Bachelorarbeit am Institut für Informatik der Freien Universität Berlin

Human-Centered Computing (HCC)

Developing Interpretability Techniques for Unsupervised Topic Modeling

Tim Korjakow

Matrikelnummer: 372862

Email: tim.korjakow@campus.tu-berlin.de

Betreuer: Jesse Josua Benjamin

Erstgutachterin: Prof. Dr. C. Müller-Birn

Zweitgutachter: Prof. Dr. K.-R. Müller

Berlin, 08.08.2019

b

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 August 8, 2019

Tim Korjakow

Abstract

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.). A wide range of research exists, particular in ML research on specific interpretability techniques (e.g., extracting and displaying information from ML pipelines). However, often a background in machine learning or mathematics is required to interpret the results of the interpretability technique itself. Therefore there is an urgent lack of techniques which may help non-technical experts in using such systems.

The grounding hypothesis of this thesis is that, especially for non-technical experts, context is an influential factor in how people make sense of complex algorithmic systems. Therefore an interaction between a user and an application assumed to be an interplay between a user and his historical context, the context of the situation in which the interaction is embedded and the algorithmic system. Interpretability techniques are the common link which bring all these different aspects together.

In order to evaluate the assumption that most of the current interpretability research is tailored to a technical audience and gain an overview over existing interpretability techniques I conducted a literature mapping study studying the state of interpretability research in the field of natural language processing (NLP). The results of this analysis suggest that indeed most techniques are not evaluated in a context where a non-technical expert may use it and that even most publications lack a proper definition of interpretability.

I propose and implement three methods for making the topic modeling pipeline more interpretable, drawing inspiration from the general strategies which were used in the sourced publications from the mapping study.

Since this thesis presumes that interpretation is a complex socio-technical process, a validation also has to take both sides into account. A first system-centered evaluation using coherence scores showed that the pipeline is indeed able to semantically link projects and show important features in a meaningful way. This theoretical analysis was followed by a human-centered validation in which a cognitive walkthrough was used to simulate the interaction between a non-technical expert and the application. This usability technique unveiled that even though the three interpretability techniques where developed with a certain strategy in mind, the context can make it possible to reinterpret the output of interpretability techniques and use them in an unintended way to understand the system. This finding speaks in favor of the hypothesis that such context-light usability techniques can not completely probe the relationship between user and application, creating a need for validations which are embedded in the same context in that the user will face the system.

Zusammenfassung

Mit der steigenden Anwendung von Machine Learning-Systemen (ML) in unserem täglichen Leben steigt auch die Nachfrage, die Nutzung und die Ergebnisse dieser Systeme für Menschen mit unterschiedlichem Hintergrund (ML-Experten, nicht-technische Experten usw.) interpretierbar zu machen. Es gibt einen breiten Fundus an Forschung, insbesondere in der ML-Forschung, zu spezifischen Interpretationstechniken (z.B. Extraktion und Darstellung von Informationen aus ML-Pipelines). Häufig ist jedoch ein Bildungshintergrund im Bereich des maschinellen Lernens oder der Mathematik erforderlich, um die Ergebnisse der Interpretierbarkeitstechnik selbst zu interpretieren. Daher fehlt es dringend an Techniken, die nichttechnischen Experten bei der Nutzung solcher Systeme helfen können.

Die grundlegende Hypothese dieser Arbeit ist, dass, insbesondere für nichttechnische Experten, der Kontext einen großen Einfluss darauf hat wie Menschen komplexe algorithmische Systeme verstehen. Daher ist eine Interaktion zwischen einem Nutzer und einer Anwendung in diesem Modell tatsächlich ein Zusammenspiel zwischen einem Benutzer und seinem historischen Kontext, dem Kontext der Situation, in die die Interaktion eingebettet ist, und dem algorithmischen System. Interpretationstechniken sind das gemeinsame Bindeglied, welches all diese verschiedenen Aspekte zusammenführt.

Um die Annahme zu evaluieren, dass der Großteil der aktuellen Interpretierbarkeitsforschung auf ein technisches Publikum zugeschnitten ist und einen Überblick über bestehende Interpretierbarkeitstechniken zu erhalten, habe ich eine systematische Literaturübersichtsstudie durchgeführt, die den Stand der Interpretierbarkeitsforschung im Bereich der natürlichen Sprachverarbeitung (NLP) untersucht. Die Ergebnisse dieser Analyse deuten darauf hin, dass die meisten Techniken in der Tat nicht in einem Kontext bewertet werden, in dem ein nicht-technischer Experte sie verwenden kann, und dass die meisten Publikationen keine angemessene Definition der Interpretierbarkeit liefern.

Daher präsentiere und implementiere ich drei Methoden, um die Topic Modeling-Pipeline besser interpretierbar zu machen, indem ich mich von den allgemeinen Strategien inspirieren lasse, die in den, aus der systematischen Literaturübersichtsstudie stammenden, Publikationen verwendet wurden.

Da diese Arbeit davon ausgeht, dass die Interpretation ein komplexer soziotechnischer Prozess ist, muss auch eine Validierung beide Seiten berücksichtigen. Eine erste systemzentrierte Bewertung mit Coherence Scores ergab, dass die Pipeline in der Tat in der Lage ist, Projekte semantisch zu verknüpfen und wichtige Merkmale aussagekräftig darzustellen. Dieser theoretischen Analyse folgte eine humanzentrierte Validierung, bei der ein Cognitive Walkthrough verwendet wurde, um die Interaktion zwischen einem nicht-technischen Experten und der Anwendung zu simulieren. Diese Usability-Technik enthüllte, dass, obwohl die drei Interpretierbarkeitstechniken mit einer bestimmten Strategie im Hinterkopf entwickelt wurden, der Kontext es ermöglicht, die Ergebnisse von Interpretierbarkeitstechniken neu zu interpretieren und in einer neuen

Weise zur Interpretierung des Systems zu verwenden. Dieses Ergebnis spricht für die Hypothese, dass solche kontext-seichten Usability-Techniken die Beziehung zwischen Benutzer und Anwendung nicht vollständig untersuchen können, was einen Bedarf an Validierungsmethoden schafft, die in den gleichen Kontext eingebettet sind, die der Nutzer hätte.

Contents

1 Thematic Introduction and Motivation **1**
- 1.1 Project IKON . . . 1
- 1.2 Topic modeling . . . 1
- 1.3 Interpretability . . . 4
- 1.4 Working plan . . . 5

2 Literature mapping study **7**
- 2.1 Motivation . . . 7
- 2.2 Method . . . 7
- 2.3 Results . . . 12

3 Implementation of the Topic Modeling Pipeline **17**
- 3.1 General setup . . . 17
- 3.2 Data and Preprocessing . . . 17
- 3.3 The existing pipeline . . . 19
- 3.4 Components . . . 20
  - 3.4.1 Document embedding . . . 21
  - 3.4.2 Topic extraction . . . 22
  - 3.4.3 Clustering . . . 24
  - 3.4.4 Visualization . . . 25

4 Implementation of Interpretability Techniques **27**
- 4.1 Techniques . . . 27
  - 4.1.1 Top words . . . 27
  - 4.1.2 Cluster topography . . . 28
  - 4.1.3 Linearization . . . 30
- 4.2 Validation . . . 32
  - 4.2.1 System-Centered . . . 32
  - 4.2.2 Human-Centered . . . 35

5 Conclusion **39**
- 5.1 Outlook . . . 40

Appendix **43**
- 5.1.1 Protocol of the cognitive walkthrough . . . 43

Literatur54

List of Figures

1.1 Screenshot of the cluster view of the IKON visualization 2
1.2 Components of a general topic extraction pipeline 2
1.3 Proposed model by the researchers showing the interplay of explanation strategies, interpretability techniques and explanations 5

2.1 Barplot displaying the distribution of publishers occurring in the meta search results 8
2.2 Barplot displaying the distribution of publishing dates occurring in the results after the inclusion-exclusion step 10
2.3 List of the 20 most used tags and their absolute frequency . . . 11
2.4 Mapping of the type of publication and its Gamut+ classification 13
2.5 Mapping of applicability and Gamut+ classification 14
2.6 Mapping of pipeline step and Gamut+ classification 15

3.1 Histogram showing the distribution of text lengths in the dataset 18
3.2 Histogram showing the distribution of text lengths in the dataset excluding duplicates and projects without a description 18
3.3 Visualization of a training step of a Doc2Vec network [WYX+18] 22
3.4 Graph showing the training and validation loss of the autoencoder over progressing epochs 23
3.5 Boxplot showing the distribution of the quotients between WMD and EuD for all documents 24
3.6 Screenshot showing the exemplary interface 26

4.1 Screenshot showing the interpolated topography on an exemplary plot 31
4.2 Screenshot showing the linearization on an exemplary plot . . . 32
4.3 Graph showing the quality of the topic modeling while varying the embedding, topic extraction and clustering model 33
4.4 Plot for the parameter and models with the best coherence score 34

5.2 Cognitive Walkthrough step 1 43
5.3 Cognitive Walkthrough step 2 44
5.4 Cognitive Walkthrough step 3 45
5.5 Cognitive Walkthrough step 4 46
5.6 Cognitive Walkthrough step 5 47
5.7 Cognitive Walkthrough step 6 48
5.8 Cognitive Walkthrough step 7 49
5.9 Cognitive Walkthrough step 8 50

5.10 Cognitive Walkthrough step 9 51
5.11 Cognitive Walkthrough step 10 52
5.12 Cognitive Walkthrough step 11 53
5.13 Cognitive Walkthrough step 12 54
5.1 BPMN process diagram of the existing topic modeling pipeline . 55

List of Tables

1.1 Table showing the sourced questions and the pipeline step which could provide an answer 4

2.1 Table showing all used inclusion and exclusion criteria 9

3.1 Table summarizing the key features of different document embedding techniques 22

4.1 Table showing the top five similar words for two queries by word 28
4.2 Table showing the top five similar words for two queries by document 29
4.3 Table showing the top words for Figure 4.4 35
4.4 Table summarizing the found usability design issues 38

1 Thematic Introduction and 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.). Considering that even the developers sometimes struggle to understand the reasoning and deductions of their own systems [Dee], it is easy to imagine that the general public has a hard time understanding the capabilities and shortcomings of machine learning.

1.1 Project IKON

One project which operates in the cross section between machine learning and a non-technical target group is project *IKON*. It tries to explore potentials for knowledge transfer activities at a research museum and was started in co-operation with the German Natural History Museum in Berlin which houses more than 300 scientists, PhD students and other staff [Tea18]. 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[BMKng]. 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 [14018]. 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, to the final target group. Potentials for knowledge exchange are made explicit by visualizing projects not in the organizational hierarchy of the museum, but instead in semantic relation to each other. This approach has become evident from qualitative work done by the project researchers [BMKng]. Therefore, I have developed a topic modeling process consisting of four major components, as seen in Figure 1.2. In the following section, I discuss the basic principles of such a pipeline, before moving on to the specific challenge of interpretability in this context.

1.2 Topic modeling

A generic topic modeling pipeline consists of four steps:

1. Document embedding

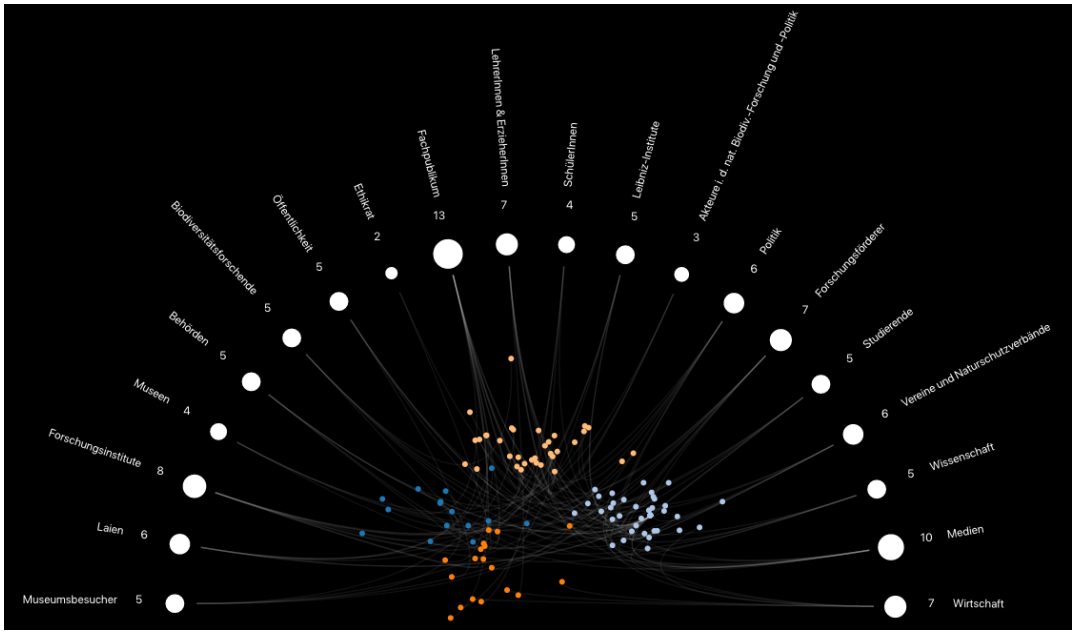

Figure 1.1: Screenshot of the cluster view of the IKON visualization

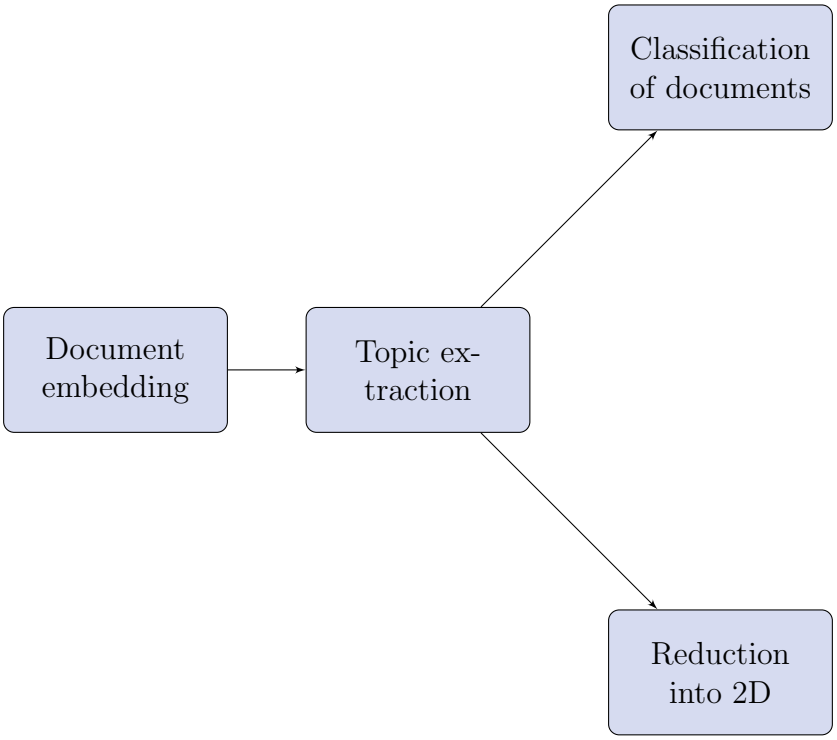

Figure 1.2: Components of a general topic extraction pipeline

2. Topic extraction

3. Classification of documents

4. Reduction into 2D

Given an unlabeled corpus $C = \{D_1, ..., D_n\}$ consisting of n documents $D_i = (t_1, ..., t_m)$, which in turn consists of a sequence of m strings, also called tokens or words, the document embedding step assigns to each document a vector $v_D \in \mathbb{R}^e, e \in \mathbb{N}^+$. Semantically similar documents should also be closer in the embedded vector space with respect to a given distance measure than documents which are semantically not related. Therefore this step transforms a corpus into a matrix $(v_1, ..., v_n) \in \mathbb{R}^{e \times n}$.

Consuming the output from the previous step the topic extraction tries to uncover k latent structures. We call these structures *topics*. Mathematically speaking a topic is a probability distribution over a fixed set of input features. [LTD+16] These features can correspond to tokens, as it is the case in the later discussed Tfidf-BOW embedding, but this does not have to be the case. Therefore this step transforms the corpus from the embedding space of dimensionality $e \times n$, where each document is described as linear combination of features, to the latent space of dimensionality $k \times n$, where each document is described as a linear combination of latent topics. Since most often $k < e$ holds true, this can also be seen as a form of dimensionality reduction, which is again a form of feature extraction.

Using the document vectors in the latent space each document is assigned a label. This may happen in a supervised way if there are labels available for training purposes, but in most cases an unsupervised classification, also known as clustering, is used to group the documents.

The last step is formally not part of the topic modeling itself, but if the use case demands a way to visualize the high dimensional distribution of documents in the latent space, another dimensionality reduction is used to project the documents to 2D. Since this is the case I include it in the process as well.

Each component in Figure 1.2 comes with its own set of parameters which influence the results generated by the pipeline. Therefore the researchers from project IKON hypothesize, based in first interviews and conceptual work, that the museum's staff, as non-technical experts without knowledge of the capabilities and shortcomings of the models used in topic modeling, will have a hard time interpreting and understanding the output 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 an exploratory 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, hypothesize during the interaction between user and visualization. Followed by an explanation of Figure 1.2 we discussed how these challenges may correlate with specific components and steps of the topic extraction pipeline. Following a description of the key questions each question

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?	Classification
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

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.

1.3 Interpretability

Due to the long-standing issue of visualizing the museum research projects, there was already an existing NLP pipeline in project IKON which I had implemented as a proof-of-concept.. However, due to the concerns discussed above, the researchers became focused on the challenge of interpretability, which I discuss in this section. In the same workshop we also developed a way of structuring commonly used terminology in the field of explainability/interpretability research, where terms explainability and interpretability are commonly used as synonyms. Our notion builds upon the findings of Lipton, who suggested that, interpretability as a term is 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. Furthermore 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 such as explainability or causality research.

Therefore the researchers from project IKON hypothesize that the context in which interpretation is performed is essential to the outcome of the interpretative process. In this discussion the term 'context' considers the situational context of the interaction between user and system as well as the historical experiences of the user. The IKON researchers therefore proposed [1] a relational model in which users posses a set of *a priori* preferred explanation strategies (e.g. comparing two entities) given the context of the interaction and their previous experiences. As a consequence of having this preferred set, they only consider a specific subset of all possible types of explanations as valid. Interpretability techniques on the other hand are conditioned in regard to an algorithmic system, e.g. a specific model or class of models. These

[1] Forthcoming publication.

techniques can be described by algorithms and deliver concrete explanations given a model and a model output. Therefore an interpretability technique serves as the missing link between high-level explanation strategies, an algorithmic system and explanations. The algorithmic system upon which these interpretability techniques are build on was implemented as a proof-of-concept by me before the start of this thesis. Building the pipeline brought up serious concerns and uncertainty concerning the meaningfulness of approaches such as parameter manipulation or choosing between algorithms for dimensionality reduction [BMG18] in our use case. Therefore I will explore the space of alternative methods. I developed a working plan for my thesis, which I will outline in the following section.

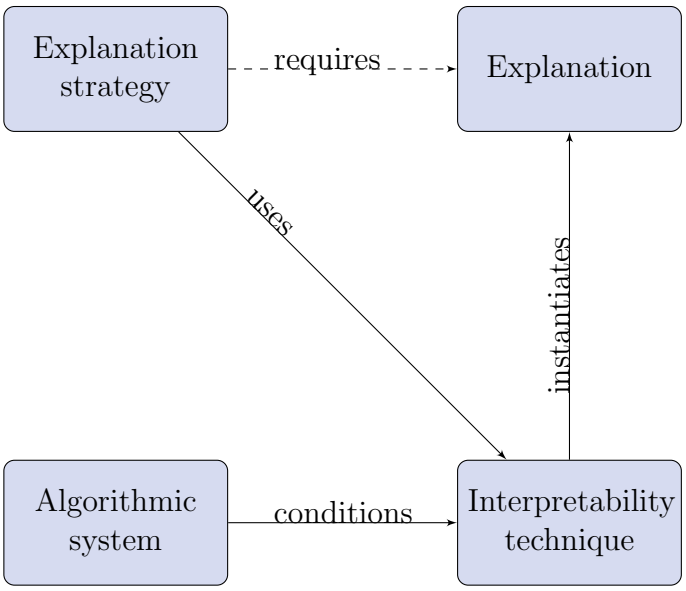

Figure 1.3: Proposed model by the researchers showing the interplay of explanation strategies, interpretability techniques and explanations

1.4 Working plan

In order to research which interpretability techniques could be applied in project IKON, I conducted the following three steps:

1. Since I as a developer did not possess an exhaustive list of techniques to enhance interpretability for unsupervised NLP models exist, a thorough and reproducible literature analysis on the status of XAI research in the field of NLP according to Petersen et al. [PFMM] is going to be conducted. This method should result in a number of papers which are, according to the process, good representatives of the literature base and therefore also of current research efforts. A quantitative analysis of these

papers should summarize occurring XAI methods and categorize them according to an applicable categorization.

2. The currently existing topic extraction pipeline can be generalized into the following four components: document embedding, dimensionality reduction into a topic space, clustering and another dimensionality reduction into 2D. Based on the results of the previous step for each component either a directly applicable method (e.g. a clustering algorithm) from a paper or a model which supports most collected methods (e.g. a neural network for document embedding) is chosen and implemented. Since the new pipeline should capture at least as much information as the old one, each component will be quantitatively accessed according to applicable measures e.g. ([RBH15]). This is necessary to ensure that one is actually interpreting existing and captured semantic relations and not random artifacts generated by the various methods.

3. A full user study would normally be necessary to assess how the implemented methods may support a non-technical expert in interpreting the results of the pipeline, but in order to keep the volume of this thesis in a feasible frame I will resort to a cognitive walkthrough from the point of view of a researcher from the national natural history museum. Since ensuring robustness in such qualitative tests is always a concern, information from previous interviews with domain experts from the museum will be used to derive meaningful tasks. The walkthrough should show how the user interacts with the implemented interpretability techniques.

2 Literature mapping study

2.1 Motivation

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

2.2 Method

The process from Petersen et al. 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 for NLP, I chose the following questions:

   a) What categories of interpretability techniques are mentioned in the corpus?

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

   c) Which interpretability techniques are applicable to the pipeline or to intermediate results of it?

2. Construction of a search string:

   Based on the questions I developed 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 processing" 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 the majority of the research is still very technical and research from social sciences rarely influences it. Even though Arxiv is a preprint server and not a 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.

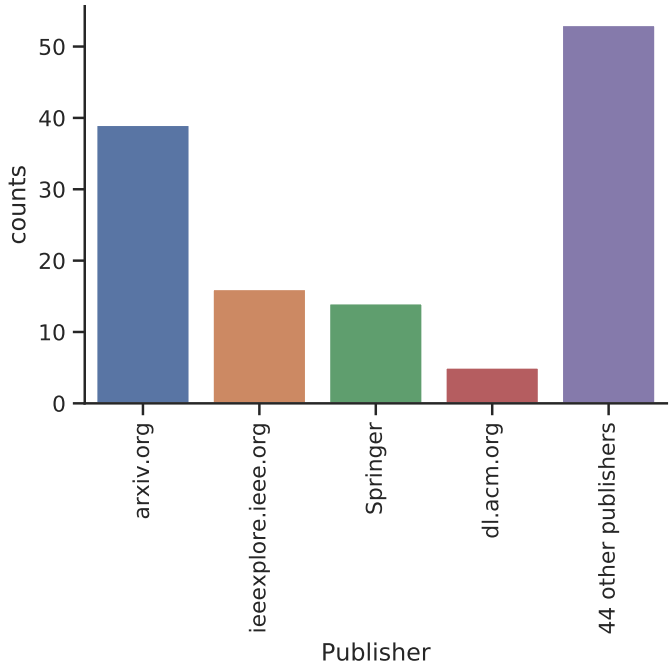

Figure 2.1: Barplot displaying the distribution of publishers occurring in the meta search results

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

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

Table 2.1: Table showing all used inclusion and exclusion criteria

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. 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. In order to support my decision making and minimize the amount of work to classify each paper I developed a Jupyter-based interface, which takes a bibliography and a set of inclusion and exclusion criteria and iterates over all contained publications, shows its title and abstract and allows the user to select criteria which apply. If a closer examination is needed it opens the paper on demand. Furthermore it sorts each publication into either a bibliography for the next stage, a bibliography with rejected publications depending on the applying criteria or a bibliography containing interesting, but not directly relevant literature. I opensourced this framework on GitHub [1].

[1] https://github.com/wittenator/limap

6. Intermediate assessment of the corpus:

   Looking at the distribution of tags in Figure 2.3 it appears that the chosen keywords represent the field well. There are no tags in the first five entries which are not constructable by the query. Plotting the distribution of publishing dates of the papers from the corpus in Figure 2.2 reveals that the first publications were already written in 1999, while there is a surge of interest and research in the last four years. This speaks in favor of the premise that interpretability research is not necessarily a young, but a recently thriving field. Investigating the earliest paper [TG99] reveals that it does not differ significantly from the rest of the corpus concerning the techniques or the understanding of interpretability it uses.

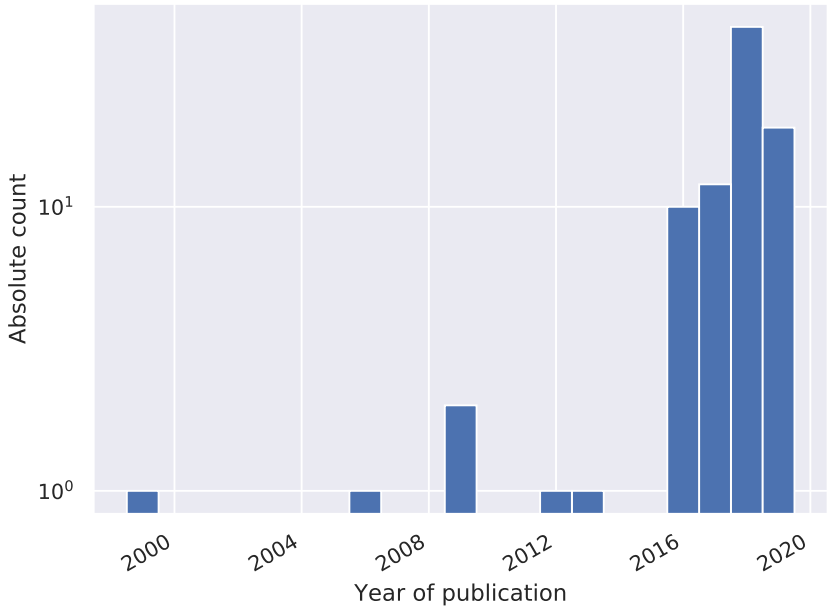

Figure 2.2: Barplot displaying the distribution of publishing dates occurring in the results after the inclusion-exclusion step

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 in order to answer the initially posed questions. To answer the first question I categorized them according to the proposed categories of Hohman et. al. [HHC+19], who also conducted a literature study.. They propose a set of categories which fit the description of explanation strategies in the initially proposed relational model. These categories, henceforth referenced as Gamut classification, are not a perfect fit for a thesis dealing with interpretability for non-technical experts since it also categorizes techniques according to their mathematical inner workings, but Hohman et al. extended the

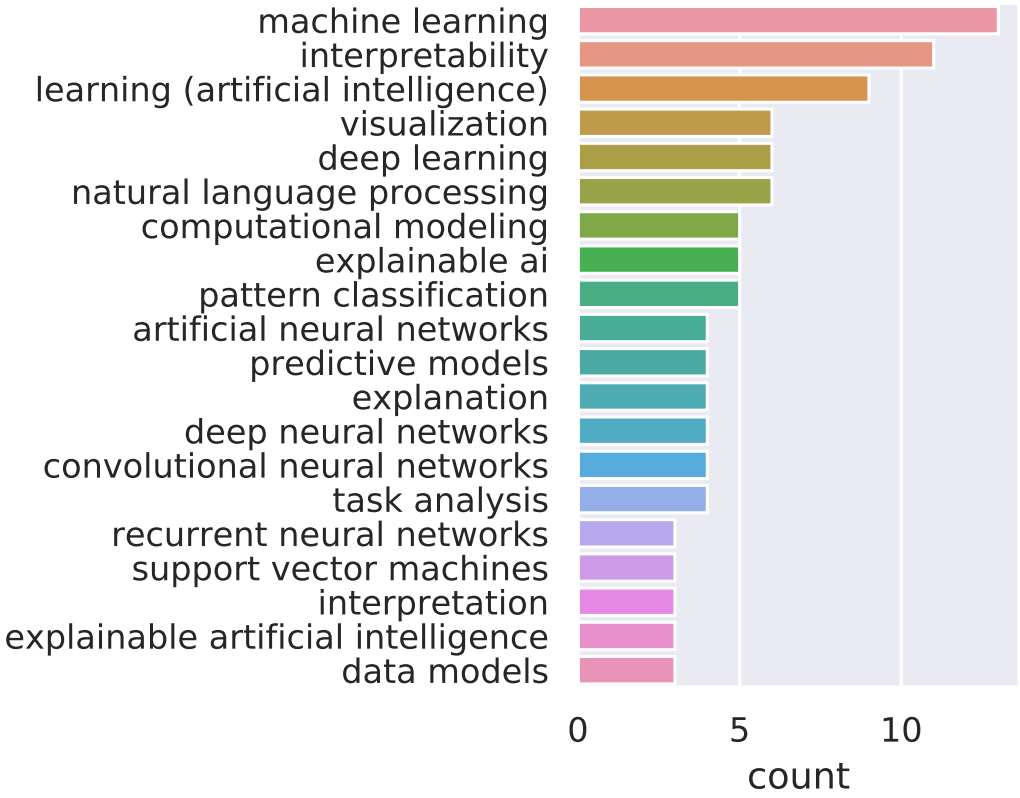

Figure 2.3: List of the 20 most used tags and their absolute frequency

categories proposed by Lipton [Lip16], which formulated the starting hypothesis for this thesis and is the closest to a nontechnical assessment of interpretability research I could find. Furthermore there are not many such classifications to begin with, so I chose one which was recently introduced and presented at a major conference (CHI 2019). Additionally, I expanded the proposed set of categories by one strategy called 'decomposition', because during the analysis of the corpus I encountered the situation of not being able to classify a technique in one given bin. Investigating the non-assigned techniques led me to understand that most of them explain the interplay between components in the model they try to explain or decompose a model and use existing techniques to explain the components. For the rest of this thesis I refer to this set of categories as Gamut+. Each publication was therefore assigned the type of explanation strategy it supports, the type of model to which the technique is applicable, the component to which the technique could be applied in the topic extraction pipeline and each paper was classified as either "Theory", "Method", "Study" or "Report". A "Method" paper presents a single interpretability techniques 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 interface evaluation which visualizes the output of interpretability techniques. Publications from the last category are therefore less technical and more concerned with the HCI aspects of interpretability 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. This reduced the final corpus to a size of 72 publications.

2.3 Results

In order to answer the posed question in the beginning of the literature mapping study I visualize the resulting corpus as a mapping on a grid. Each dimension of this grid stands for one of the selected categories and these dimensions are divided into all possible tags in the respective categories. The intersections contain bubbles which visualize, by size and as an explicit number in the middle of the bubble, the amount of publications which were assigned both tags.

Mapping the type of paper and the classification according to Gamut+ each on an axis (Figure 2.4) shows clearly that there is a trend towards developing methods which explain single decision instances (38 paper). Furthermore most developed methods are tested on real world data (61 paper), but their application in an interface is rarely studied (6 paper). This speaks in favor of the

hypothesis that most interpretability techniques are developed as mathematical theories and influences from HCI are rarely taken into consideration and that the findings of Miller [MHS17] also hold true in the subdomain of NLP.

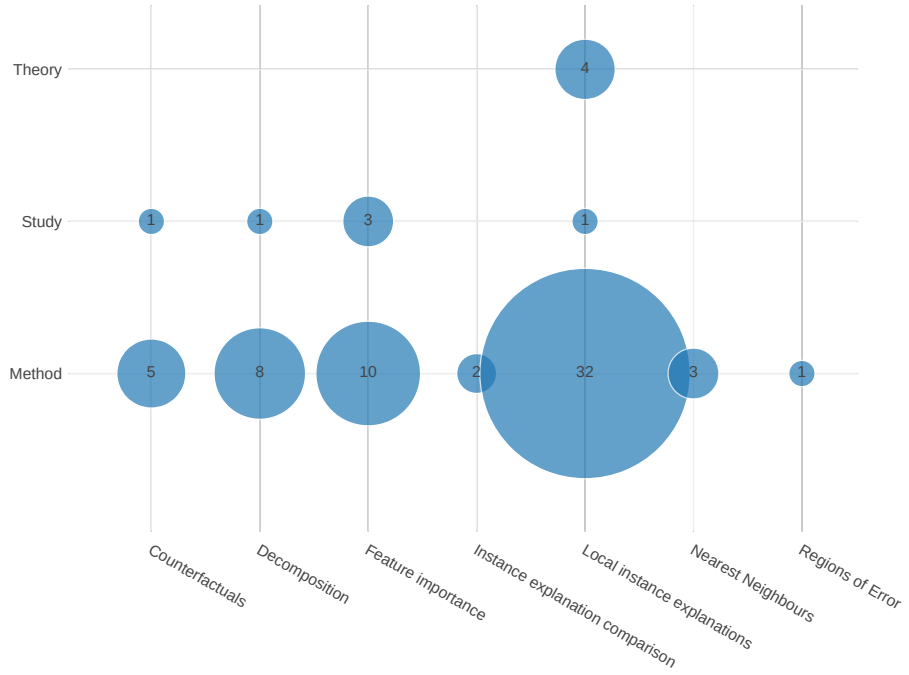

Figure 2.4: Mapping of the type of publication and its Gamut+ classification

The second question was concerned with the type of models which are enhanced by interpretability techniques. In Figure 2.5 it is visible that neural architectures (NN, CNN, FNN, RNN, GCNN) dominate the field (40 paper). 19 papers try to explain a given model in an agnostic way as a black box, while a minority of publications deals with the interpretability of clustering results, decision trees or linear models.

The third mapping in Figure 2.6 shows the relation between the applicability of a method in the general topic extraction pipeline and its Gamut+ classification. Surprisingly, 51% of the sourced publications are not applicable to the general topic extraction pipeline in any form. The two main reasons why a publication falls into this category is that it either presents a method in a subdomain of NLP which is not directly applicable [GMPB16] [IST+18] or its presented use case and context is too far off in order to be applied [MWM18] [GCJC]. The second biggest category consists of techniques which could be applied to the document classification step using labeled data to train a model. Since any neural network can be used to classify vectorized documents, most

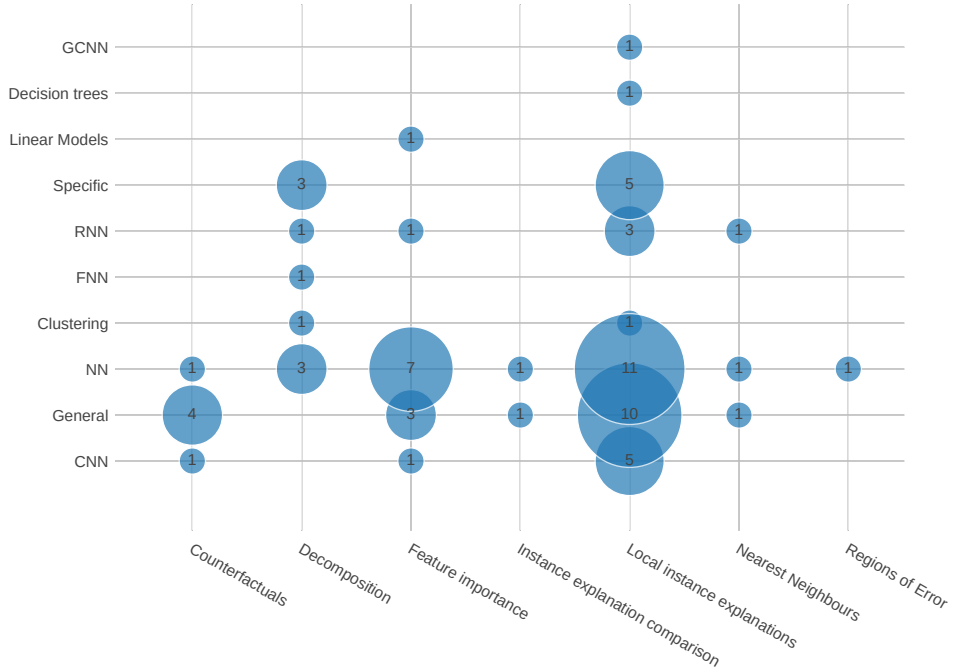

Figure 2.5: Mapping of applicability and Gamut+ classification

of the publications on the "NN" axis in Figure 2.5 fall into this bucket as well. All in all, 9 publications remain which could be applied to an unsupervised topic extraction pipeline. The document embedding step could be made interpretable by decomposition or by explaining the embedding of single instances. The corpus delivers no techniques which could be applied to a unsupervised dimensionality reduction, but unveils 7 potential techniques for unsupervised classification, also known as clustering. This step can has also the most variety of potential interpretability techniques only missing a technique which uses instance explanation comparison as an explanation strategy.

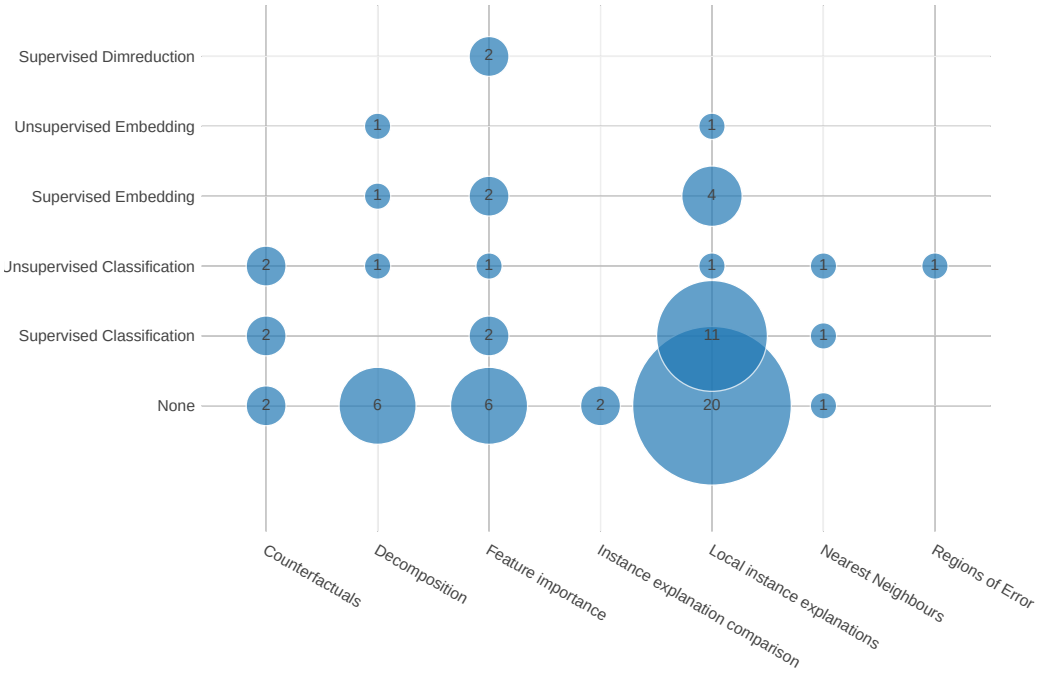

Figure 2.6: Mapping of pipeline step and Gamut+ classification

The literature mapping study revealed a few interesting insights into the field of interpretability research in NLP, while yielding surprisingly few applicable techniques. This may be due to the selection of only the first 100 publications in the scientific database or due to introducing a false bias to the results by including key words like 'neural network' or 'deep learning'. These techniques are normally connected to supervised learning which is not applicable to this use case from the start. These findings, especially the analysis showing the distribution of explanation strategies over models, makes it possible to tackle the problem of the proof-of-concept pipeline having theoretical shortcomings while simultaneously improving the interpretability of each component.

3 Implementation of the Topic Modeling Pipeline

3.1 General setup

In order to ensure that the results of this thesis are usable for further work and research, it was one of my priorities to integrate all my code into the existing project as well as possible. Since the IKON project [1] uses a Docker-based microservice architecture to develop and manage their servers, I decided to integrate the Jupyter Notebook, which I used as my main tool for code development and documentation, into this network. Doing this also enabled the Notebook to dynamically fetch data from the Postgres database which serves as the main source of information. In anticipation of huge computational loads the Docker container was designed to make use of a potential graphic card. Therefore I built all my work on top of the official Tensorflow Docker image which comes with all the drivers for NVidia GPUs.

3.2 Data and Preprocessing

Since one of the main aims of project IKON is to connect projects semantically instead of by using the rigid hierarchy of the museum, I was able to use the project's abstract 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 [Spa18] 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 150 words, while also having smaller peaks at ca. 70 and 350 words. The shortest description has a length of one word and the longest 983 words.

[1] Needs access to a private repository for keys. If access is wished for, please contact the author.

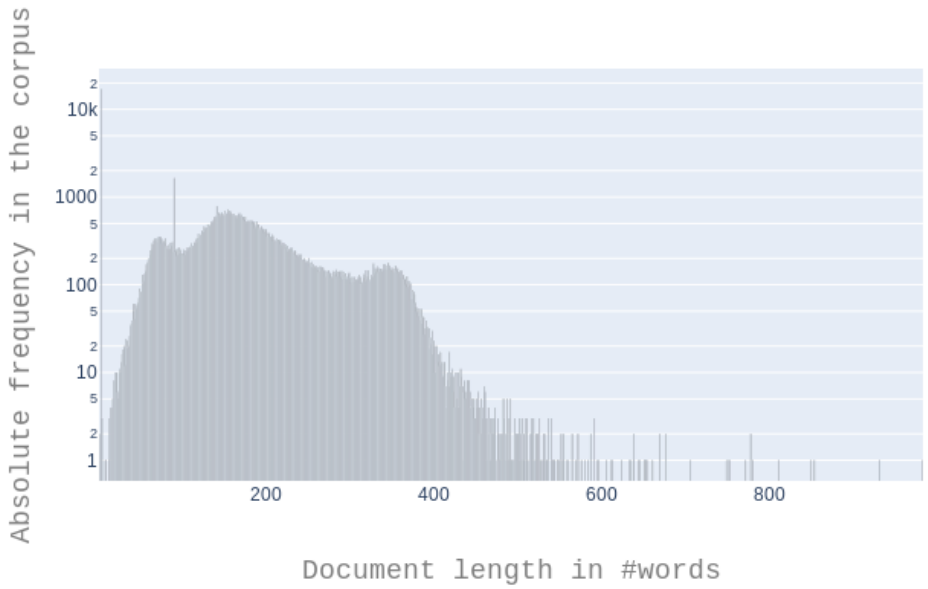

Figure 3.1: Histogram showing the distribution of text lengths in the dataset

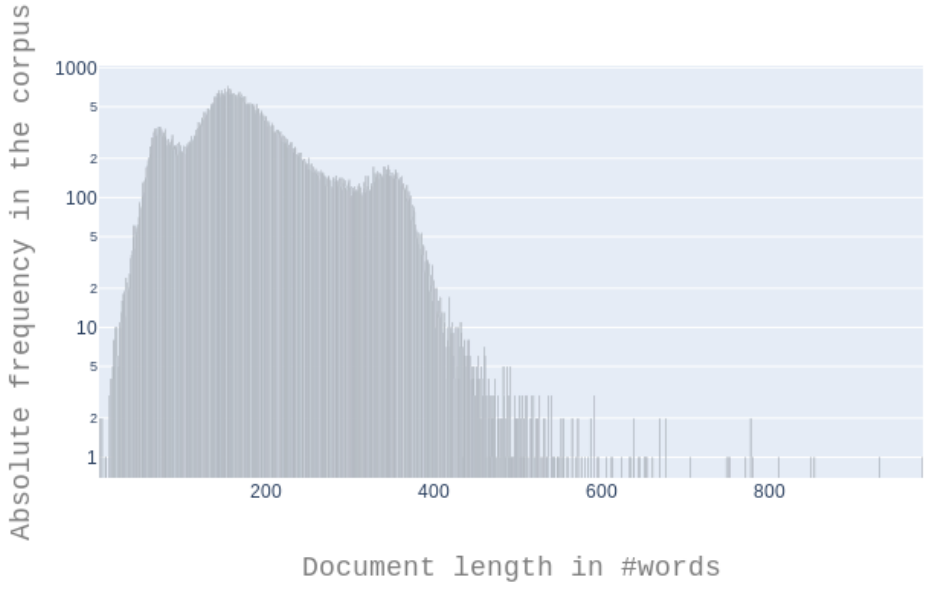

Figure 3.2: Histogram showing the distribution of text lengths in the dataset excluding duplicates and projects without a description

Following the analysis of Matthew et al. [Den17] the texts were preprocessed by a S-P-N-W scheme. Following this, according to the categories of Matthew et al., a stemming step (S) is performed first, which uses lemmatization to find the lemmas of words by using vocabularies and the context of each word. Then 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. 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.

3.3 The existing pipeline

As discussed above, 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 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. 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 there is a vector representation of each document, the clustering of the documents could be performed in this vector space using K-Means. A common problem that occurs in such spaces is 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 metric

for the k-Means algorithm due to it using the Euclidean distance, becomes effectively meaningless and making it necessary to reduce the dimensionality of the vector space. That suggests that using K-Means directly in the embedding space leads to the clustering algorithm failing to perform. For that reason a dimensionality reduction technique needs to project the data in a vector space of lesser dimensionality first, before further analyses can be conducted.

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 might correspond to 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 Deerwester et al. 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. The existing pipeline uses a k-Means algorithm to discover clusters and classify the documents as a next step. Finally, in order to visualize the high dimensional topic space in 2D a linear discriminant analysis is used using the clustering as labels. This process is formalized as a BPMN diagram in Figure 5.1.

3.4 Components

As mentioned in the Introduction, the previously described proof-of-concept topic modeling pipeline suffers from limitations due to the very nature of the methods. Arras et al. showed for example that the SVD shares a limited amount of expressiveness with different other linear methods due to its purely linear nature [AHM$^+$17]. Since interpretability techniques , which work on top of these models, can only convey as much information as the underlying model captured, it is worth exploring the space of alternatives. Furthermore the literature mapping study in chapter 2 showed that the existing techniques may not be well suited for augmentation by interpretability techniques which suggests that a different choice may improve interpretability as well.

3.4.1 Document embedding

A short survey of document embedding techniques

Since 1972, the year when the Idf measure was proposed for the first time, [Rob04] a number of other techniques appeared, which are able to vectorize documents in a corpus.

Le and Mikolov [LM14] proposed *Paragraph vectors* almost a decade later using the newest advances in neural networks. This technique, also known as *Doc2Vec*, because it expands the idea of Word2Vec [MSC+] to documents, utilizes a shallow neural network to run over each document with a sliding window and predict a token in this window using the other tokens and a paragraph id as a special token as context. Using a standard backpropagation algorithm to train the weights of the network the final paragraph vector consists of the weights which are used for the paragraph id. The intuition is that the paragraph vector acts as an additional storage for context information and since the connected paragraph ID is unique for each document it contains semantic information for the entire document. Choosing a low dimension as an embedding dimension also corresponds to the embedding and topic extraction step at once, but the authors recommend an embedding dimensionality of at least 100.

A rather new method was presented by Wu et al. [WYX+18] using a new distance metric called *Word Mover's distance*(WMD). This metric uses pre-trained word vectors and word alignment in order to compute more meaningful distances. Because the computation of this metric is quite expensive, Wu et al. develop an approximative kernel which embeds a corpus into a vector space using the WMD, which can be used instead of computing the full kernel with all the training data.

Another approach would be to not train a model on the specific dataset, but rather use a model which was pretrained on a huge and very general dataset. One of the state-of-the-art techniques for that is BERT [DCLT18]. Devlin et al. present a new model architecture based on the popular Transformer model [VSP+17] and train it in the first version on a concatenated corpus of BookCorpus and the English Wikipedia ($3,3 \cdot 10^9$ words in total). Having such a huge amount of data as context knowledge one is now able to train another model for downstream tasks on top of BERT and utilize the knowledge extracted from the corpus in a transfer learning fashion. It is also possible to extract the raw document embeddings from BERT directly, but the sequence length is capped to 512 characters.

Selection of a document embedding technique

Summarizing the previously discussed methods by two of their main characteristics - maximum processable document length and type of model results in Table 3.1.

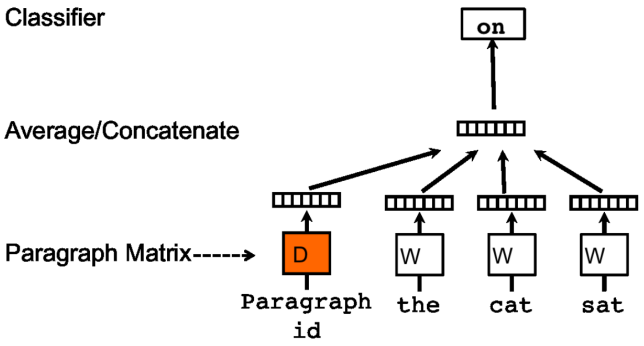

Figure 3.3: Visualization of a training step of a Doc2Vec network [WYX+18]

Technique	Max. document length	Type
Tf-Idf BOW	unlimited	Probabilistic
Doc2Vec	unlimited	NN
Word mover's embedding	unlimited	Kernel method
BERT	512 characters	NN

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

The model is now selected by exclusion. Since our database contains documents which are longer than 512 tokens and each token has a length of at least 1 character, BERT is eliminated as a potential document embedding technique. It would be possible to take word embeddings from BERT and average them in order to get a document embedding as it was proposed and further developed in [DBVCDD16] for Word2Vec embeddings, but there was no scientific or non-scientific literature that suggested that this works for the case of contextualized BERT embeddings. Furthermore the previous literature mapping study showed that there is not a lot of work done for explaining probabilistic models or models utilizing kernel tricks, therefore TF-Idf BOW, the LDA and the Word Mover's embedding are not of interest in this case. Only Doc2Vec remains, supporting both an unlimited document length and being of type 'NN' and therefore potentially being able to support at least 7 interpretability techniques sourced from the mapping study. Since neural networks are a focus for interpretability research at the moment, as visible in Figure 2.5, this technique is possibly more future-proof in regard to being able to support interpretability techniques which are yet to be developed.

3.4.2 Topic extraction

Currently the only used topic extraction method is the LSA. Since the underlying SVD is a purely linear technique, the questions stands if the results of the topic extraction improve when nonlinear features are taken into account. One

technique to perform an nonlinear, unsupervised dimensionality reduction is the *Autoencoder*. This type of neural network consists of an encoder network and a decoder network. The encoder maps an input to an intermediate layer, while the decoder maps a vector from its representation in the intermediate layer to a vector in the vector space of the original input.

The composition of encoder and decoder is then trained to reconstruct the input from its intermediate layer via a standard backpropagation algorithm. Choosing the intermediate layer of lesser dimensionality compresses the input vectors, which constitutes as a dimensionality reduction. As Wang et al. [WYZ16] pointed out, an autoencoder can emulate the results of a PCA/SVD by choosing a linear activation function for all neurons and may even outperform it for other nonlinear activation functions.

Adding a sparsity constraint to the encoding network can help to describe a project by as little features as possible. Since the features could be interpreted as topics, this constraint helps the clustering task downstream.

Training the model with an embedding dimension of 50, binary crossentropy as a loss function and Adadelta optimizer shows that the model achieves a loss of 0.005 after 75 epochs on the training set and approximately 0.01 on the validation set Figure 3.4.

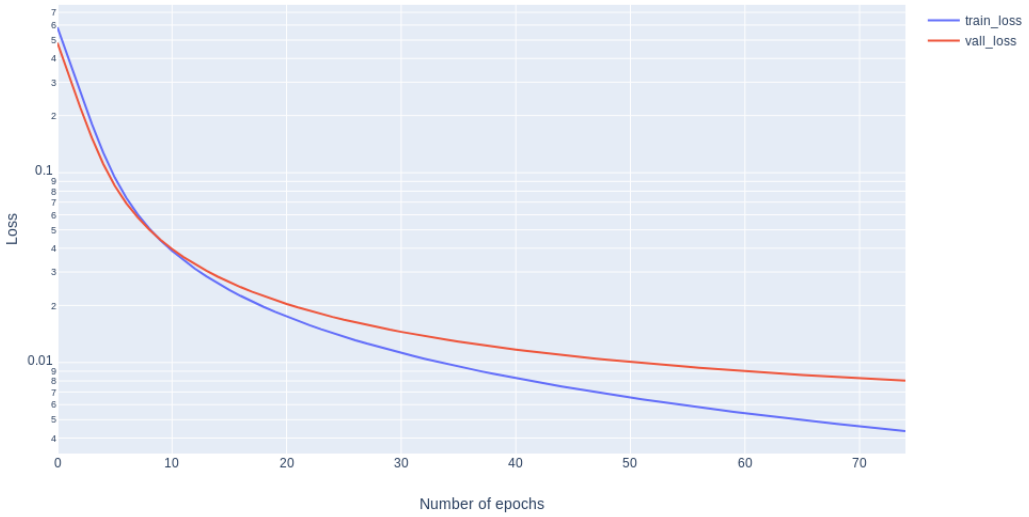

Figure 3.4: Graph showing the training and validation loss of the autoencoder over progressing epochs

3.4.3 Clustering

As described in the beginning of this chapter a K-Means clustering would now classify the documents in the latent topic space. A problem that this approach poses is that the assumption that the Euclidian distance (EuD), which is the inert similarity measure of the K-Means algorithm is meaningful in our vector space may not be true. Another distance measure which may may encode more semantic meaning could be the previously mentioned *Word Mover's Distance*. Since it uses the generated word embeddings and their order in the document to compute distances, it may be more suited for comparing texts and subsequently also yield better results for the clustering. This weighs even heavier if the corpus is vectorized as a sparse matrix since the Euclidian distance, as described in chapter 1, looses its meaning . In our case, having embeddings from a Doc2Vec model, we don't have to deal with this additional problem, but the question remains if the two distance measures differ on our dataset. Comparing the distances between documents generated by both the Word Mover's distance and the Euclidian distance in Figure 3.5 it is apparent that there is indeed a difference. Therefore it is worth investigating if using this information improves the clustering results.

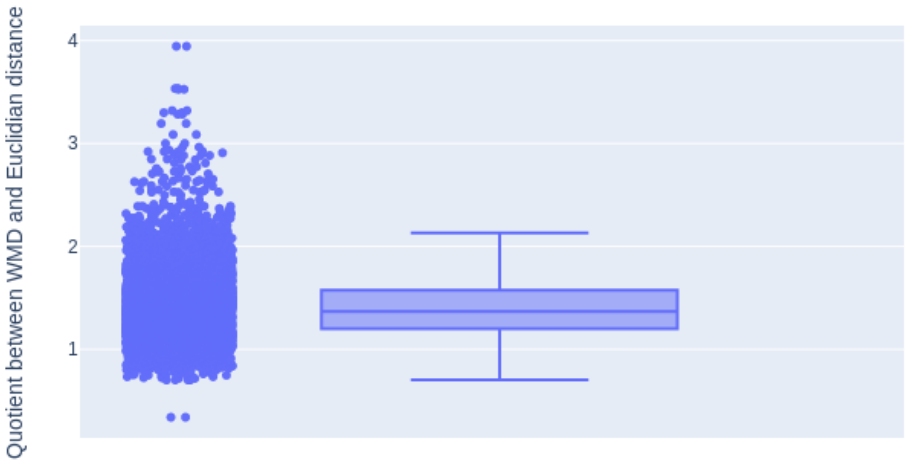

Figure 3.5: Boxplot showing the distribution of the quotients between WMD and EuD for all documents

Inspired by Liu et al. [LHLH18] I chose a hierarchical clustering approach, specifically *Agglomerative Clustering*, as a contending method to the K-Means

algorithm. This method works bottom-up since in the beginning it considers every data point to be its own cluster. Now in every step two clusters are merged which minimize a given linkage metric. The distance calculations between points are performed by a lookup in a precomputed distance matrix which enables the usage of any given distance metric. Doing this until only one cluster remains, creates a binary tree, which describes the hierarchy of the data given the used metric.

3.4.4 Visualization

In order to visualize all the results from the topic modeling pipeline I developed a D3.js-based interface, which is embedded in the Jupyter notebook in which the code for all the numerical computations resides. Since Jupyter is embracing the browser as a frontend, there is the possibility to embed arbitrary Javascript code in a cell and inject any kind of visualization. In order to do that I construct a JSON object containing all the results from the topic pipeline (vectorized documents as scatter points and linearized points, top words, the interpolated topography, model parameters etc.) and pass it into the Javascript code via Jupyter's *Javascript* function and string interpolation which renders it in the browser.

The interface consist of a discrete slider controlling the granularity of the clustering, which translates to the number of clusters which are computed in the topic modeling pipeline. Furthermore there is a dropdown selection where one can switch between the scatter and the linearized variants of the scatter plots. Beneath the parts of the interface which are connected to parameter selection, the top words for all clusters are displayed and even further down the plot is located in which each project is visualized as a circle. Changing any of these parameters result in an immediate change of the computed plot and the top words. Additionally one can hover over projects in order to reveal additional metadata about the project (project ID, title, top words).

Now that alternatives for the components in the pipeline where explored, it is worth investigating how the findings from the literature mapping study could be applied to this set of methods in order to improve interpretability.

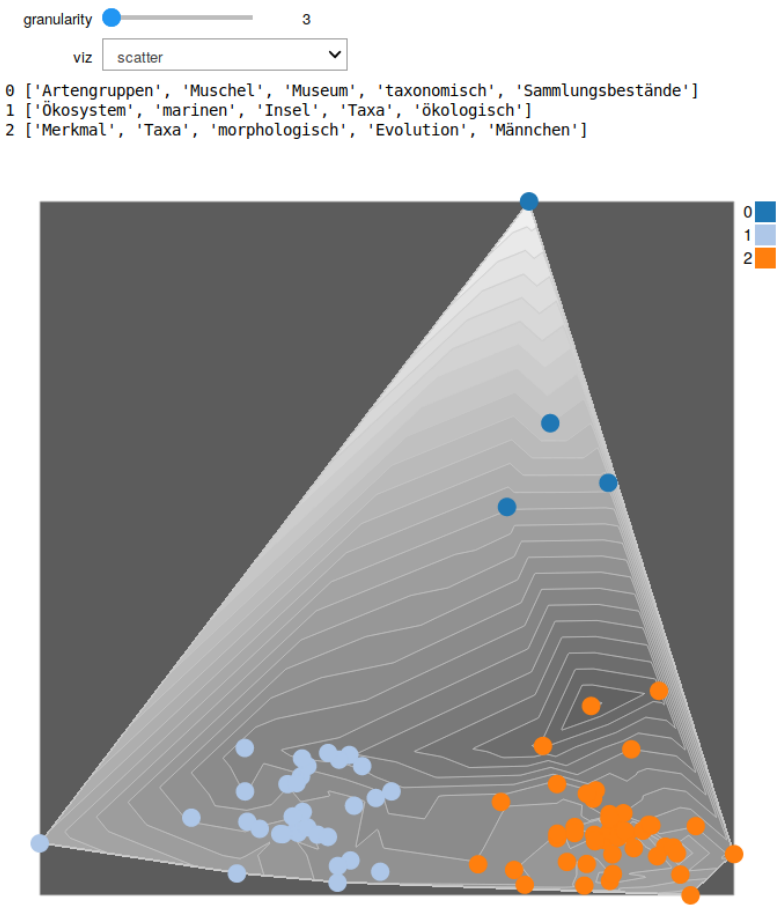

Figure 3.6: Screenshot showing the exemplary interface

4 Implementation of Interpretability Techniques

4.1 Techniques

Based on the improvements to the existing pipeline, numerous possibilities for the implementation of interpretability techniques now became available. Analyzing the sourced publications from chapter 2 reveals that none of the nine potentially applicable algorithms is technically applicable to the models which were selected in the previous steps. Even if that is the case, the ideas and concepts from these techniques can be used to develop methods for this particular use case. This also reflects the relational model developed by the IKON researchers, which shows interpretability techniques as supporting specific explanation strategies.

4.1.1 Top words

Looking at Figure 2.6 shows that local instance explanations as a strategy to explain the output of an unsupervised embedding algorithm are most prevalent among all sourced publications. Factoring in the two questions from Table 1.1 which deal with the most prominent topics and the similarity between projects and clusters suggest to answer these questions by ranking the input features to the document embedding and topic extraction step for each document and cluster. According to the Gamut+ classification this would be a local instance explanation. Applied to the augmented/new pipeline of IKON, this can be achieved as follows.

Assume that a document is described as a vector $c \in \mathbb{R}^k$ in the latent topic space. In order to show the most important features in the document there are two possible ways. The first option factors in how well the topic extraction projects the documents. Taking the vector in the latent space, firstly the inverse dimensionality reduction is applied in order to transform the vector into the embedding space. Both the LSA and the autodecoder approach have this opportunity, but they differ greatly in quality of this reconstruction. Since the LSA is a linear method, the back projection yields all documents on a hyperplane, while the autodecoder is able to minimize the reconstruction loss through its inherent nonlinearity. The second option disregards the capabilities of the topic extraction model to reconstruct a vector in the embedding space from a vector in the latent topic space. Instead for projects the embedding vector is taken directly or for cluster centers the average from all assigned projects in the embedding space is calculated as a representative. For the sake of implementational simplicity the second option was used for this thesis despite its shortcoming of disregarding the topic extraction.

”Evolution”	”Diversität”
1. evolutionären	1. Artenzusammensetzung
2. evolutionäre	2. Biodiversität
3. evolutionärer	3. Taxa
4. phylogenetische	4. taxonomisch
5. Artbildung	5. Lebensräumen

Table 4.1: Table showing the top five similar words for two queries by word

Now that the document vector is in its embedding space we will make use of a special ability of the Doc2Vec model. As described in the previous section the model does not only train document vectors, but it also generates word embeddings in the same space. The revolution the Word2Vec model, as a base of the Doc2Vec model, presented was that the generated embeddings and their relations to each other encoded semantic relations. I hypothesize that this behavior also applies to the embedding of document and word vectors into one space. This leads to the possibility of describing a document by its nearest word vectors. An exemplary analysis suggests that there seems to be a valid semantic structure in the relations between tokens and between tokens and documents. As a German native speaker it is easy for me to verify that the word queries in Table 4.1 are indeed semantically well connected. The results of the document queries in Table 4.2 on the other hand are hard to verify since most documents are very specific, scientific texts. I picked two projects and their top words which I was able to understand without relying on external information. The first three top words are indeed well connected to the topic of the corresponding project, but the last two ones seem to be off. This exemplary analysis speaks in favor of the hypothesis that there are indeed semantic connections document vectors and word vectors, though this should be seen critically and only in regard to this specific dataset and use case.

Extracting top words for the existing TfIdf method works in a similar fashion, because every entry in the embedding vector has a one-to-one correspondence to words. Taking the biggest n entries yields n top words due to the TfIdf measure directly being a indicator for the amount of contained information and subsequently importance.

4.1.2 Cluster topography

The next proposed technique is inspired by Gamut's ”Region of Error” category, which is according to the literature mapping study the sparsest category with

'Ambitionierte Amateure' - Europäische Filmclubs in den langen 1960er Jahren	Netzwerke im europäischen Handel des Mittelalters
1. Kulturpraxis	1. Opportunitätskosten
2. Kulturzentren	2. Diskursteilnehmer
3. alltagsweltlich	3. evoluieren
4. pain	4. schloss
5. Ceuta	5. Staphylococcen

Table 4.2: Table showing the top five similar words for two queries by document

only one paper using that strategy.

The hypothesis on which this technique is grounded is that clusters in a high dimensional space are inherently hard to interpret, because the cluster does not exist as an explicit object. A cluster is rather an abstract concept and solely consists of the points which are connected to it via a membership assignment. Especially once the points which form a cluster get reduced into a 2D or 3D space distances and neighborhoods get distorted. Information on the position and form of the clusters is not easily obtainable anymore. One way to mitigate this is to extract artifacts in the high dimensional space and carry it over into the 2D space where the points can be visualized.

Motivated by the exploratory nature of the interaction which the non-technical experts will carry out according to the IKON researchers, the quality of each point fitting into its cluster could be an interesting pointer to make potentially contrastive deductions concerning the structure of the clusters.

This technique computes an uncertainty measure for each point which describes how well it fits into its assigned cluster. A straightforward idea would be to assign each point the Euclidean distance between the point and its cluster centroid. Small distances speak in favor of the hypothesis that a point fits well into its cluster, while bigger distances speak in favor of the contrary. After the projection into 2D a topography is interpolated with the points and their Euclidean distances as fulcrums. This intentionally invokes the notion of a geographical surface since most scientists, especially at a natural history museum, work often with such kind of maps. Making sure that the peaks of this topography corresponds to points which fit very well into their clusters enlarges the metaphor and ensures that people are intuitively able to understand the visualization.

The problem with this approach is that, firstly the Euclidean distance may

not be well suited for such vector spaces, as discussed earlier, and points in different clusters cannot be compared since volume-wise larger clusters exhibit a higher distance for each point than ones which are more compact, albeit having a similar fitness in relation to the points of the same cluster. Secondly points may exhibit the same fitness and lie close together in 2D, but may be on two opposing ends of an n-dimensional sphere in the latent space.

The second problem is harder to tackle since that is one of the inherent problems of dimensionality reductions and therefore I will present an argument for an alternative similarity measure.

The main information that this technique should convey is how well the clustering method captured the structure of the high dimensional space which can be expressed by measuring how sure the method is about the cluster assignment for each point. One measure capturing this information is the silhouette score [Rou87]. The silhouette score takes, for each sample, also the nearest, non-assigned cluster into account, computes the mean intra-cluster distance a and the mean nearest-cluster distance b and scores them as $\frac{(b-a)}{\max a,b}$. This leads to a score between -1 and 1, where a negative value denotes that the point was assigned to the wrong cluster since the nearest, not-assigned one is closer than the assigned cluster. The normalization of this score makes it possible to compare the fitness of points between clusters as well.

I developed this technique in cooperation with Christoph Kinkeldey and Jesse Josua Benjamin, presented it at TrustVis19 in Porto, Portugal, as well as published it in [CTJ19].

4.1.3 Linearization

As Lipton [Lip16] pointed out, a visualization is already an interpretability technique in itself. Therefore we can use the vast amount of interaction design research to optimize the existing interface.

One of the main problems of scatter graphs is overdrawing, also called clutter. This problem occurs when glyphs are close enough in a scatter plot so that they overlap each other. The higher the density of a region is, the harder it gets to perceive the total number of points and the harder it is to annotate the glyphs with additional data [MG13].

Normally this problem is hard to solve since the position of points in respect to each other encodes important information. Considering that in this case a high dimensional space is reduced to 2D, there are inherent reduction errors introduced into the distances between points and the absolute distances in 2D loose part of their semantic importance. Therefore it is possible to assume that instead of distances, the neighborhood of a point is more important, which leads to the possibility to map the scatter graph on a more regularized structure while preserving the neighborhood of points as well as possible. One way to do so, is to map the points in 2D onto a grid. This problem reduces to a linear assignment problem (LAP), where each point should be assigned to its nearest point in the grid while minimizing the global displacement er-

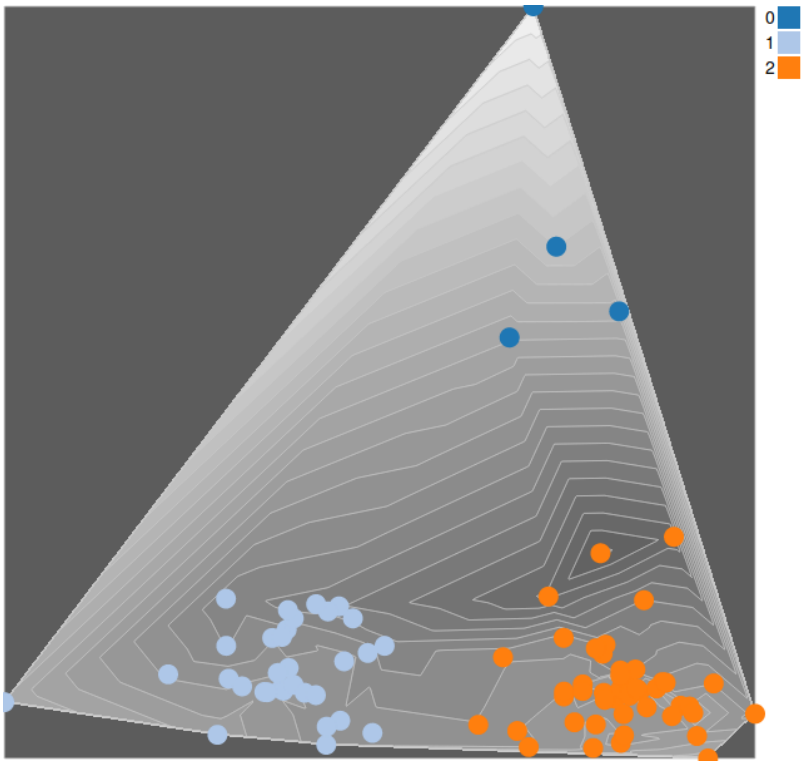

Figure 4.1: Screenshot showing the interpolated topography on an exemplary plot

ror. A popular algorithm for that is the Jonker-Volgenant algorithm [JV87]. Jonker and Volgenant rephrase the LAP as a shortest path problem and by using augmenting paths improve the previously popular Hungarian algorithm to cubic worst-case complexity. As we will see in one of the next subsections, this technique also synergizes well with the cluster topography method, because the interpolation also get linearized and the topography gets expanded presumably making it easier to see differences in the relief.

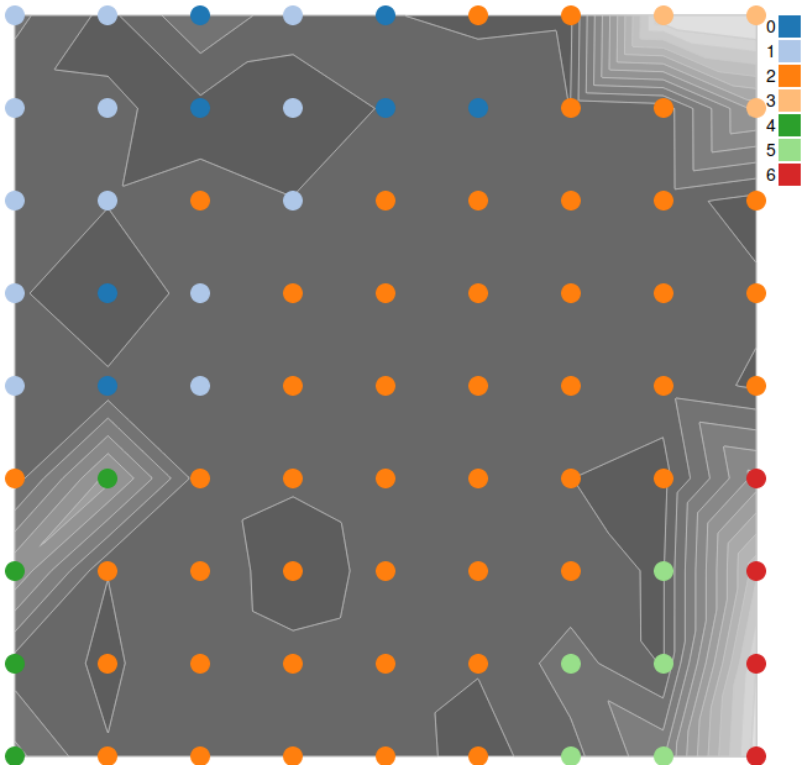

Figure 4.2: Screenshot showing the linearization on an exemplary plot

4.2 Validation

The core idea of the relational model, which was introduced in chapter 1, is that both humans and algorithmic system play an integral role in interpretative processes. Therefore it is also necessary to analyze both parts if one wants to validate such compound systems.

4.2.1 System-Centered

Now that there is an option to choose between two options for the embedding, topic extraction and classification step of the generic topic modeling pipeline, it is interesting to see which combination of these methods performs the best. A popular method to measure the quality of a topic modeling pipeline is the

Coherence Score. Röder et al.[RBH15] proposed an agnostic way to evaluate topic models, unifying several contending methods at that time. Their generic pipeline consists of four steps: segmentation, probability calculation, confirmation measure and aggregation. The segmentation step consists of segmenting the dictionary, the set of all occuring words, into subsets. Then for every word a probability is calculated using a reference corpus, which is used in the subsequent step where a confirmation measure scores the word pairs concerning how often they appear together in documents or in sliding windows over documents. A last step aggregates all the agreement scores and calculates an overall coherence score for the segmentation on the supplied reference corpus. In our case the segmentation is done by supplying the top words for all clusters. In this analysis the c_v measure was chosen, because it displayed the highest correlation (0.731) with human ratings in the experiments of Röder et. al.

Plotting the coherence scores over different embedding, topic extraction and clustering models and different number of clusters reveals (Figure 4.3) , surprisingly, that the combination with the best scores is an TfIdf embedding, an LSA topic extraction, followed by an Agglomerative Clustering with 10 clusters. The combination of TfIdf embeddings and an autoencoder couldn't be tested since in order to feed sparse matrices to a Keras neural network there is a considerable amount of work involved. In order to keep this thesis reasonable, this analysis was therefore excluded and is recommended for further work.

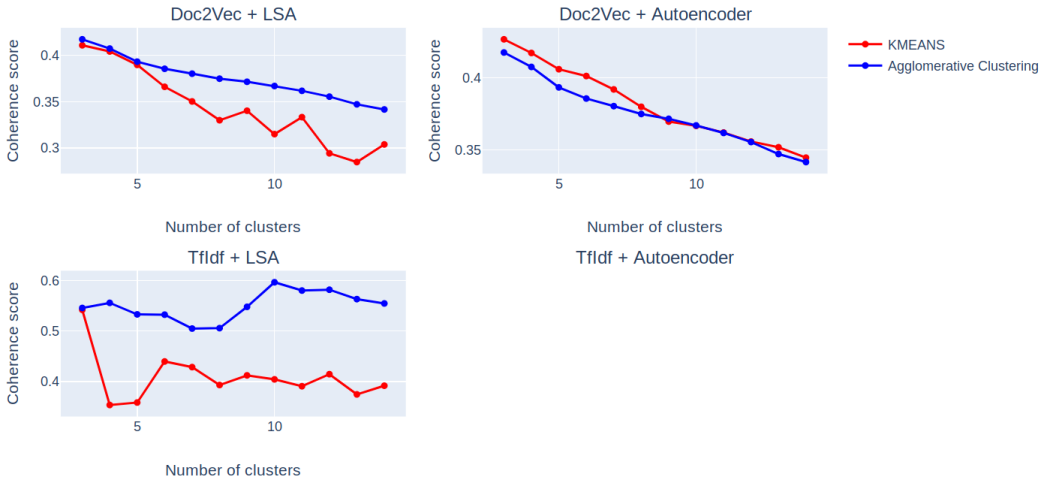

Figure 4.3: Graph showing the quality of the topic modeling while varying the embedding, topic extraction and clustering model

Plotting this combination of models and parameters in Figure 4.4 reveals that there is one huge cluster (cluster 0) while the other 9 clusters contain maximally five projects. A further analysis of the dominating cluster also

showed that although its top words suggest projects connected to evolution and biology, there are also a number of projects which deal with geology and paleontology. Interestingly, changing the parameters of all models does not change the fact that such a huge cluster forms speaking in favor of the hypothesis that these projects indeed form a huge cluster in the high embedding space.

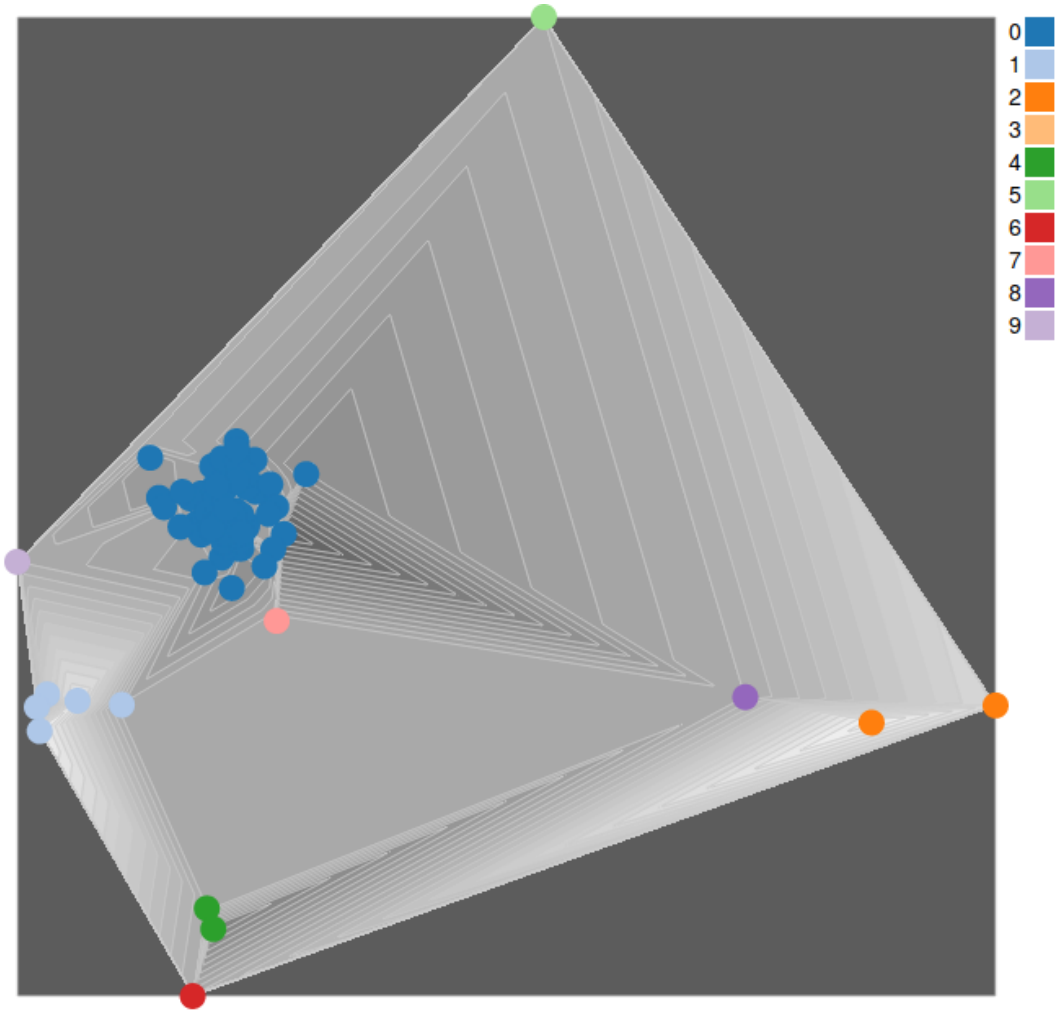

Figure 4.4: Plot for the parameter and models with the best coherence score

At this point I gathered feedback from the researchers from project IKON and their answers suggested that problems may occur trying to visualize such dominating structures in scatter plots. Therefore I turned to the second best performing combination of models which is a TfIdf embedding, an LSA topic extraction and a K-Means clustering - which is effectively the topic modeling pipeline which was existing at the beginning of the thesis. Even though this meant a return to the already existing pipeline I had implemented as a proof

0	Datum, Evolution, morphologisch, Taxa, ökologisch
1	Sackflügelfledermaus, Abwanderungsverhalten, Weibchen, Männchen, Harem
2	Western, Mountains, Kaokoveld, Fauna, Escarpment
3	Tansania, biostratigraphische, palynologisch, Tendaguru, Palynologie
4	Acentropinae, Philippinen, Trichoptera, Lepidoptera, Reliktendemiten
5	SCHRANK, gesichert, Flora, Sauropoden, Gymnospermen
6	Praktik, kolonisieren, Mediziner, Gesundheitsbehörden, Malaria
7	Magmaozeans, Magmaozean, Planetare, Impaktprozess, Impaktors
8	Kurzexpedition, Basilosauridae, Pabdeh, Ablagerungen, Iran
9	Hauptexporteur, Kieselalgendiversität, Känozoikum, Silikatverwitterung, Kieselalgen

Table 4.3: Table showing the top words for Figure 4.4

of concept, developing the interpretability techniques and the following analysis made me realize that the more complex models are not necessarily the better fitted ones and sticking to systems out of theoretical arguments may actually decrease the perceived interpretability. Since all the implemented interpretability techniques also work for this combination of methods, I stuck with it for the next validation step.

4.2.2 Human-Centered

In order to show how the implemented interpretability techniques may help a non-technical expert understanding the output generated by the pipeline, a proper user study would be needed. Since that is a task which would fill a bachelor thesis on its own, I resorted to a strategy which involves less work, but also delivers qualitative insights into the interaction between a user and the topic modeling pipeline and its visualization - the cognitive walkthrough. Performing this method involves seeing things from the perspective of a fictive user and interacting with the application in their stead. Since the nature of this visualization supports exploratory interactions in the first place, standard approaches for cognitive walkthroughs, like they are formulated in [WRLP94], do not work well due to the necessity of coding an interaction sequence prior to the simulated interaction. Allendorf et al. [AAP+05] adapted the well-established method of cognitive walkthroughs to this kind of use case. Their method consists of defining a persona, goals for the interaction and possible steps which can be taken in the visualization. With this setup an action is performed which seems most applicable to reach the current goal and afterwards the following four questions are answered:

1. What effect was the user trying to achieve by selecting this action?
2. How did the user know that this action was available?
3. Did the selected action achieve the desired effect?

4. When the action was selected, could the user determine how things were going?

Setup

The fictive user is a postdoctoral researcher at the Museum für Naturkunde Berlin. Their background is characterized by the following features:

- **Education** Is a postdoctoral researcher of biology specializing in evolutionary theory
- **Relevant work experience** Currently working on a project called "Variabilität von MHC-Genen bei der Sackflügelfledermaus Saccopterix bilineata" investing the genetic variability of bats
- **Experience with user interface design and usability assessment** Has no prior knowledge of interface design or usability assessment
- **Operating systems and software packages used frequently** Microsoft Windows; Microsoft Office (Word, Excel, PowerPoint); Microsoft Outlook; Mozilla Firefox; Microsoft Media Player; LaTeX; Zotero

As described in the Introduction, there are no common meeting rooms for the scientific staff at the museum. The interface is therefore positioned on a location which has the biggest throughput in the museum - in this case the side entrance which is exclusively used by the museum's staff. As discussed in the paragraph concerning interpretability in the Introduction, context is an essential component for interpretability. Therefore the fictive user interacts with the application in a well defined scenario:

> One day after work, the fictive user is coming down the wide stairway of the side building they work in and sees once more the display with the visualization they pass every day on their way to and from work. This time the curiosity is stronger than the urge to go home and since they already heard that the museum financed a huge initiative to foster intra-organizational, scientific exchange, they decide to see what that application has to offer.

Looking at the questions formulated in the beginning of this thesis in chapter 1, we can now derive specific tasks this user may want to complete to answer the questions:

1. Identify dominating research areas
2. Find their own project
3. Explore the projects in the same cluster

4. Explore the projects in the vicinity of their own cluster

The prototypical interface provides the following actions in the visualization:

1. Investigate metadata for a project (title, ID and top words)
2. Investigate top words for all clusters
3. Change the number of clusters
4. Switch between the linearized view and the scatter view

Cognitive Walkthrough

Simulating the full interaction, which is traceable in the appendix using the protocol (starting from page 43), reveals that all three implemented interpretability techniques can be used to help a fictive user understand the output of the system. The way they are used differ significantly on the other hand. The cluster top words were either used to quickly see what kind of topics prevail or to decide if an acceptable level of granularity was reached. If that was the case the user was able to pinpoint in which cluster the object of interest lies which reduces the search space nominally. The linearization was used only one time. After the user interacted with all the projects which were not hidden by clutter, they changed into the linearized view using this technique just as intended. During the cognitive walkthrough the cluster topography was of less relevance. The user primarily used this interpretability technique to generate candidates for the next inspections via top words and not gain a global view over the clustering.

This speaks in favor of the hypothesis that interpretability techniques can be reinterpreted with another explanation strategy, as it happened with the cluster topography. This suggests that there is not, as initially assumed, a 1:n correspondence between explanation strategies and interpretability techniques, but rather a n:n connection. Since the reinterpretation is a context-dependent step, the cognitive walkthrough, as a context-light technique, may not be the best suited method for investigating the way users interpret the output of such systems. In order to research such dependencies it would be necessary to conduct a full user study in the particular environmental and mental context of the use case.

Furthermore it unveiled a number of usability issues with the visualization (Table 4.4). The first found problem is dismissable since the presented interface is not the final one which is going to go live, but the second may hinder the interaction with the visualization even in the actual prototype.

Description	Usability Impact
The label for the dropdown selection between the scatter plot and the linearized view does not properly describes what it does.	Uncertainty about the usage of a tool may disturb a user in the inference task, therefore a descriptive name for this selection should be chosen.
There is no visual connection between views while changing the cluster or view parameters.	Perturbing these parameters does not change the underlying displayed corpus, but the recomputation of the full pipeline may lead, due to the random initialization of the K-Means algorithm, to dramatically different outputs. Tracking these changes is quite hard and therefore after each change the user has to orient himself in the visualization anew. Adding animated transitions could help alleviating this problems by introducing object permanence in the views.

Table 4.4: Table summarizing the found usability design issues

5 Conclusion

As stated in the Introduction, this thesis was conducted in order to study what kind of interpretability techniques for NLP exist and how they could support a non-technical expert in understanding the output from the system.

An exploratory workshops with the researchers from project IKON laid the groundwork for this thesis by defining key questions and a theoretical model of how interpretative processes may form in non-technical experts. The proposed theory puts the context of a interpretative context and connects high-level explanation strategies to concrete explanations via algorithmic interpretability techniques.

The first step was a systematic literature mapping study according to Petersen et al. [PFMM] which enabled me to confirm the findings of Lipton [Lip16] and Miller [Mil17] in the domain of NLP. Furthermore the results from the literature mapping study suggest that most of the current research focuses on supervised methods, such as neural networks, and these models are mainly made interpretable through local instance explanations. A proper definition of interpretability or an analysis of how a method influences interpretability lacks in a majority of publications.

Based on these findings I was now able to take each component of the general topic extraction pipeline in Figure 1.2 and propose and implement a contending method. Each method was evaluated according to standard measures in order to ensure proper performance.

Having the choice between two models for each step still not enabled me to implement any method from the sourced literature. Therefore I developed three techniques which follow the strategies found in the literature and are tailored to the use case of IKON. The first technique (Top Words) explains projects or clusters locally by supplying the user with the most influential words connected a project or a cluster. The second technique (Cluster Topography) allows the user to make global deductions concerning the fitness of each project in its assigned cluster. In order to do this a similarity measure is calculated in the latent topic space and is visualized in 2D by interpolating a relief on the corresponding scatter plot. The last implemented technique solves the long-standing problem of clutter in scatter graphs in this use case. Since the distances in 2D have only limited expressiveness due to the performed dimensionality reductions the scatter plot is mapped onto a grid structure. This allows to add additional metadata more easily. Additionally this technique synergizes well with the cluster topography since cluttered relief changes are also expanded and are more easily to grasp.

Using the top words and coherence scores I measured the performance of the quantitatively and made a model selection based on these computed scores.

Ultimately, the previously existing topic modeling pipeline appeared to be superior to the newly developed techniques. A cognitive walkthrough simulating a researcher performing an exploratory interaction unveiled a number of usability issues, but also showed how the implemented techniques support the user in making inferences about the output of the topic modeling pipeline. In contrast to what was expected I found out that the output of these interpretability techniques can be reinterpreted hinting at the possibility that one technique may support multiple strategies. As a consequence, context-less or context-light usability techniques, as cognitive walkthroughs, may not be well suited to study the impact of interpretability techniques on the interpretation of non-technical experts since the concrete materialization of the explanation technique depends heavily on the context.

5.1 Outlook

As discussed in chapter 2 and visible in the results of the literature mapping study, there are a number of additional explanation strategies which could be applied to the augmented topic modeling pipeline.

Although the performance of Agglomerative Clustering didn't seem to satisfy the needs of the application, the idea of explaining a model by an induced taxonomy is still very interesting [LHLH18]. Factoring in that the majority of the staff at the museum are trained biologists and taxonomies are widely used in this scientific discipline, these structures may be a very useful metaphor to present information to these non-technical experts.

Furthermore during the work on this thesis another potential question, additional to the ones defined in Table 1.1, arose:

> What kind of potential projects exist in the space between projects?

One of the already used techniques could be used to deliver potential answers . If the current LDA reduction gets replaced by a special kind of autoencoding, called variational autoencoding (VAE), it should be possible to generate meaningful vectors in the latent topic space and via the previously discussed methods also top words for these potential projects.

Asides from additional interpretability strategies and further model tuning, the whole system needs to be subjected to complete and rigorous user test with non-technical experts from the museum. The cognitive walkthrough included in this thesis does deliver a few insights into the usability of the application and the interaction with the topic modeling pipeline, but only a test in the situational context of the environment of the museum can convey reliable information concerning the interpretability of the used algorithms and the inferences the users are able to make using the system.

Accordingly, and based in part on the work of this thesis, the researchers in project IKON are currently developing qualitative research methods to study the contextual interpretability of the topic modeling pipeline. This may show

how reinterpretation of interpretability techniques, and possible refinement of the developed interpretability techniques of this thesis, may be taken up in future work.

Appendix

5.1.1 Protocol of the cognitive walkthrough

1. Step
    - *Which action was selected?*
      2
    - *What effect was the user trying to achieve by selecting this action?*
      The user sees the visualization for the first time and tries to connect the cluster top words, which are displayed above the visualization and the clusters in the visualization.
    - *How did the user know that this action was available?*
      This action follows from the immediate presentation of top words and scatter plot.
    - *Did the selected action achieve the desired effect?*
      The user deduces that all the projects can be clustered into three clusters - a taxonomic cluster, one connected to ecosystems and one concerned with evolution.
    - *When the action was selected, could the user determine how things were going?*
      This action did not change the visualization.
    - *Which interpretability technique was used?*
      Top words

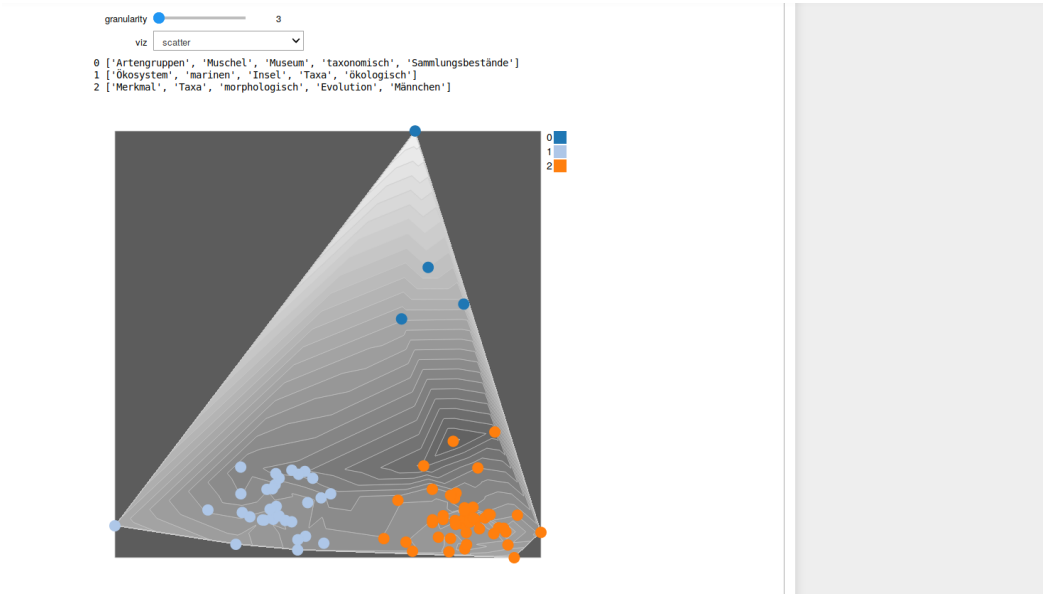

Figure 5.2: Cognitive Walkthrough step 1

2. Step

- *Which action was selected?*
  3
- *What effect was the user trying to achieve by selecting this action?*
  The user concluded that his project must be in the 'evolution' cluster and therefore they change the granularity by moving the slider to the middle of the selection range.
- *How did the user know that this action was available?*
  It was the only available slider and its label suggested that this is the proper action.
- *Did the selected action achieve the desired effect?*
  The user sees a more granular clustering over all projects.
- *When the action was selected, could the user determine how things were going?*
  Since there is no transition, the user does not know what is happening.
- *Which interpretability technique was used?*
  Top words

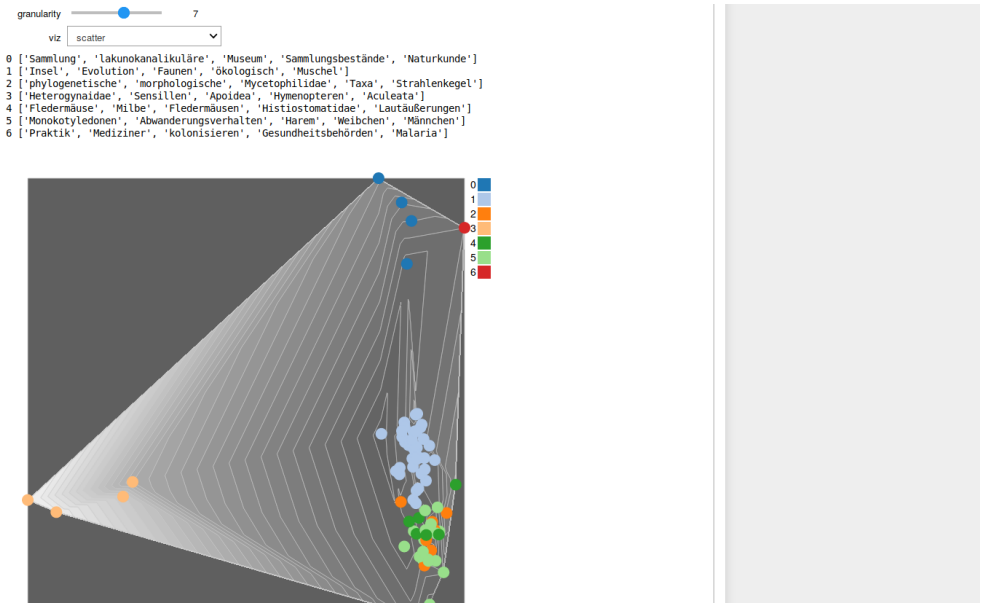

Figure 5.3: Cognitive Walkthrough step 2

3. Step
    - *Which action was selected?*
      2
    - *What effect was the user trying to achieve by selecting this action?*
      After changing the granularity, the user is trying to pinpoint the cluster to which his project is now assigned.
    - *How did the user know that this action was available?*
      As in the beginning the immediate presentation of the topwords and the visualization makes it inevitable to read and connect both.
    - *Did the selected action achieve the desired effect?*
      The user determined that his project is probably in cluster 4.
    - *When the action was selected, could the user determine how things were going?*
      This action did not change the visualization.
    - *Which interpretability technique was used?*
      None

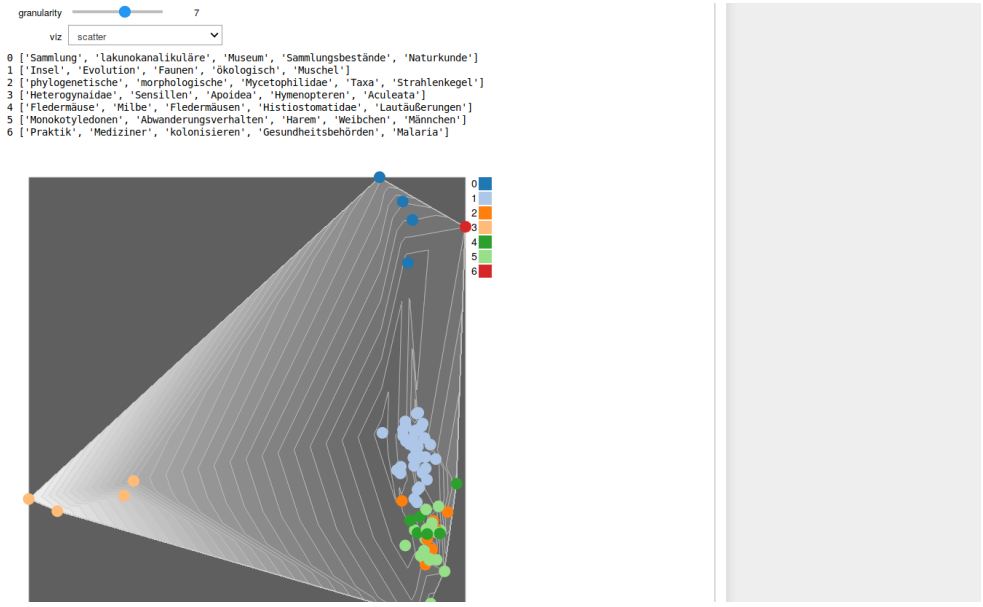

Figure 5.4: Cognitive Walkthrough step 3

4. Step

- *Which action was selected?*
  1
- *What effect was the user trying to achieve by selecting this action?*
  The user is trying to find his project in cluster 4.
- *How did the user know that this action was available?*
  Hovering over a glyph to display metadata is a common design strategy and therefore the user tried this first.
- *Did the selected action achieve the desired effect?*
  The selected project was not the correct one.
- *When the action was selected, could the user determine how things were going?*
  Since the metadata is displayed right alongside the project and the rest didn't change, there was no confusion.
- *Which interpretability technique was used?*
  Top words

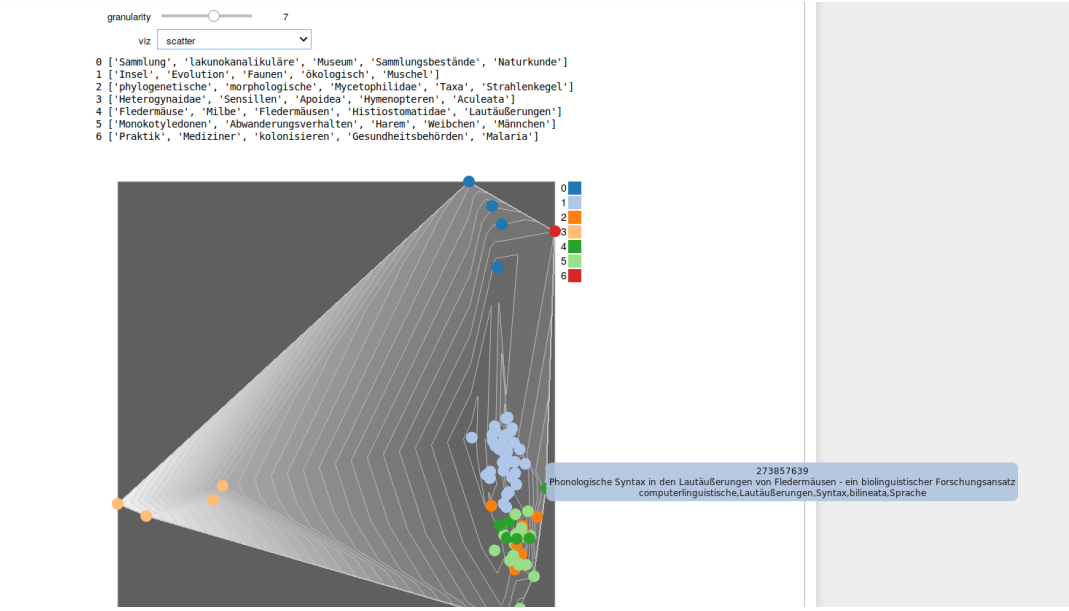

Figure 5.5: Cognitive Walkthrough step 4

5. Step

- *Which action was selected?*
  1
- *What effect was the user trying to achieve by selecting this action?*
  The user is trying to find his project in cluster 4.
- *How did the user know that this action was available?*
  Hovering over a glyph to display metadata is a common design strategy and therefore the user tried this first.
- *Did the selected action achieve the desired effect?*
  The selected project was again not the correct one.
- *When the action was selected, could the user determine how things were going?*
  Since the metadata is displayed right alongside the project and the rest didn't change, there was no confusion.
- *Which interpretability technique was used?*
  Top words

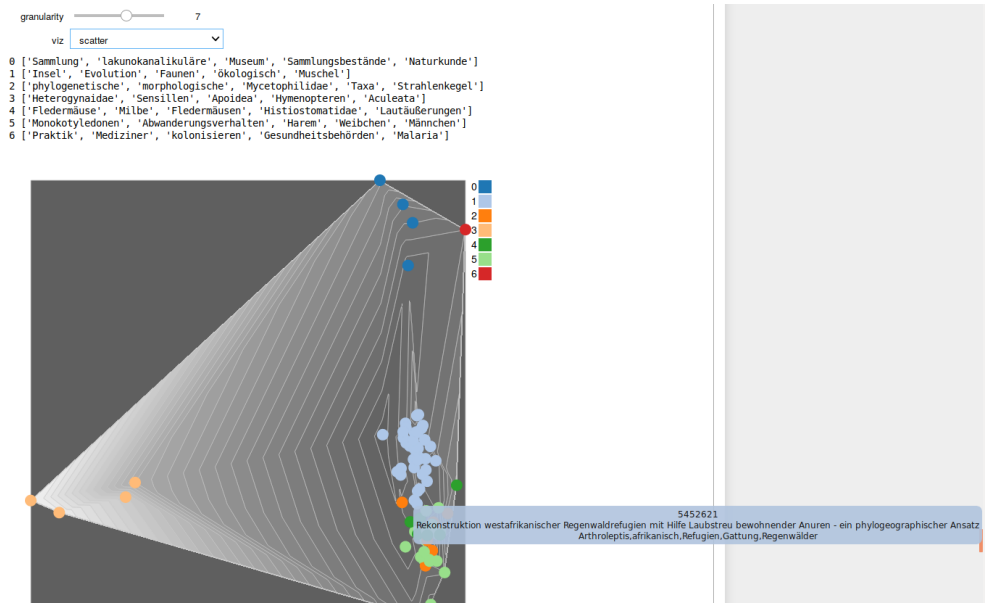

Figure 5.6: Cognitive Walkthrough step 5

6. Step

- *Which action was selected?*
  4
- *What effect was the user trying to achieve by selecting this action?*
  Since the rest of the cluster is extremely cluttered, the user decides to switch into the linearized view
- *How did the user know that this action was available?*
  This is the only remaining interaction option, but the naming of this selection makes it hard to intuitivly understand what it does.
- *Did the selected action achieve the desired effect?*
  The scatter plot got uncluttered by linearization.
- *When the action was selected, could the user determine how things were going?*
  Since there is no animated transition, the user does not really know how the scatter plot and this view are connected. Furthermore the cluster assignments and the top words changed, because the whole pipeline was computed from ground up. This adds into the confusion.
- *Which interpretability technique was used?*
  Linearization

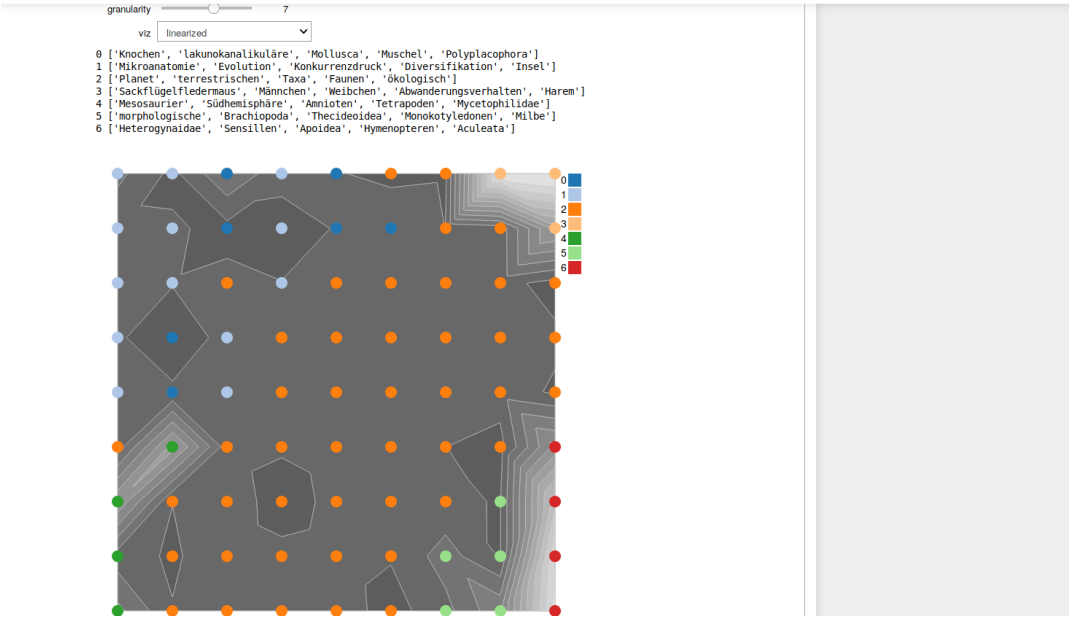

Figure 5.7: Cognitive Walkthrough step 6

7. Step

- *Which action was selected?*
  2
- *What effect was the user trying to achieve by selecting this action?*
  Now that the clustering and assignments changed again, the user is again searching for the cluster in which his project may lie.
- *How did the user know that this action was available?*
  As in the previous instances of this action this is the only available, logical action.
- *Did the selected action achieve the desired effect?*
  The user is able to pinpoint cluster 3 as the cluster connected to bats.
- *When the action was selected, could the user determine how things were going?*
  Nothing changed, therefore there was no room for confusion.
- *Which interpretability technique was used?*
  Top words

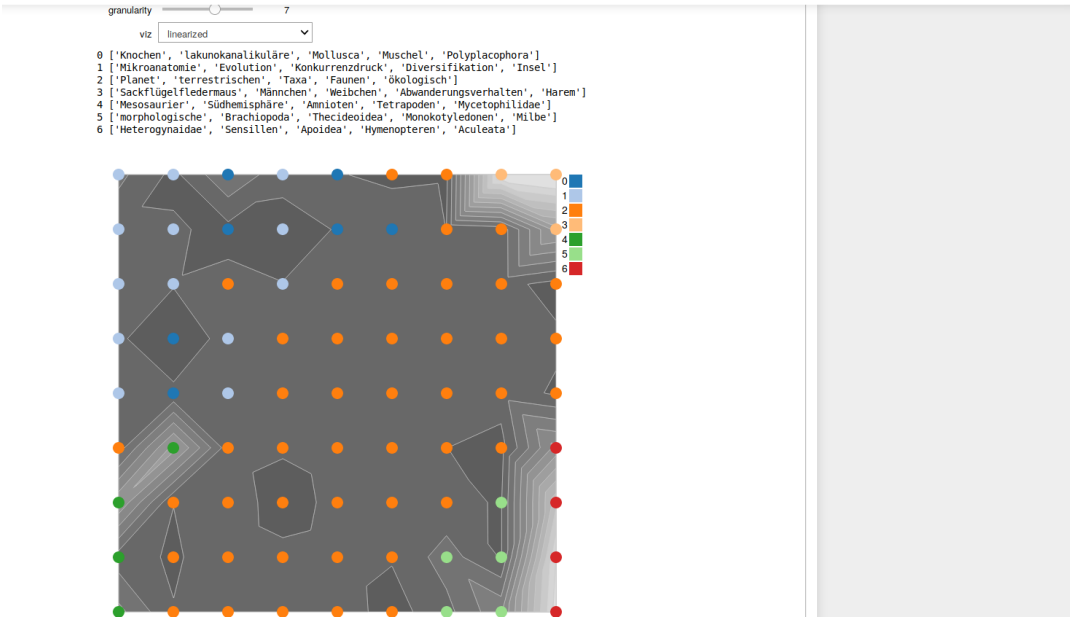

Figure 5.8: Cognitive Walkthrough step 7

8. Step

- *Which action was selected?*
  1
- *What effect was the user trying to achieve by selecting this action?*
  The user is searching for his project and selects the first visible glyph.
- *How did the user know that this action was available?*
  In the previous steps the user verified that hovering for metadata is a possibility.
- *Did the selected action achieve the desired effect?*
  The project they selected was indeed his own project. All of the top words make sense in the context of his project.
- *When the action was selected, could the user determine how things were going?*
  Again there was no confusion.
- *Which interpretability technique was used?*
  Top words

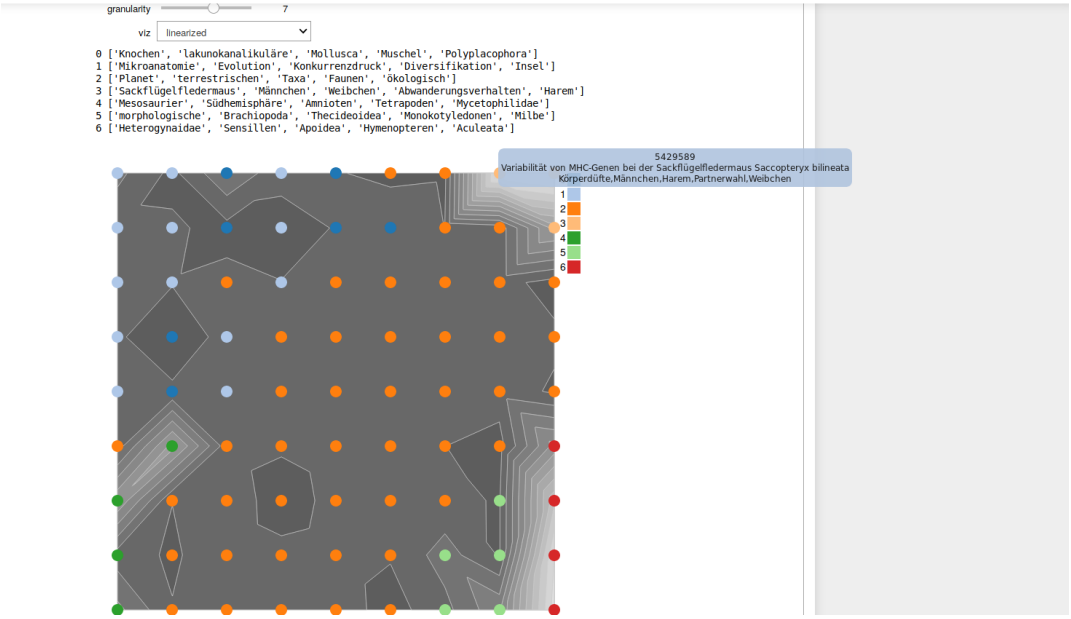

Figure 5.9: Cognitive Walkthrough step 8

9. Step

- *Which action was selected?*
  1
- *What effect was the user trying to achieve by selecting this action?*
  Now that they found his project the user is interested what kind of projects are also in the cluster which was assigned to his project. Therefore they select the next available project in the same cluster.
- *How did the user know that this action was available?*
  In the previous steps the user verified that hovering for metadata is a possibility.
- *Did the selected action achieve the desired effect?*
  The next project is also connected to the very same research subject. Therefore the clustering makes sense and the top words also are closely related to the top words of his own project.
- *When the action was selected, could the user determine how things were going?*
  Again there was no confusion.
- *Which interpretability technique was used?*
  Top words

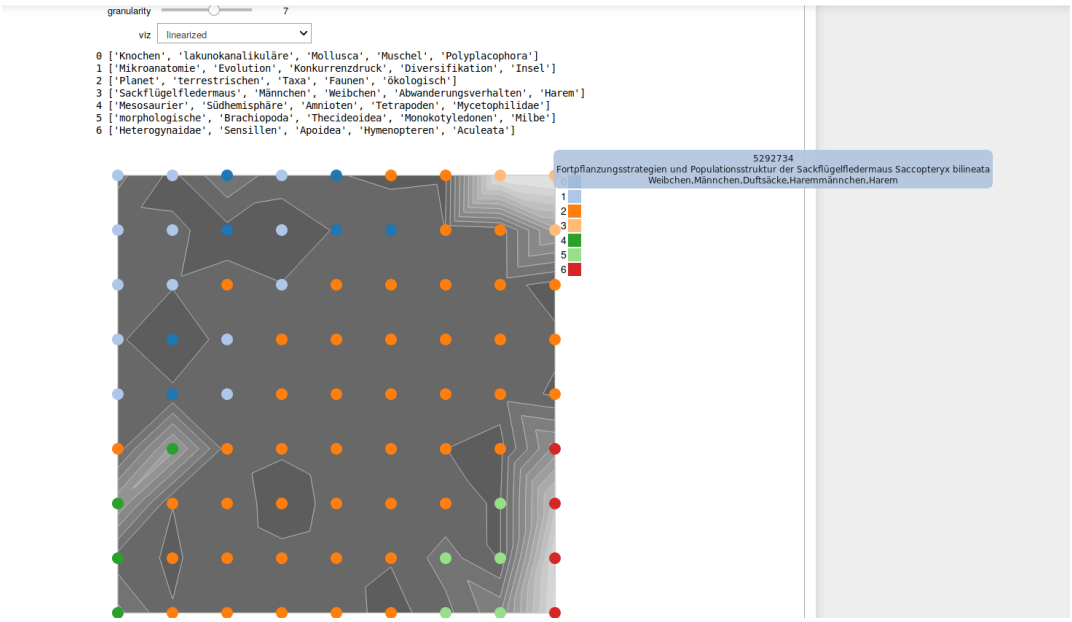

Figure 5.10: Cognitive Walkthrough step 9

10. Step

- *Which action was selected?*
  1
- *What effect was the user trying to achieve by selecting this action?*
  The user selects the last remaining project in the same cluster to see if it also fits into the cluster since the topography suggests that it may fit less well into the overarching topic.
- *How did the user know that this action was available?*
  In the previous steps the user verified that hovering for metadata is a possibility.
- *Did the selected action achieve the desired effect?*
  The last project is also connected to a similar research subject, although it differs a bit due to it rather being concerned with migration of bats than procreation. Since the topwords also suggest this, it further validates the clustering.
- *When the action was selected, could the user determine how things were going?*
  Again there was no confusion.
- *Which interpretability technique was used?*
  Top words, Cluster Topography

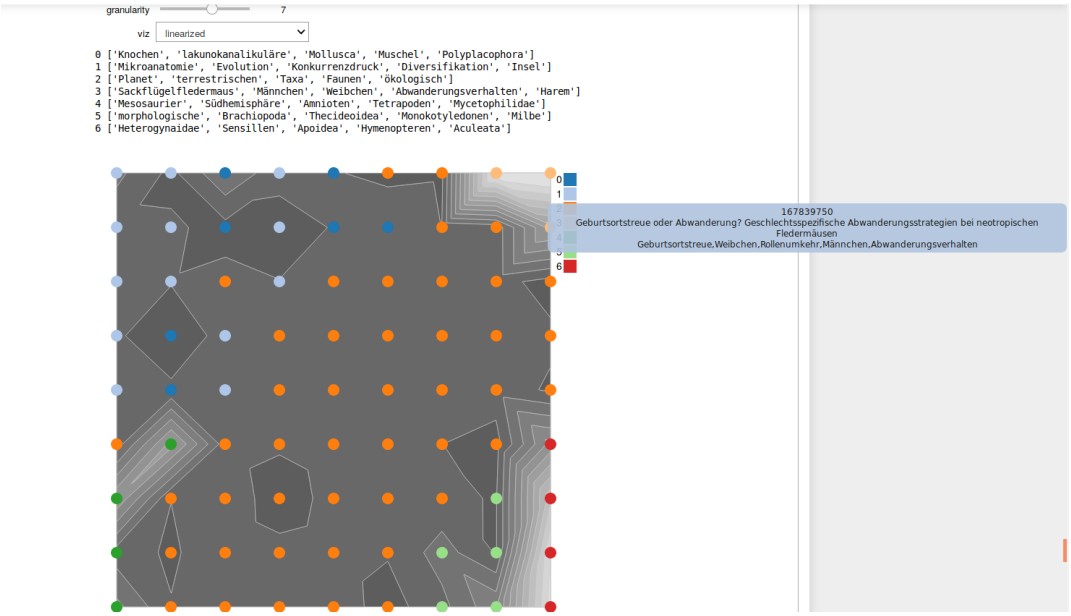

Figure 5.11: Cognitive Walkthrough step 10

11. Step

- *Which action was selected?*
  1
- *What effect was the user trying to achieve by selecting this action?*
  Since the cluster does make sense the user decides to have a look at the neighbouring projects. Especially since the selected one seems to fit worse into its cluster than the rest.
- *How did the user know that this action was available?*
  In the previous steps the user verified that hovering for metadata is a possibility.
- *Did the selected action achieve the desired effect?*
  The first project they select does investigate a completly different field.
- *When the action was selected, could the user determine how things were going?*
  The user is able to tell why the project lies in another cluster using the top words.
- *Which interpretability technique was used?*
  Top words, Cluster topography

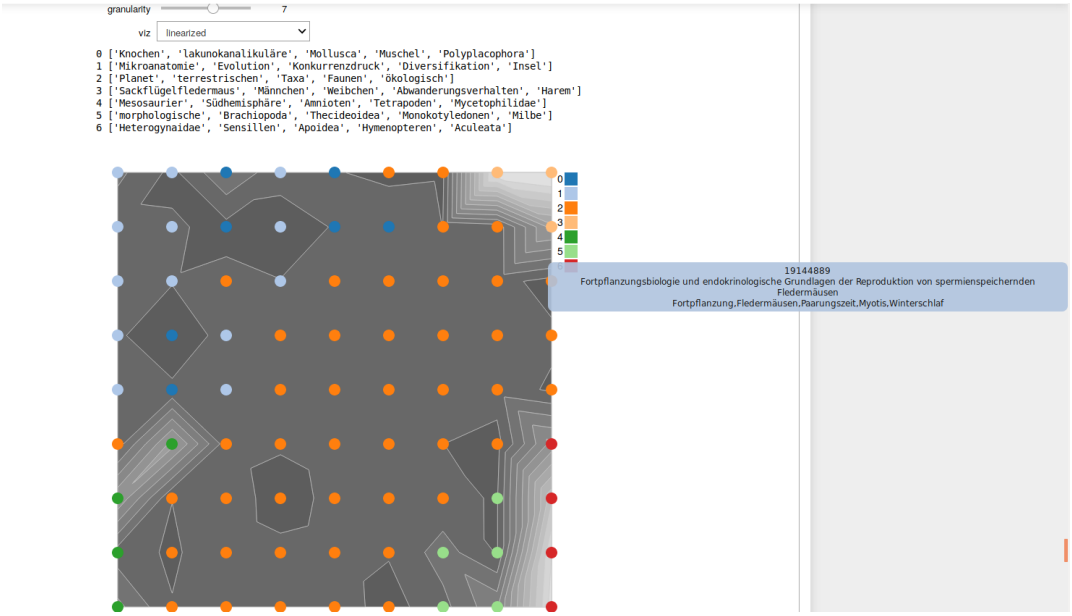

Figure 5.12: Cognitive Walkthrough step 11

12. Step

- *Which action was selected?*
  1
- *What effect was the user trying to achieve by selecting this action?*
  The user still thinks that there may be another similar project, because the top words of cluster 2 are a concerned with ecology, faunas and taxonomies.
- *How did the user know that this action was available?*
  In the previous steps the user verified that hovering for metadata is a possibility.
- *Did the selected action achieve the desired effect?*
  The next project they selects is surprisingly also connected to bats.
- *When the action was selected, could the user determine how things were going?*
  The user is not entirely sure why this project was not categorized in his own cluster. The top words suggest that the work is rather specialzed on the biological processes of procreation using bats as a use case.
- *Which interpretability technique was used?*
  Top words

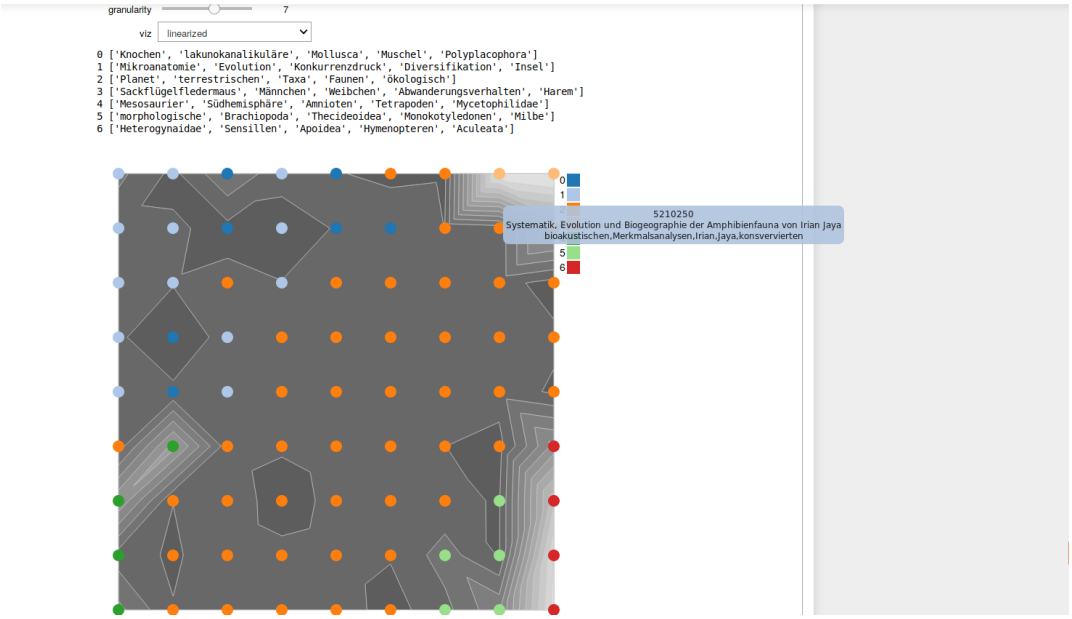

Figure 5.13: Cognitive Walkthrough step 12

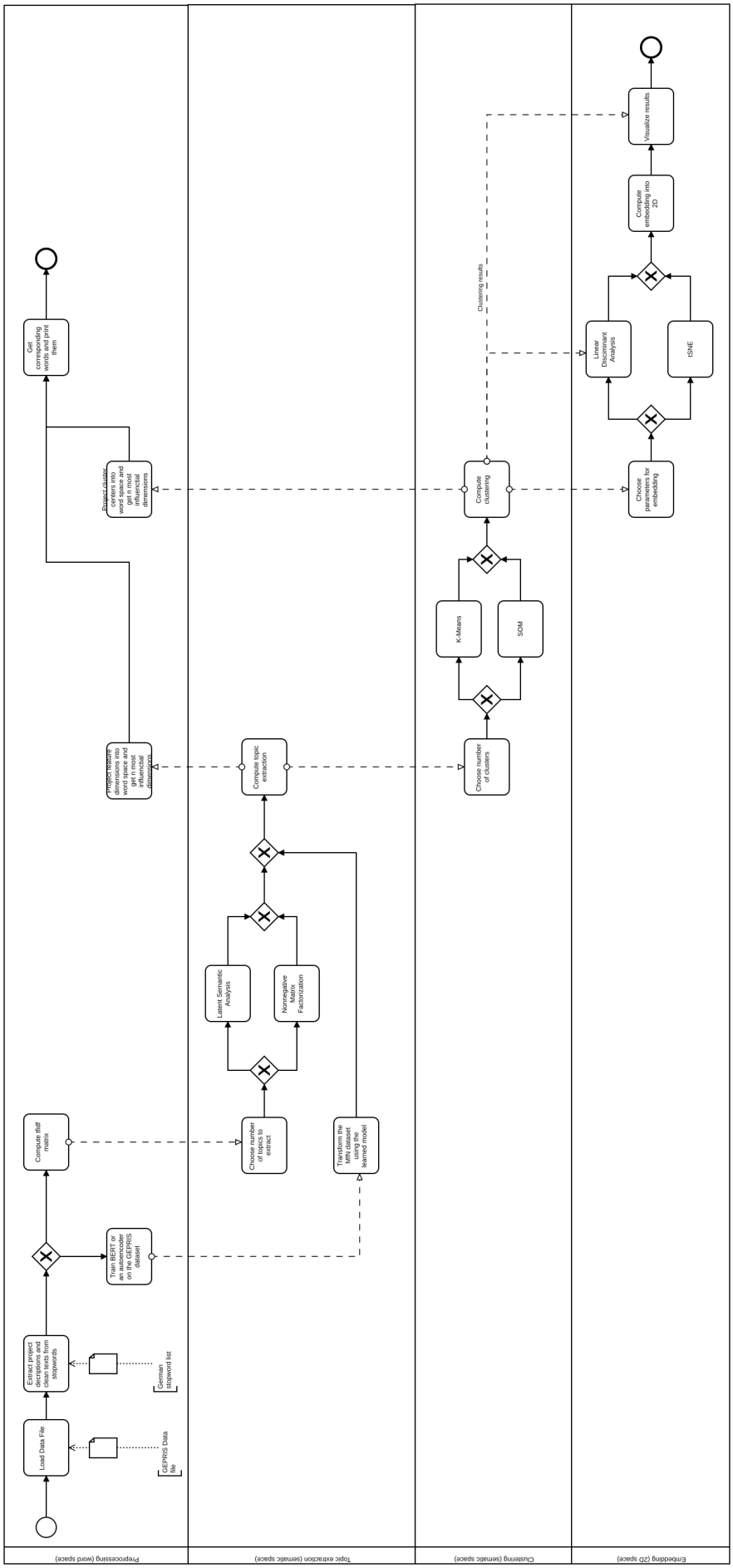

Figure 5.1: BPMN process diagram of the existing topic modeling pipeline

Bibliography

[14018] Ein 140 Jahre altes Baukonzept bereitet den Weg zum Naturkundemuseum des 21. Jahrhunderts. https://www.museumfuernaturkunde.berlin/de/about/bau/ein-140-jahre-altes-baukonzept-bereitet-den-weg-zum-naturkundemuseum-des-21-jahrhunderts, May 2018.

[AAP+05] K. Allendoerfer, S. Aluker, G. Panjwani, J. Proctor, D. Sturtz, M. Vukovic, and Chaomei Chen. Adapting the cognitive walkthrough method to assess the usability of a knowledge domain visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pages 195–202, October 2005.

[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.

[AHM+17] Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. "What is Relevant in a Text Document?": An Interpretable Machine Learning Approach. *PLOS ONE*, 12(8):e0181142, August 2017.

[BMG18] Jesse Benjamin, Claudia Müller-Birn, and Rony Ginosar. Transparency and the Mediation of Meaning in Algorithmic Systems. October 2018.

[BMKng] Jesse Josua Benjamin, Claudia Müller-Birn, and Christoph Kinkeldey. Understanding Knowledge Transfer Activities at a Research Institution through Semi-Structured Interviews. Technical Report TR-B-19-02, Freie Universität Berlin, Berlin, forthcoming.

[CTJ19] Christoph Kinkeldey, Tim Korjakow, and Jesse Josua Benjamin. Towards Supporting Interpretability of Clustering Results with Uncertainty Visualization. In *TrustVis19*, June 2019.

[DBVCDD16] Cedric De Boom, Steven Van Canneyt, Thomas Demeester, and Bart Dhoedt. Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recognition Letters*, 80:150–156, September 2016.

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, October 2018.

[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.

[Dee] Deep Phenomena. https://deep-phenomena.web.app/.

[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.

[GCJC] Yuxia Geng, Jiaoyan Chen, Ernesto Jimenez-Ruiz, and Huajun Chen. Human-centric Transfer Learning Explanation via Knowledge Graph. page 4.

[GMPB16] Yash Goyal, Akrit Mohapatra, Devi Parikh, and Dhruv Batra. Towards Transparent AI Systems: Interpreting Visual Question Answering Models. *arXiv:1608.08974 [cs]*, August 2016.

[HHC+19] Fred Hohman, Andrew Head, Rich Caruana, Rob DeLine, and Steven Drucker. Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. January 2019.

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

[IST+18] Tomoki Ito, Hiroki Sakaji, Kota Tsubouchi, Kiyoshi Izumi, and Tatsuo Yamashita. Text-Visualizing Neural Network Model: Understanding Online Financial Textual Data. In Dinh Phung, Vincent S. Tseng, Geoffrey I. Webb, Bao Ho, Mohadeseh Ganji, and Lida Rashidi, editors, *Advances in Knowledge Discovery and Data Mining*, Lecture Notes in Computer Science, pages 247–259. Springer International Publishing, 2018.

[JV87] R. Jonker and A. Volgenant. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340, December 1987.

[LHLH18] Ninghao Liu, Xiao Huang, Jundong Li, and Xia Hu. On Interpretation of Network Embedding via Taxonomy Induction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 1812–1820. ACM, 2018.

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

[LM14] Quoc V. Le and Tomas Mikolov. Distributed Representations of Sentences and Documents. *arXiv:1405.4053 [cs]*, May 2014.

[LTD+16] Lin Liu, Lin Tang, Wen Dong, Shaowen Yao, and Wei Zhou. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, 5(1):1608, September 2016.

[MG13] A. Mayorga and M. Gleicher. Splatterplots: Overcoming Overdraw in Scatter Plots. *IEEE Transactions on Visualization and Computer Graphics*, 19(9):1526–1538, September 2013.

[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.

[Mil17] Tim Miller. Explanation in Artificial Intelligence: Insights from the Social Sciences. *arXiv:1706.07269 [cs]*, June 2017.

[MSC+] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. page 9.

[MWM18] D. L. Marino, C. S. Wickramasinghe, and M. Manic. An Adversarial Approach for Explainable AI in Intrusion Detection Systems. In *IECON 2018 - 44th Annual Conference of the IEEE Industrial Electronics Society*, pages 3237–3243, October 2018.

[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/.

[RBH15] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search*

and Data Mining - WSDM '15, pages 399–408, Shanghai, China, 2015. ACM Press.

[Rob04] Stephen Robertson. Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*, 60(5):503–520, October 2004.

[Rou87] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, November 1987.

[Spa18] Spaude, Daniel. *Implementierung Einer Software Zum Regelmäßigen Crawling Der DFG Förderungsdaten Unter Besonderer Berücksichtigung Der Datenqualität.* PhD thesis, FU Berlin, Berlin, October 2018.

[Tea18] Team. https://www.museumfuernaturkunde.berlin/en/about/team, January 2018.

[TG99] I. A. Taha and J. Ghosh. Symbolic interpretation of artificial neural networks. *IEEE Transactions on Knowledge and Data Engineering*, 11(3):448–463, May 1999.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.

[WRLP94] Cathleen Wharton, John Rieman, Clayton Lewis, and Peter Polson. Usability Inspection Methods. pages 105–140. John Wiley & Sons, Inc., New York, NY, USA, 1994.

[WYX+18] Lingfei Wu, Ian E. H. Yen, Kun Xu, Fangli Xu, Avinash Balakrishnan, Pin-Yu Chen, Pradeep Ravikumar, and Michael J. Witbrock. Word Mover's Embedding: From Word2Vec to Document Embedding. *arXiv:1811.01713 [cs, stat]*, October 2018.

[WYZ16] Yasi Wang, Hongxun Yao, and Sicheng Zhao. Auto-encoder based dimensionality reduction. *Neurocomputing*, 184:232–242, April 2016.