.

[Int] Introducing the Museum für Naturkunde in Berlin. https://pro.europeana.eu/post/introducing-the-museum-fur-naturkunde-in-berlin.

[Lip16] Zachary C. Lipton. The Mythos of Model Interpretability. *arXiv:1606.03490 [cs, stat]*, June 2016.

[MHS17] Tim Miller, Piers Howe, and Liz Sonenberg. Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. *arXiv:1712.00547 [cs]*, December 2017.

[PFMM] Kai Petersen, Robert Feldt, Shahid Mujtaba, and Michael Mattsson. Systematic Mapping Studies in Software Engineering. page 10.

[Piv] Pivoted document length normalisation | RARE Technologies. https://rare-technologies.com/pivoted-document-length-normalisation/.