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 4, 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.

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. Based on these findings and a workshop with the main researchers of project *IKON* I was able to extract a set of questions that a subject tries to answer while interacting with the visualization:

1. How does the research landscape look like and on what kind of topics are prominent?

2. What does a cluster mean?

3. What does the distance between clusters/projects mean?

4. How similar are two projects/clusters?

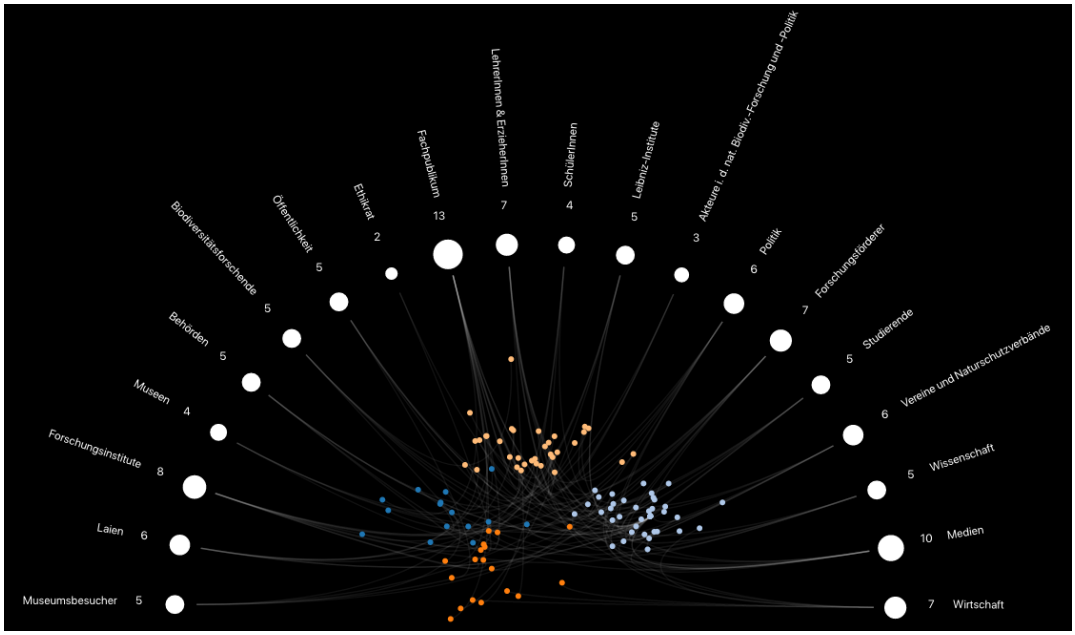

Figure 1.1: BPMN process diagram of the existing topic extraction pipeline

1.2 Interpretability

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

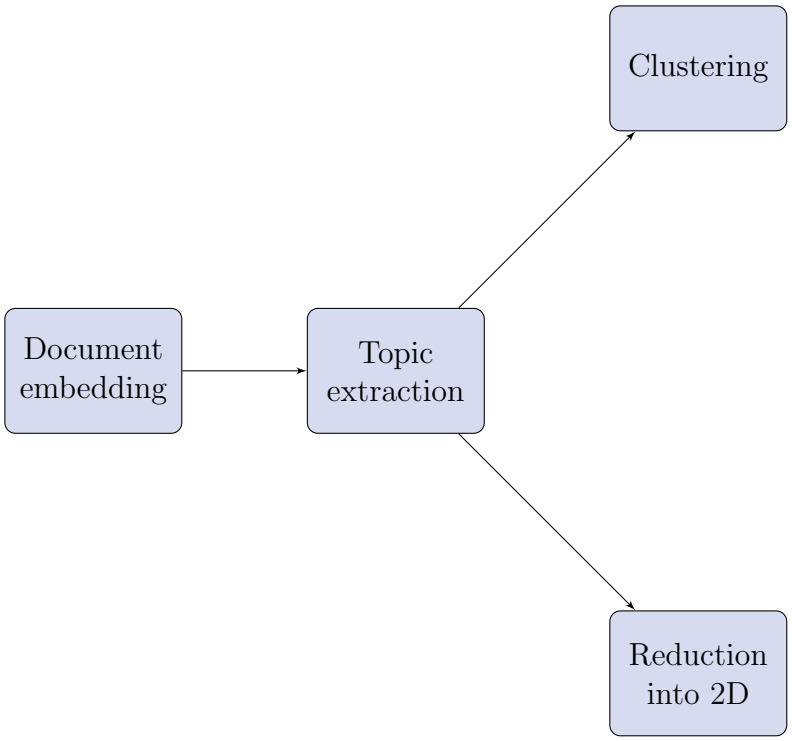

Figure 1.2: Components of a general unsupervised topic extraction pipeline

2 Literature mapping study

2.1 Motivation

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.

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) On 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")

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.

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.

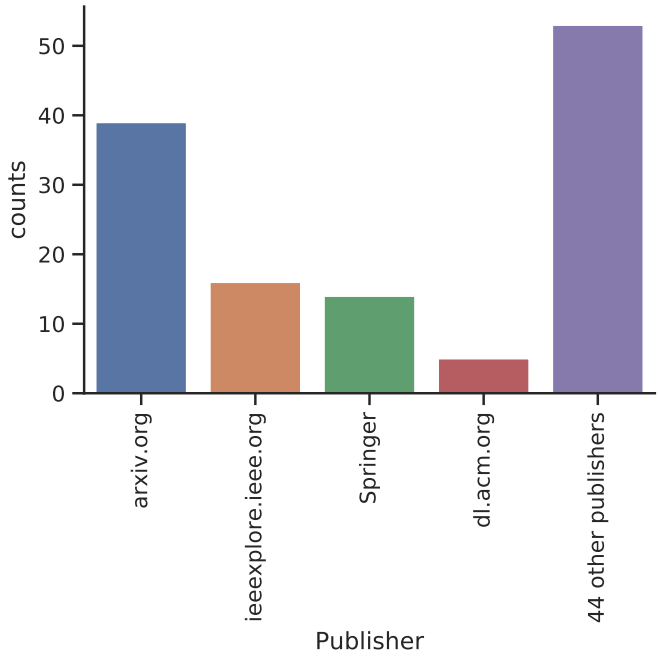

Figure 2.1: Barplot displaying the distribution of publishers occurring in the meta search results

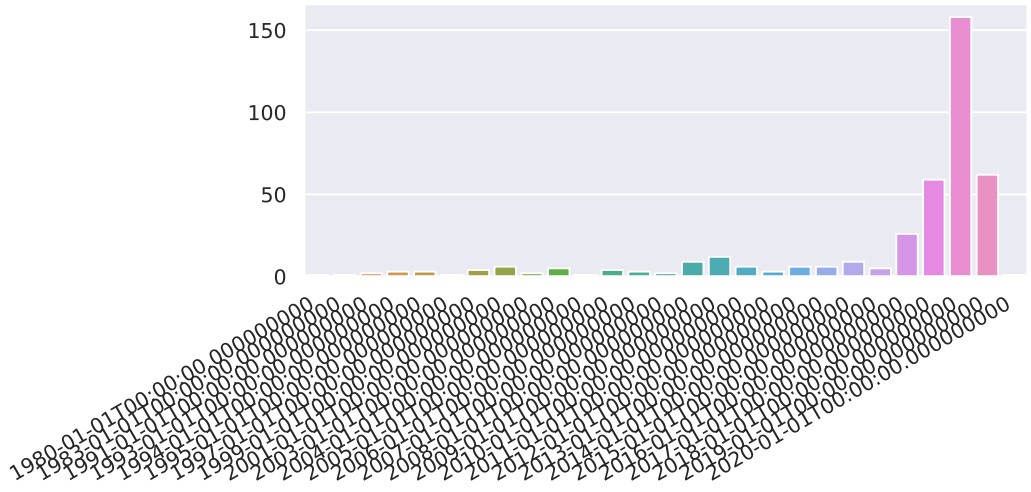

Figure 2.2: Barplot displaying the distribution of publishers occurring in the meta search results

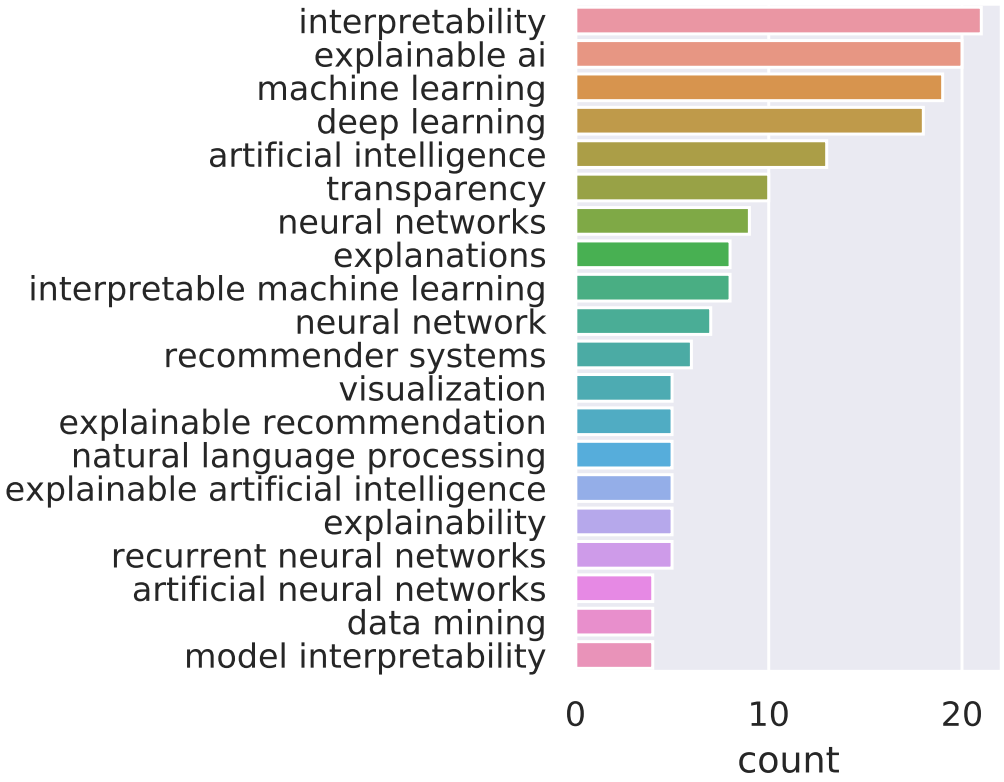

Figure 2.3: List of the 20 most used tags and their absolute frequency

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

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. 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.

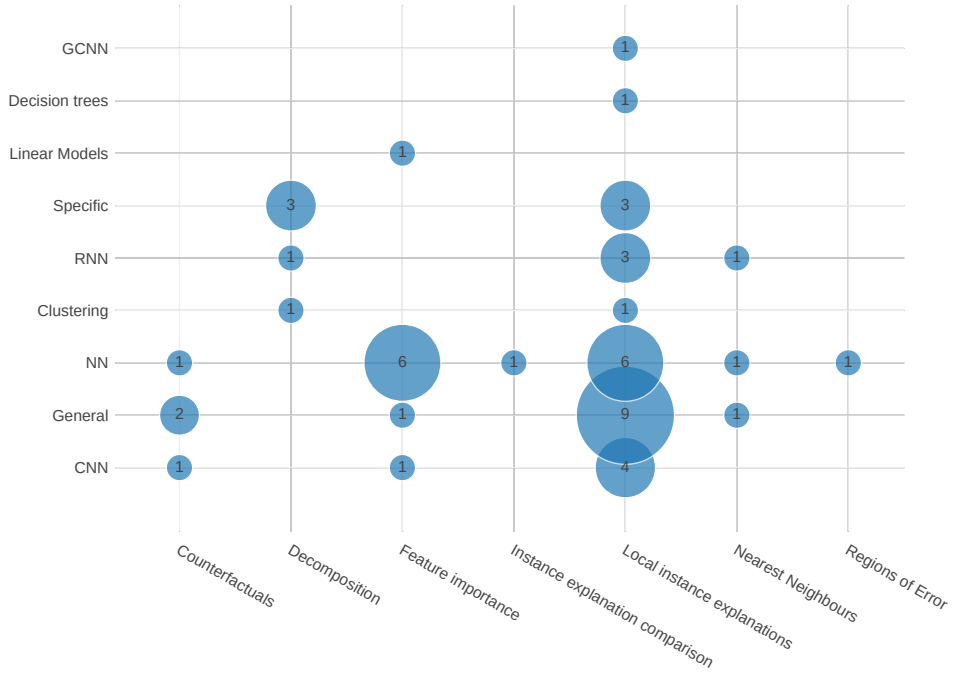

Figure 2.4: Mapping of applicability and gamut classification

2.3 Results

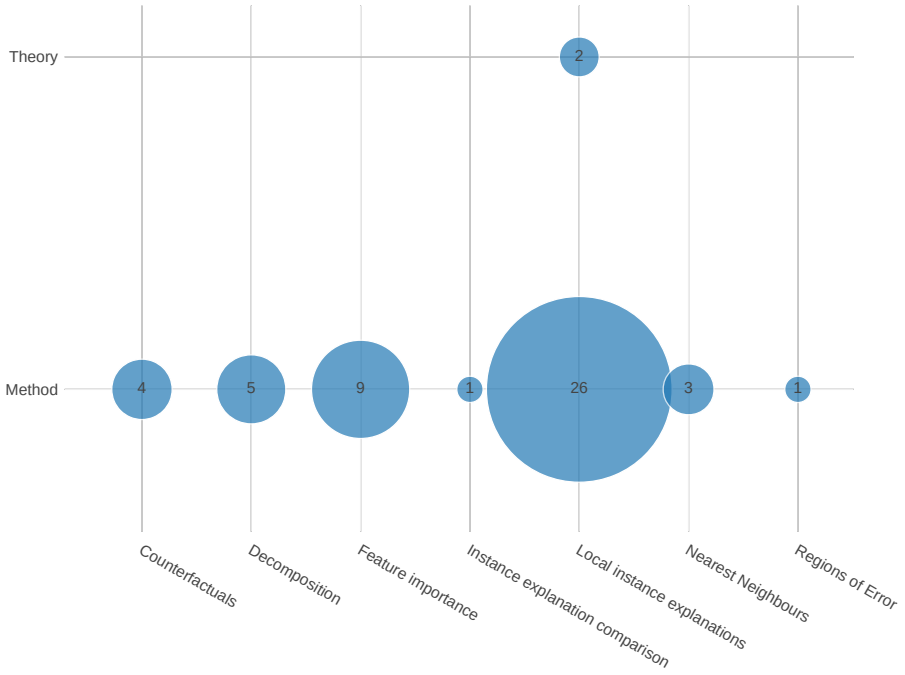

Figure 2.5: Mapping of applicability and gamut classification

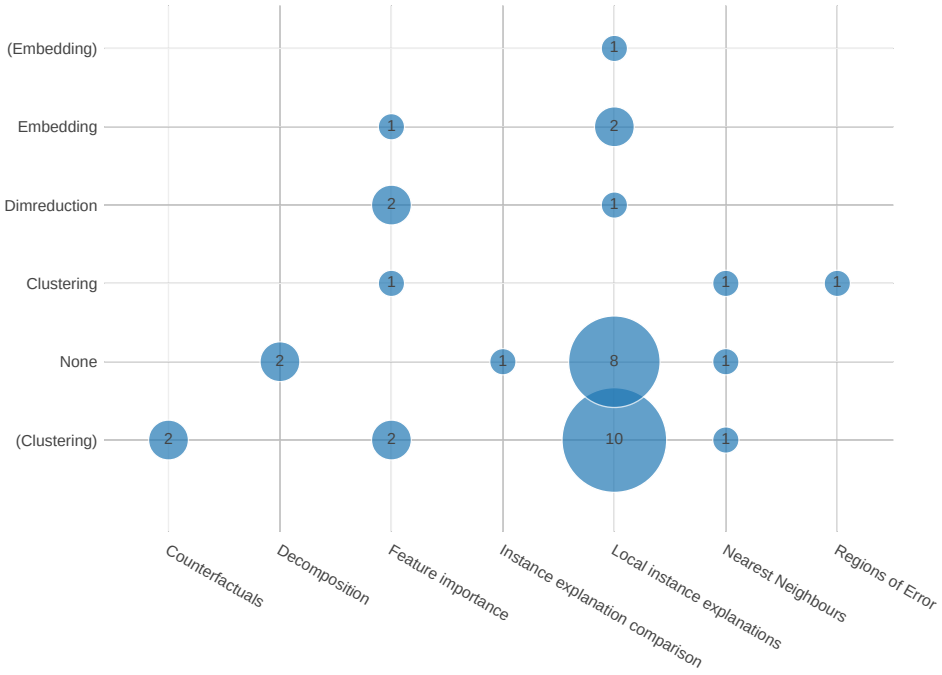

Figure 2.6: Mapping of applicability and gamut classification

3 Implementation

3.1 The existing pipeline

3.2 Preprocessing

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

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

[**TK:** Add diagram for IKON pipeline]

Figure 3.1: BPMN process diagram of the existing topic extraction pipeline

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

[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.