Bachelorarbeit am Institut für Informatik der Freien Universität Berlin

Human-Centered Computing (HCC)

Comparing interpretability techniques for unsupervised topic modeling

Tim Korjakow

Matrikelnummer: 372862

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

Betreuer: Jesse Jonas Benjamin

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

Zweitgutachter: Prof. Dr. K. 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 2, 2019

<Name>

Abstract

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

Zusammenfassung

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

Contents

List of Figures

List of Tables

1 Introduction

1.1 Project IKON

This thesis has a direct application in a project which tries to explore potentials for knowledge transfer activities at a research museum. Project *IKON* was started in cooperation with the German Natural History Museum in Berlin which houses more than 600 [**TK:** Right number?] scientists, PhD students and other staff. With that size of scientific staff the institution is a global player in research on evolution and biodiversity [Int]. Despite its importance in the research landscape, the museum is challenged with a lack of shared knowledge across working groups and organizational structures such as departments. In interviews researchers from the project were able to trace these problems back to the very intricate and complex layout of rooms and halls in the building which was originally constructed in 1810 [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. Knowledge transfer is made explicitly visible by showing projects not in the predefined taxonomy of the museum, but instead in semantic relation to each other. This is accomplished by running all project abstracts through a topic modeling process consisting of four major components, as seen in Figure 1.2.

1.2 Topic modeling

A general topic modeling pipeline consists of four steps:

1. Document embedding
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

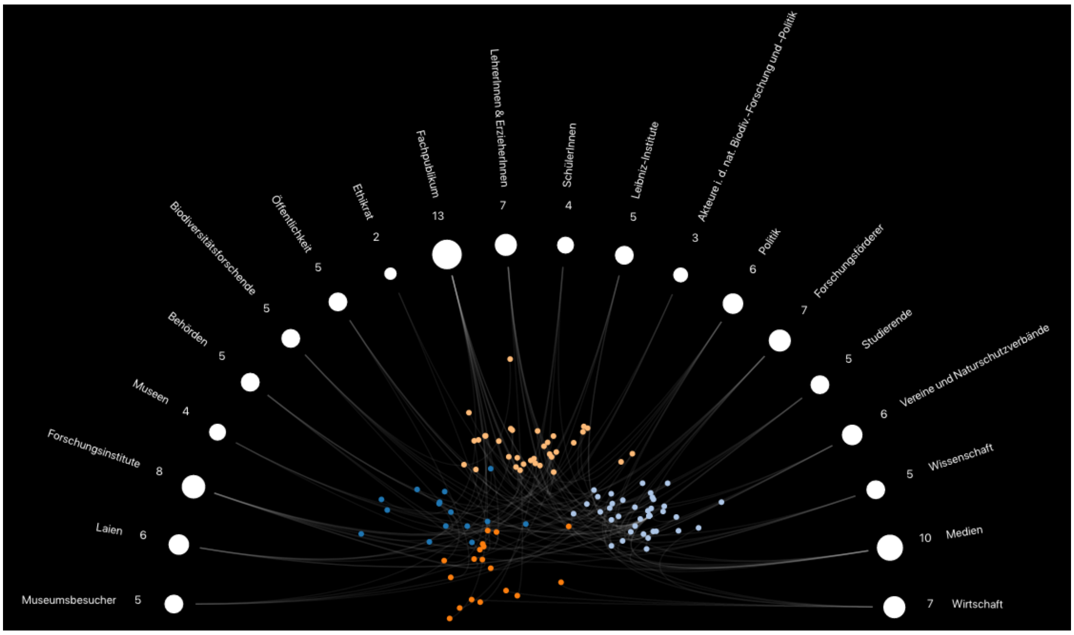


Figure 1.1: Screenshot of the cluster view of the IKON visualization

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.

Finally in order to visualize the high dimensional distribution of documents in the latent space another dimensionality reduction is used to project the documents to 2D.

Based in first interviews and conceptual work the researchers from project IKON hypothesize that the scientists from the museum, as non-technical experts, will have a hard time interpreting and understanding the output generated by the pipeline. Furthermore each component in Figure 1.2 introduces additional parameters which influence the results generated by the pipeline.

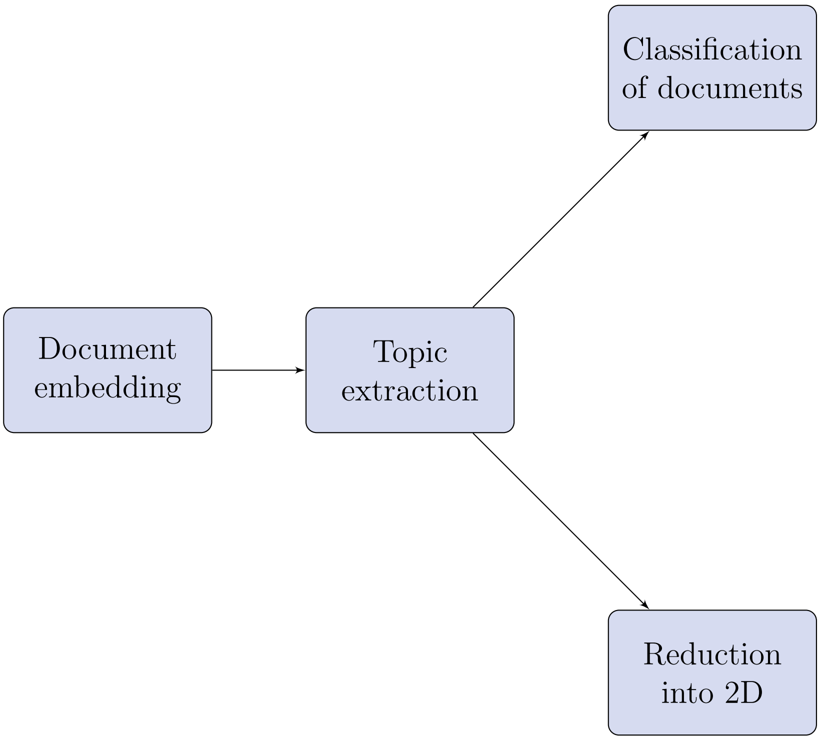


Figure 1.2: Components of a general topic extraction pipeline

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

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

1.3 Interpretability

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

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

2 Literature mapping study

2.1 Motivation

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

2.2 Methodology

The 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 explainability techniques are mentioned in the corpus?

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

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

2. Construction of a search string:

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

   (”explainability” OR ”explainable” OR ”explanation” OR ”explaining” OR ”interpretability” OR ”interpretable” OR ”interpretation” OR ”interpret” OR ”understanding”) AND (”machine learning” OR ”neural network” OR ”neural networks” OR ”AI” OR ”XAI” OR ”artificial intelligence” OR ”model”) AND (”text” OR ”document” OR ”NLP” OR ”nat-

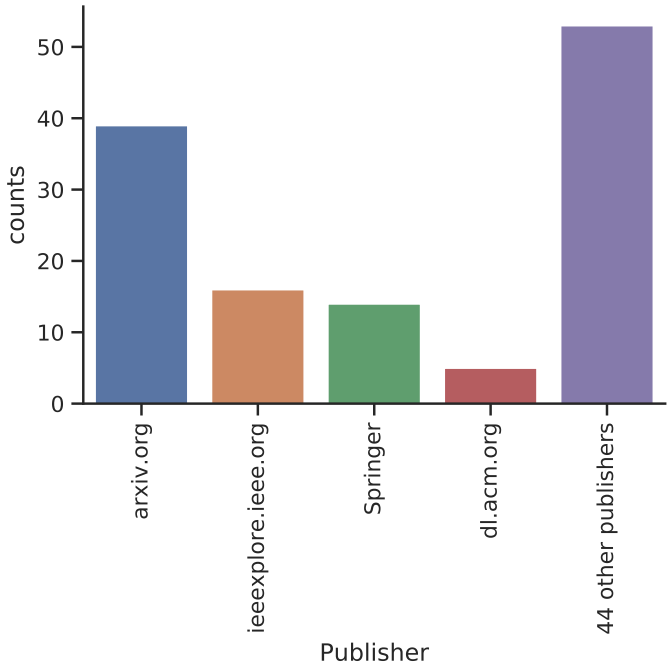


Figure 2.1: Barplot displaying the distribution of publishers occurring in the meta search results

ural language programming" OR "review" OR "method" OR "technique" OR "visualization")

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

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

4. Sourcing of publications in scientific databases:

   Based on the insights from the previous step each of the main publisher's databases is scraped using the search string and their respective 'advanced search' interfaces or their APIs. Since most searches result in

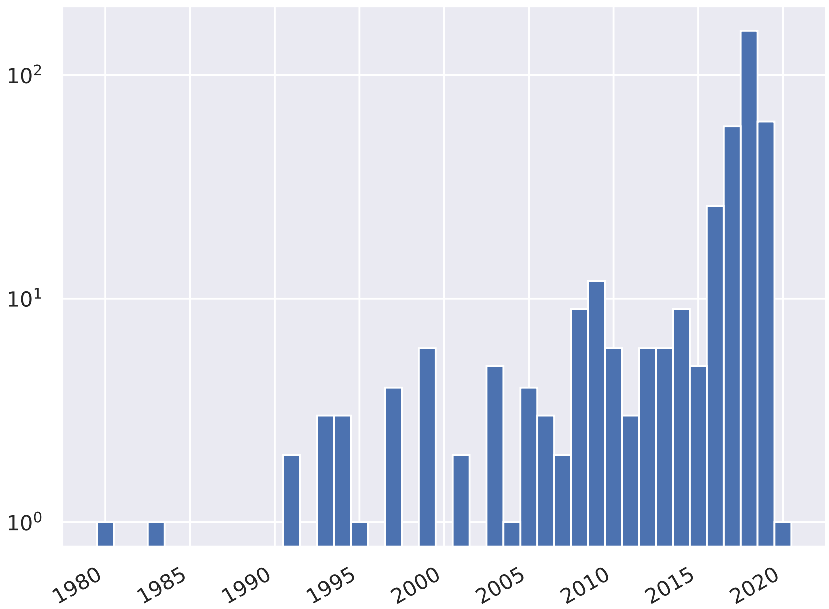


Figure 2.2: Barplot displaying the distribution of publishers occurring in the meta search results

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. Intermediate assessment of the corpus:

   Looking at the distribution of tags in Figure 2.3 it is apparent that the chosen keywords represent the field well. There are no tags in the first 5 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 1980, while there is a surge of interest and research in the last 4 years. This speaks in favor of the premise that interpretability research is not necessarily a young, but a recently thriving field.

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

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

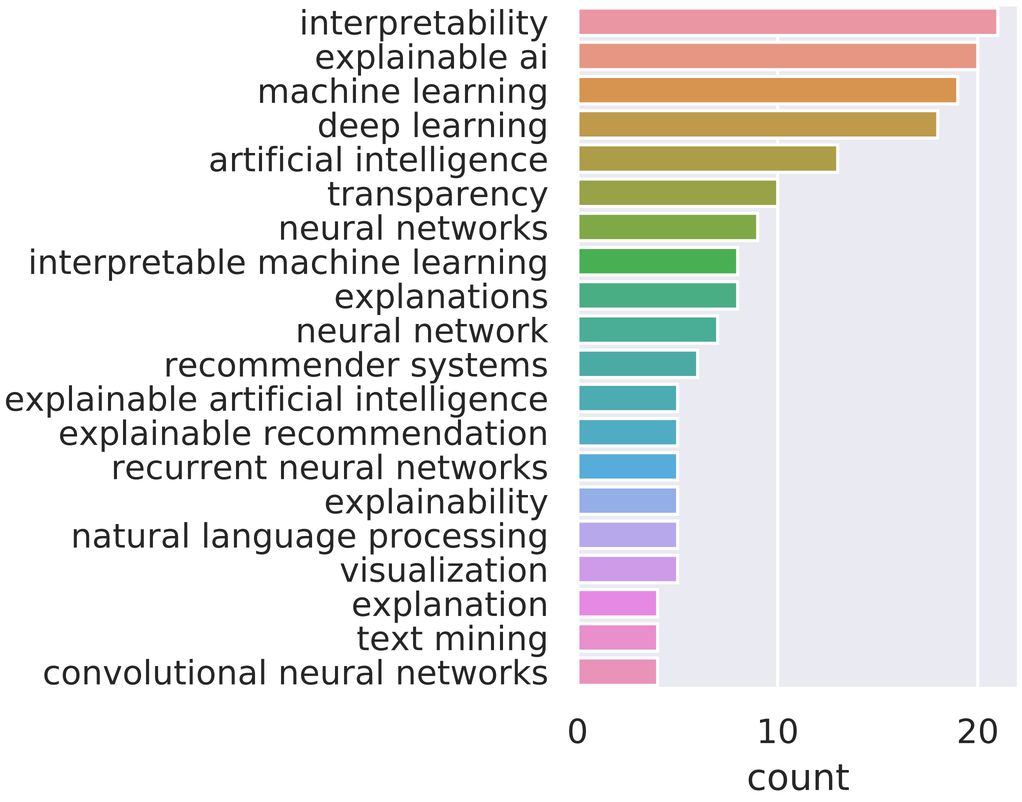


Figure 2.3: List of the 20 most used tags and their absolute frequency

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

Table 2.1: Table showing all used inclusion and exclusion criteria

port 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.

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. [?]. These categories are not a perfect fit for a thesis dealing with explainability for non-technical experts since it also categorizes techniques according to their mathematical inner workings, but Hohman et al. extended the 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 each publication was assigned 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 explainability technique and demonstrates its impact in an exemplary use case. A "Theory" paper does so as well, but misses a presented application and evaluation. A "Report" on the other hand summarizes and presents multiple techniques. Finally, a "Study" paper shows the results of an interface evaluation which visualizes the output of explainability methods. Publications from the last category are therefore less technical and more concerned with the HCI aspects of explainability techniques and their visualization.
>
> Since most of the overview papers presented a huge amount of techniques which were already covered by the "Method" papers and the corpus was already large, I decided to exclude them from the last mapping step. This reduced the final corpus to a size of 72 publications.

2.3 Results

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

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 explainability methods are developed as mathematical theories and influences from HCI are rarely taken into consideration.

The second question was concerned with the type of models which are enhanced by explainability 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 explainability 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. Suprisingly, 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 of the publications on the "NN" axis in Figure 2.5 fall into this bucket as well. All in all, 18 publications remain which could be applied to an unsupervised topic extraction pipeline. The document embedding step could be made interpretable by decomposition, feature importance visualization or by explaining

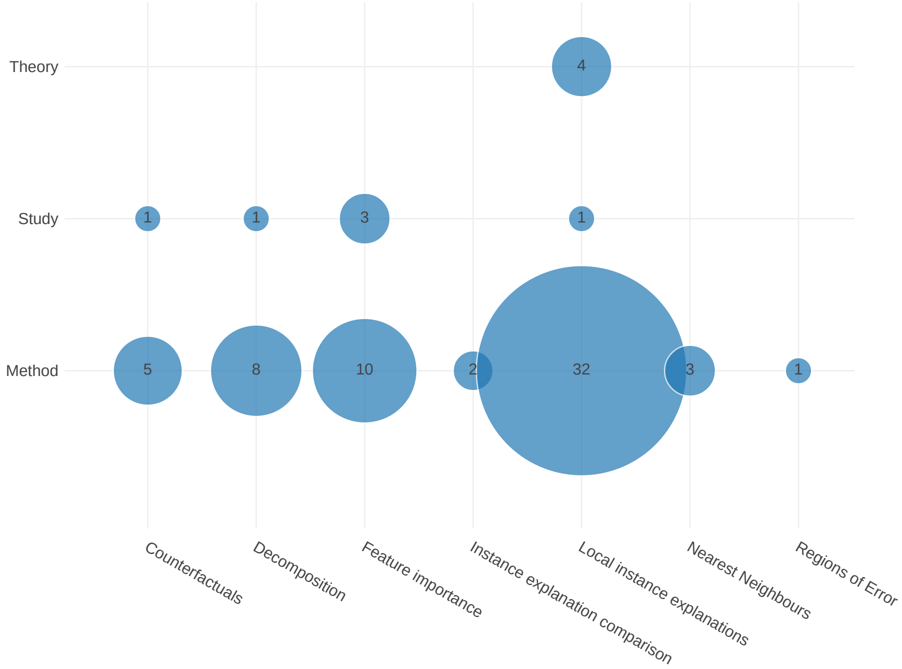


Figure 2.4: Mapping of the type of publication and its Gamut classification

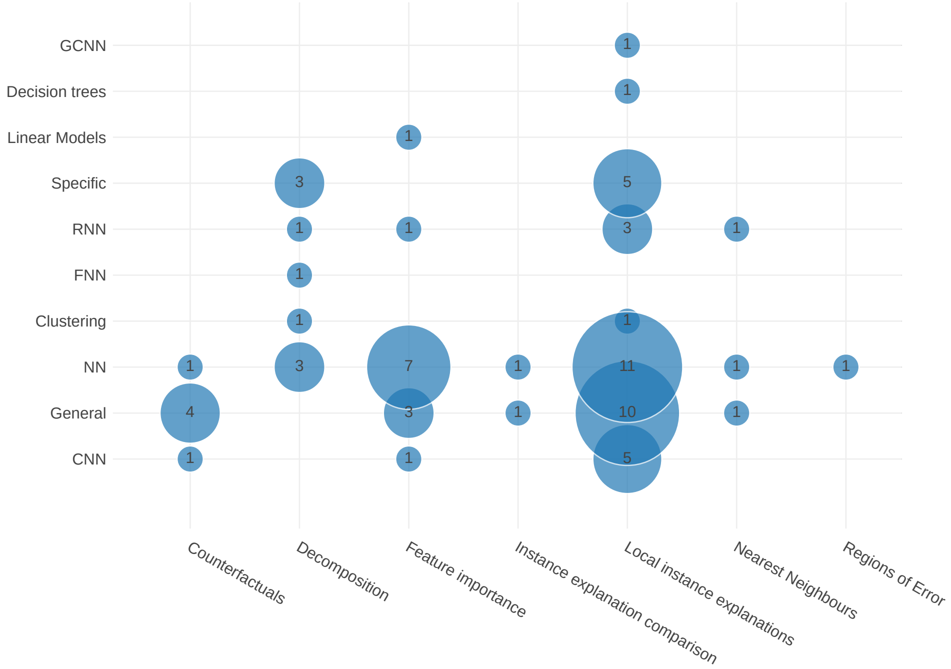


Figure 2.5: Mapping of applicability and Gamut classification

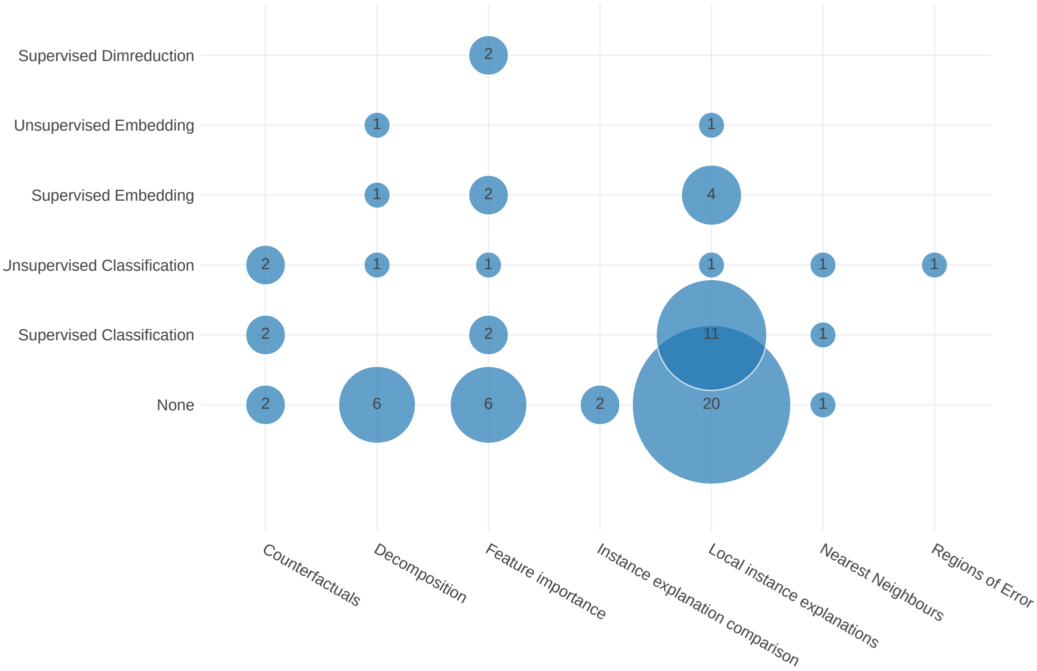


Figure 2.6: Mapping of pipeline step and Gamut classification

the embedding of single instances.

Kim et al. [KLHK19] decompose a pretrained network and extract simple features which they use as to train a neural network on another task in a transfer learning fashion. The predictions for new tasks can then be described as a combination of these extracted features.

In contrast to that, Zhang et al. [ZYL+18] train another neural network to explain the output of any given neural network in a unsupervised way. They focus on CNNs and utilize the fact that these convolutional layers contain structural information. For each input they are able to disentagle the information from the applied convolutional filters and extract features which can be applied back to the input as masks to show influential parts. Given a document embedding technique, which uses CNNs, and a corpus this explainability technique could be used to highlight influential parts of the input document.

3 Implementation

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 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. That's why I built all my work on top of the official Tensorflow Docker image which comes with all the drivers for NVidia GPUs. The problem with that image is that it either detects a graphic card and suitable drivers and works or it fails in case of no present drivers. For that reason I created a Shell which makes the use of the container easier. By activating a $--gpu$ flag while executing the script, the docker container dynamically selects an image with or without GPU support, starts the Notebook container and the database in a separate network and scans the output of the Notebook container in order to extract the Notebook credentials, which are necessary to access the program via the browser, and opens a browser session once these credentials were found.

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 taxonomy 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 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 approximately 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

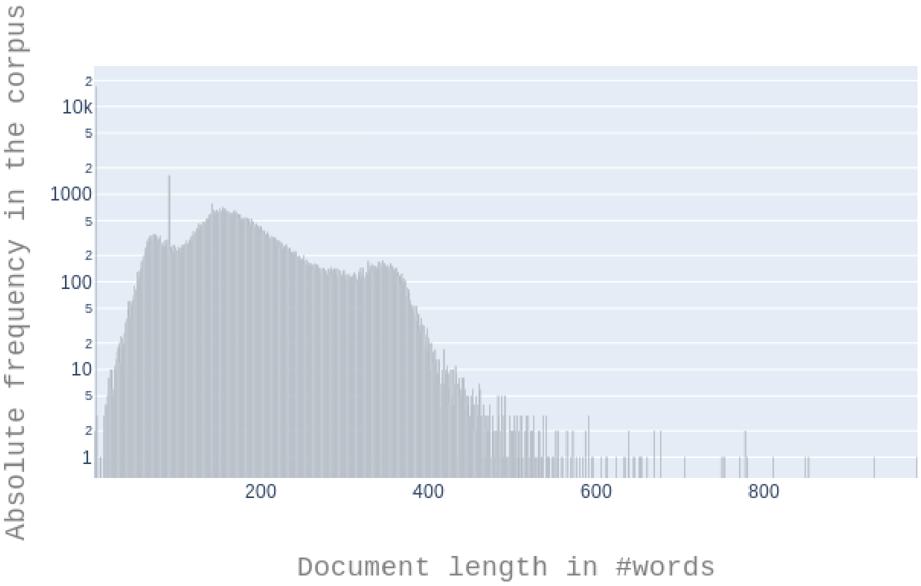


Figure 3.1: Histogram showing the distribution of text lengths in the dataset

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.

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

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

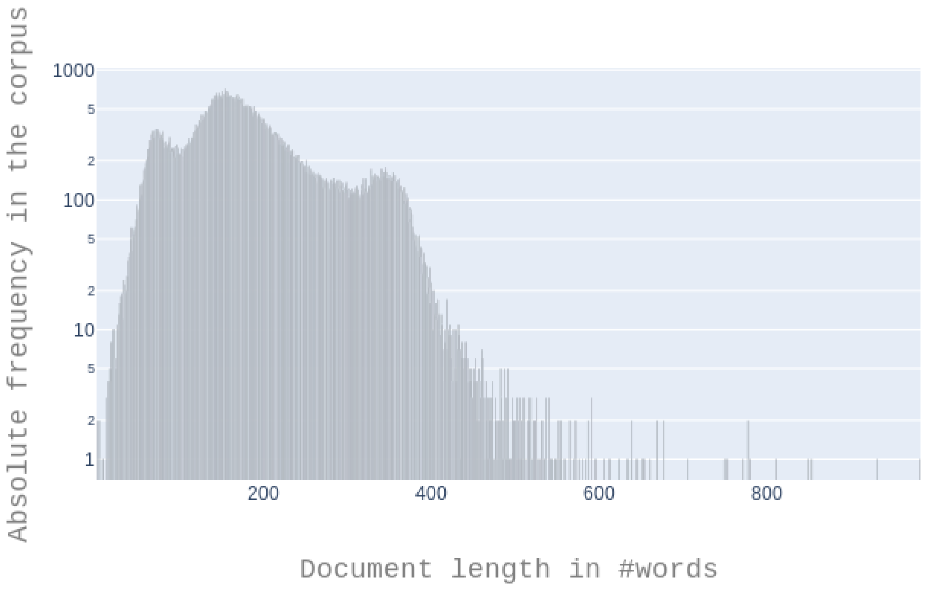


Figure 3.2: Histogram showing the distribution of text lengths in the dataset excluding duplicates and projects without a description

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

3.3 The existing pipeline

The existing pipeline was implemented by me as a proof-of-concept for project *IKON*. Following the structure of Figure 1.2 the first step is a document vectorization of the given texts in order to embed them in one common vector space. One of the simplest and still effective methods is a Tf-Idf Bag-Of-Words (TfIdf-BOW) embedding. With this procedure each text is represented as a set of terms, the bag of words. Having a whole corpus it is now possible to assign a vector to each document D in corpus $C = \{D_1, ..., D_n\}$ of length $N = |C|$, where each entry i is the number of term occurrences of term t_i

in D. That means that each document gets embedded into a vector space of dimensionality |(unique terms in C)| and the corpus becomes a matrix of size |(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. [**TK:** Needs maybe rework based on Shannon theory] This ensures that words which are specific to a small group of documents and appear often in them, get a higher weight, while terms which are infrequent or too frequent in many documents, as articles for example, get a small weight.

Now that there is a vector representation of each document, the clustering of the documents could be performed in this vector space using K-Means. The *the curse of dimensionality* suggests that this leads to to the clustering algorithm failing to perform. 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 Euclidian distance, becomes effectively meaningless and making it necessary to reduce the dimensionality of the vector space.

One popular method, which is often used in conjunction with Tf-Idf BOW embeddings, is the Latent Semantic Indexing (LSI), also known and henceforth referenced as Latent Semantic Analysis (LSA). A LSA operates on the premise that a vectorized corpus contains latent structures, which my correspond to topics for example. Such a topic would consist of several words which are semantically connected and therefore appear together more often than words which are not semantically similar. Adding constraints such as adjustable representational richness, which depicts sufficient parameterisation, explicit representation of both terms and documents and computational tractability for large datasets the authors decided to use a Singular Value Decomposition (SVD) [DDF$^+$]. The SVD is closely related to Principal Component Analysis (PCA) and reduces the dimensionality of a dataset by removing the dimensions with the least variance, effectively projecting the vector space onto the subspace with the highest variance and therefore the most information contained. Applying a SVD on the corpus changes the representation of the document from being a linear combination of words into being a linear combination of latent topics. This representation is now usable for most other methods such as clustering due to its smaller dimensionality. 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 Document embedding

3.4.1 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.

Another popular technique was published by Blei et al. [BNJ03] in 2003. *Latent Dirichlet Allocation* is a hierarchical Bayesian model, which describes documents as a finite mixture of latent topics, while topics are an infinite mixture of latent topic probabilities. The LDA therefore performs the embedding and the topic extraction step at once.

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 pretrained 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

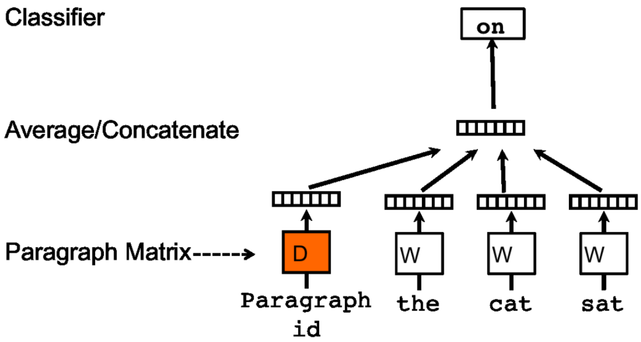


Figure 3.3: Visualization of a training step of a Doc2Vec network [WYX+18]

Technique	Max. document length	Type
Tf-Idf BOW	unlimited	Probabilistic
Latent Dirichlet Allocation	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

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.

3.4.2 Selection of a document embedding technique

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

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 analysis 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 more explainability techniques which may be developed in the future.

3.4.3 Explainability technique: 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 I will rank the input features to the document embedding and topic extraction step for each document and cluster.

Assume that I have a document 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 the idea is to go backwards. 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.

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. Although there is no literature on this, it is a hypothesis that this behaviour 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 shows 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 3.2 are indeed semantically well connected. The results of the document queries in Table 3.3 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.

Extracting top words for the existing TfIdf method is simpler, because the every entry in the embedding vector has a one-to-one correpondence 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.

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

Table 3.2: Table showing the top five similar words for two queries by word

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

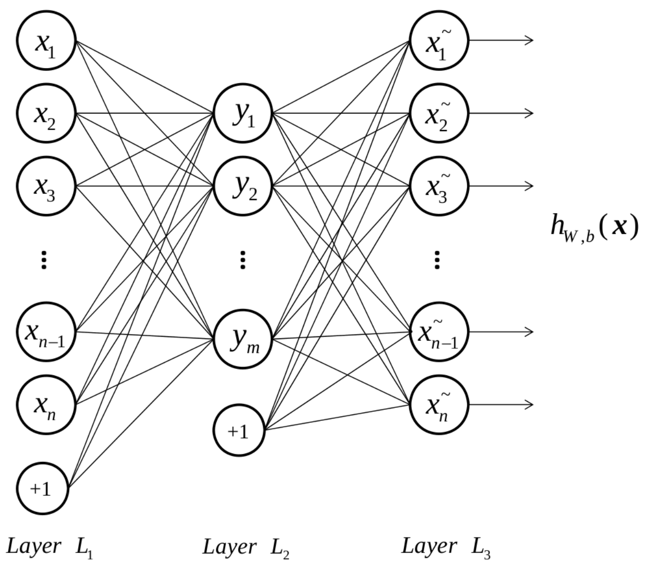


Figure 3.4: Strucuture of a simple autoencoder [WYZ16]

3.5 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 as seen in Figure 3.4. 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 3.5.

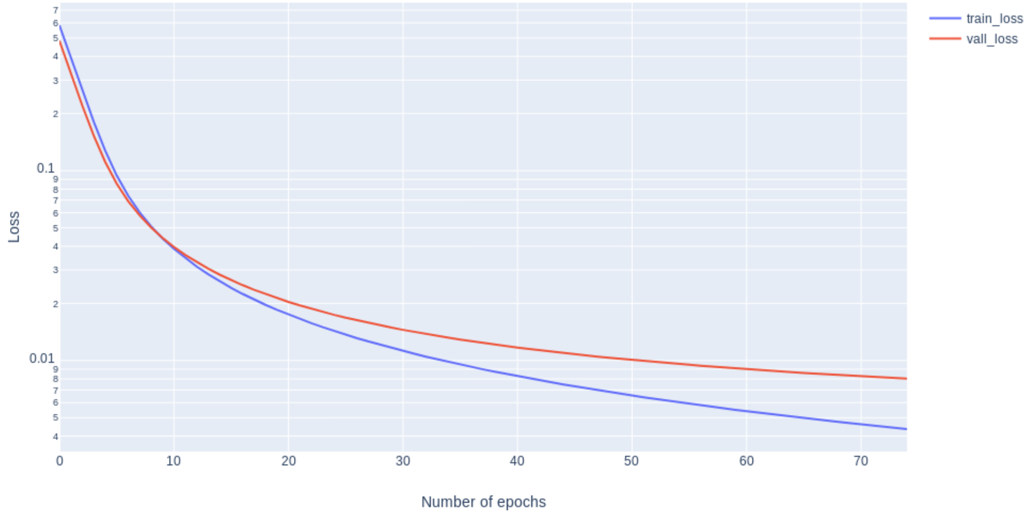


Figure 3.5: Graph showing the training and validation loss of the autoencoder over progressing epochs

3.6 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 [**TK:** Cite!] is meaningful in our vector space may not be true. [**TK:** Cite?] 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 3.6 it is apparent that there is indeed a difference. Therefore it is worth investigating if using this information improves the clustering results.

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

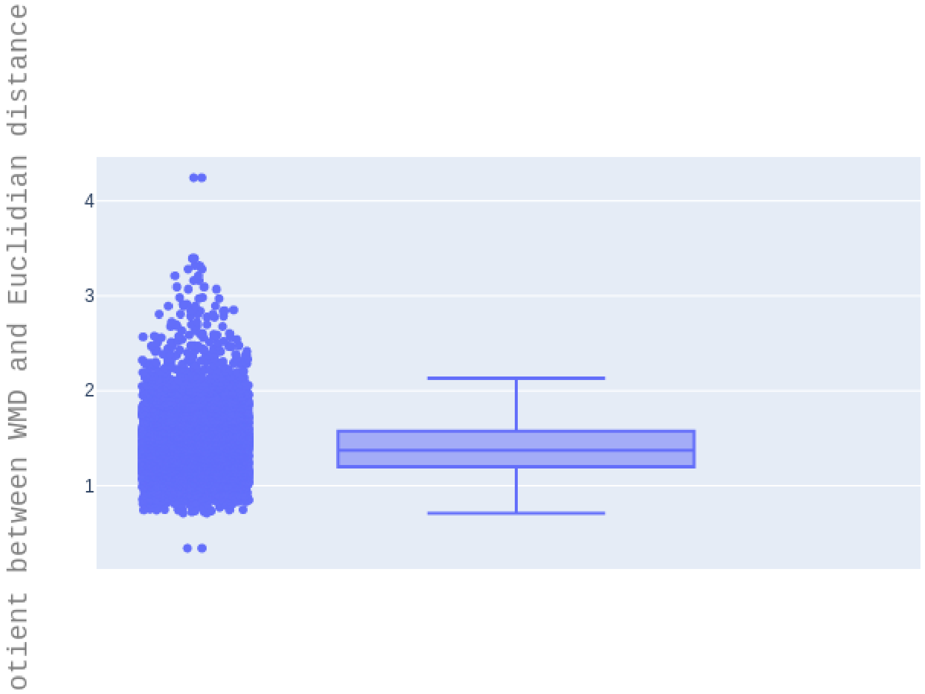


Figure 3.6: Boxplot showing the distribution of the quotients between WMD and EuD for all documents

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.

3.6.1 Explainability technique: Cluster topography

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 get distorted, a problem which we will discuss further in on of the following subsections. 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 explanatory nature of the interaction which the non-technical experts will carry out, 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.

Firstly described by Kinkeldey et al. [CTJ19], this technique computes an uncertainty measure for each point which describes how well it fits into its assigned cluster. In the original paper they 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 intendedly invokes the notion of a geographical surface since this most sciences, 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 the original 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 smaller. 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, the nearest 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.

This measure is then linearly interpolated between the 2D position of all points in order to create a relief symbolizing the uncertainty landscape of the clustering.

3.7 Assessing the quality of the topic modeling using topic coherence

A popular method to measure the quality of a topic modeling pipeline is the Coherence Score.

Plotting the coherence scores over different embedding, topic extraction and clustering models and different number of clusters reveals (Figure 3.7) , 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

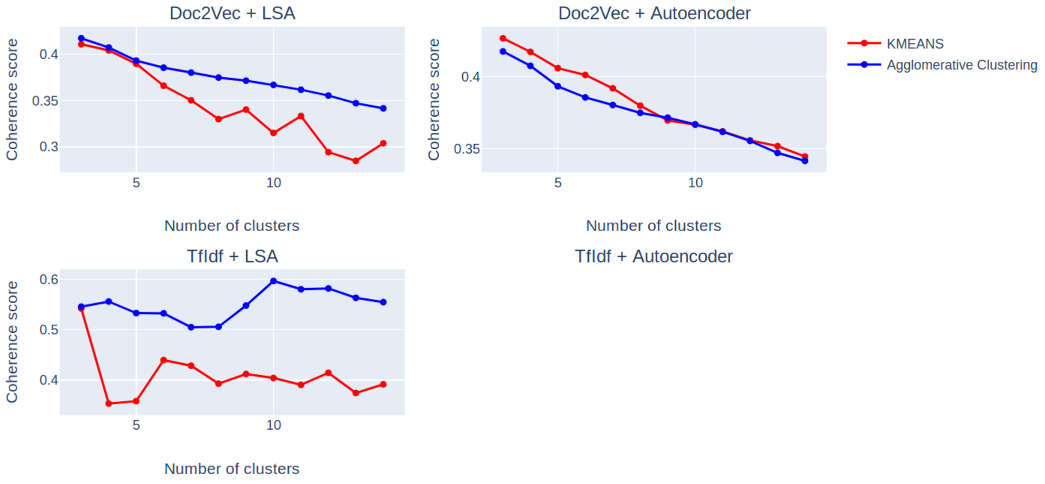


Figure 3.7: Graph showing the quality of the topic modeling while varying the embedding, topic extraction and clustering model

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 3.4: Table showing the top words for Figure 3.8

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 is considered in the Outlook.

Plotting this combination of models and parameters in Figure 3.8 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, perturbating 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.

At this point I gathered feedback from the researchers from project IKON and their answers suggested that problems may occur trying to visualize such

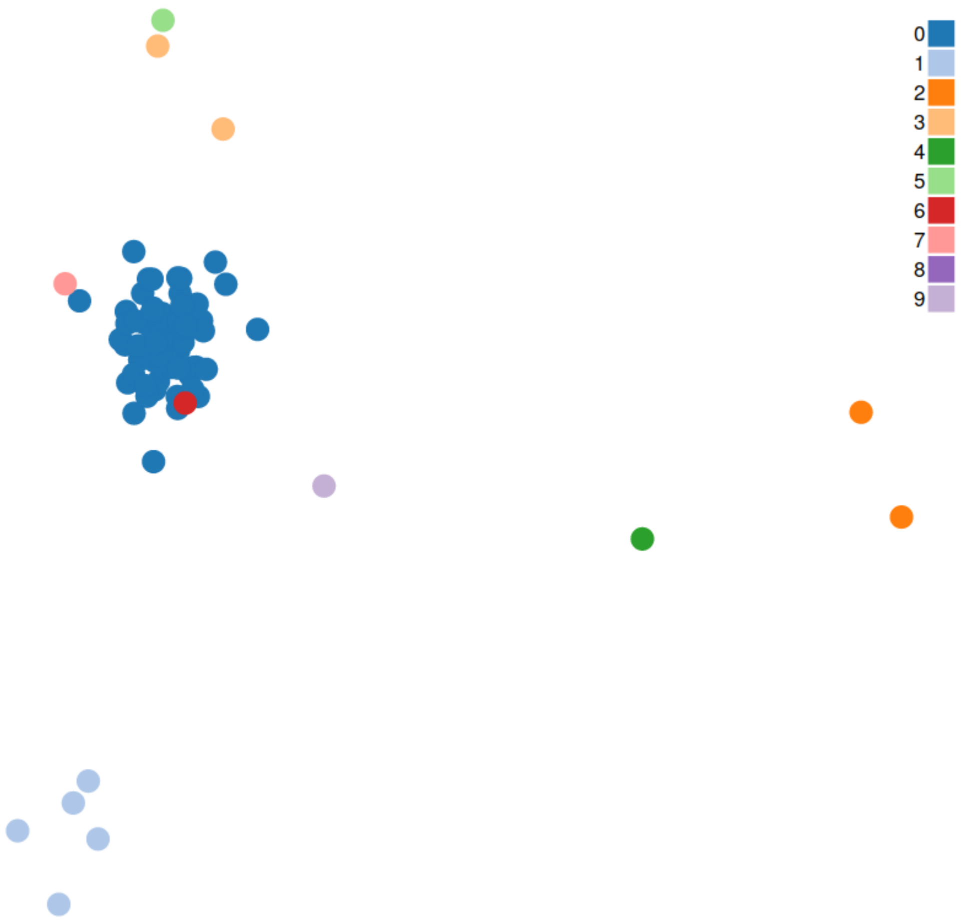


Figure 3.8: Plot for the parameter and models with the best coherence score

dominating structures in scatter plots. [**TK:** Proper argument?] Therefore I turned to the second best performing combination of models which is a Doc2Vec embedding, an autoencoder-based topic extraction and a K-Means clustering. The graph also suggests that less clusters improve the quality of the generated top words which I incorporated by constraining the range of selectable numbers of topics.

3.8 Reduction into 2D

3.9 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.

3.9.1 Explainability technique: Linearization

As Lipton [Lip16] pointed out, a visualization is already an explainability 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, as discussed in the section concerning the cluster topography. 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 error. 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. This technique also synergizes well with the previously described cluster topography method, because the interpolation also get linearized and the topography gets expanded making it easier to see differences in the relief.

4 Validation

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?

4.1 Setup

The fictive user is a postdoctoral researcher at the Museum für Naturkunde Berlin. His 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; Apple OS X; 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. One day after work, the fictive user is coming down the wide stairway of the side building he works in and sees once more the display with the visualization he passes every day on his way to and from work. This time the curiosity is stronger than the urge to go home and since he already heard that the museum financed a huge initiative to foster intra-organizational, scientific exchange, he decides to see what that application has to offer.

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

1. Identify dominating research areas
2. Find his own project
3. Explore the projects in the same cluster
4. Explore the projects in the vicinity of his own cluster

The prototypical interface provides the following actions in order to manipulate the visualization:

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

4.2 Cognitive Walkthrough

Action	1	2	3	4
2 (Figure 5.2)	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.	This action follows from the immediate presentation of top words and scatter plot.	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.	This action did not change the visualization.

3 (Figure 5.3)	The user concluded that his project must be in the ’evolution’ cluster and therefore he changes the granularity by moving the slider to the middle of the selection range.	It was the only available slider and its label suggested that this is the proper action.	The user sees a more granular clustering over all projects.	Since there is no transition, the user does not know what is happening.
2 (Figure 5.3)	After changing the granularity, the user is trying to pinpoint the cluster to which his project is now assigned.	As in the beginning the immediate presentation of the topwords and the visualization makes it inevitable to read and connect both.	The user determined that his project is probably in cluster 4.	This action did not change the visualization.
1 (Figure 5.4)	The user is trying to find his project in cluster 4.	Hovering over a glyph to display metadata is a common design strategy and therefore the user tried this first.	The selected project was not the correct one.	Since the metadata is displayed right alongside the project and the rest didn’t change, there was no confusion.
1 (Figure 5.5)	The user is trying to find his project in cluster 4.	Hovering over a glyph to display metadata is a common design strategy and therefore the user tried this first.	The selected project was again not the correct one.	Since the metadata is displayed right alongside the project and the rest didn’t change, there was no confusion.
4 (Figure 5.6)	Since the rest of the cluster is extremely cluttered, the user decides to switch into the linearized view	This is the only remaining interaction option, but the naming of this selection makes it hard to intuitivly understand what it does.	The scatter plot got uncluttered by linearization.	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.
(Figure 5.6)	Now that the clustering and assignments changed again, the user is again searching for the cluster in which his project may lie.	As in the previous instances of this action this is the only available, logical action.	The user is able to pinpoint cluster 3 as the cluster connected to bats.	Nothing changed, therefore there was no room for confusion.

1 (Figure 5.7)	The user is searching for his project and selects the first visible glyph.	In the previous steps the user verified that hovering for metadata is a possibility.	The project he selected was indeed his own project. All of the top words make sense in the context of his project.	Again there was no confusion.
1 (Figure 5.8)	Now that he found his project the user is interested what kind of projects are also in the cluster which was assigned to his project. Therefore he selects the next available project in the same cluster.	In the previous steps the user verified that hovering for metadata is a possibility.	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.	Again there was no confusion.
1 (Figure 5.9)	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.	In the previous steps the user verified that hovering for metadata is a possibility.	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.	Again there was no confusion.
1 (Figure 5.10)	Since the cluster does make sense the user decides to have a look at the neighbouring projects.	In the previous steps the user verified that hovering for metadata is a possibility.	The first project he selects does investigate a completly different field.	The user is able to tell why the project lies in another cluster using the top words.
1 (Figure 5.11)	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.	In the previous steps the user verified that hovering for metadata is a possibility.	The next project he selects is surprisingly also connected to bats.	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.

Table 4.1: Exploratory interaction simulated by a CW

This cognitive walkthrough unveiled a number of usability issues with the visualization, but also showed that the implemented explainability techniques do indeed help the fictive used to accomplish his goals.

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 th visualization anew. Adding animated transitions could help alleviating this problems by introducing object permanence in the views.

Table 4.2: 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 explainability techniques for NLP exist and how they could support a non-technical expert in understanding the output from the system.

The first step was a systematic literature mapping study according to Petersen et al. [PFMM] which unveiled that I was able 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 on the other hand.

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. Three out of four components where also augmented by explainability methods. Following an analysis investigating the interplay between all implemented methods, the decision was made to remain with the existing pipeline, but augment it by the proposed explainability techniques.

A cognitive walkthrough simulating a researcher doing 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.

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 explainability 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.

Appendix

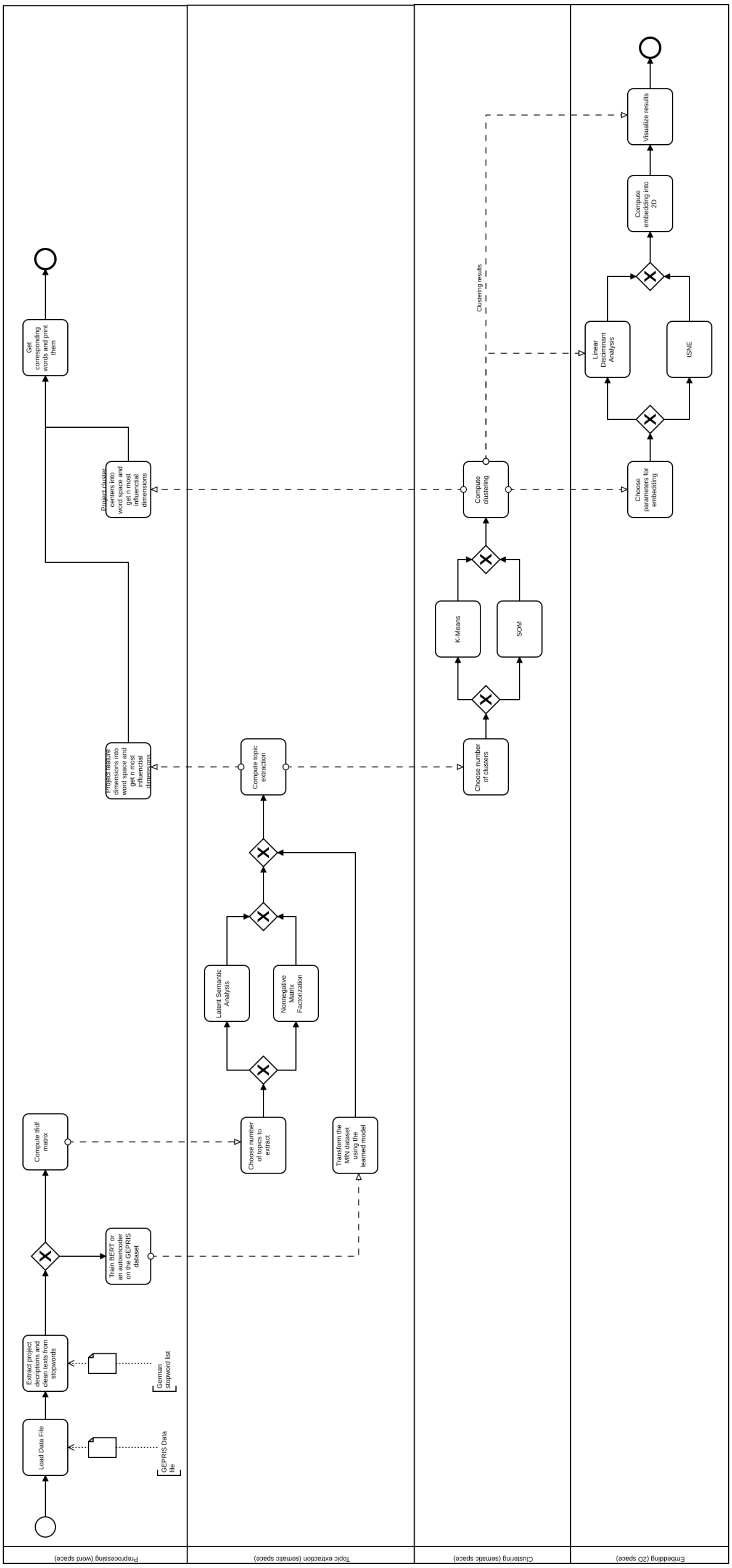
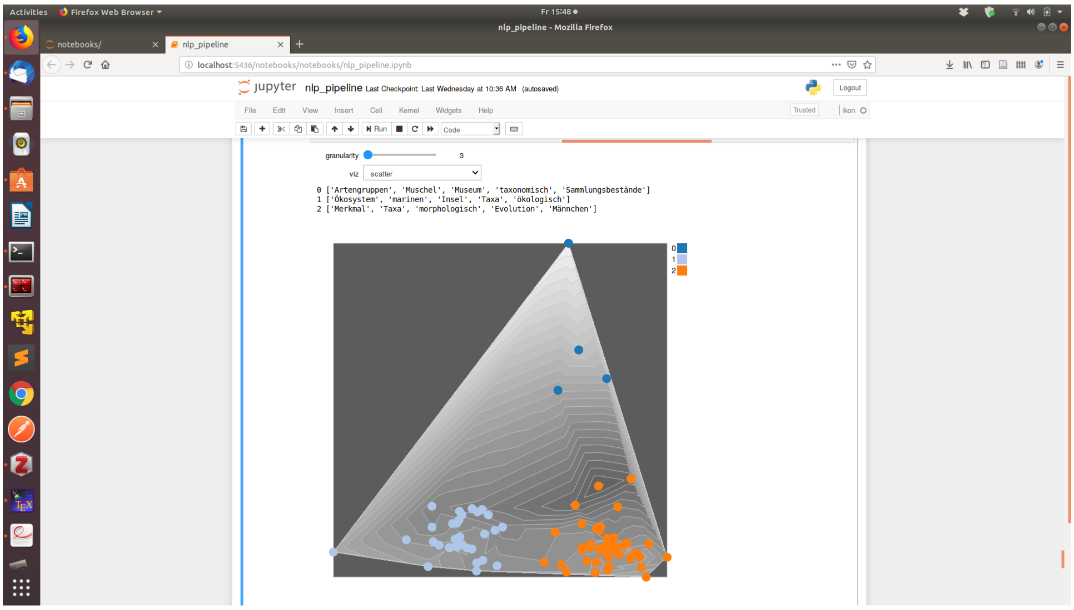

Preprocessing (word space)
Topic extraction (sematic space)
Clustering (sematic space)
Embedding (2D space)
Load Data File
GEPRIS Data file
Extract project decriptions and clean texts from stopwords
German stopword list
Train BERT or an autoencoder on the GEPRIS dataset
Compute tfidf matrix
Choose number of topics to extract
Transform the MfN dataset using the learned model
Latent Semantic Analysis
Nonnegative Matrix Factorization
Compute topic extraction
Project feature dimensions into word space and get n most influencial dimensions
Choose number of clusters
K-Means
SOM
Compute clustering
Project cluster centers into word space and get n most influencial dimensions
Get corresponding words and print them
Clustering results
Choose parameters for embedding
Linear Discriminant Analysis
tSNE
Compute embedding into 2D
Visualize results

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



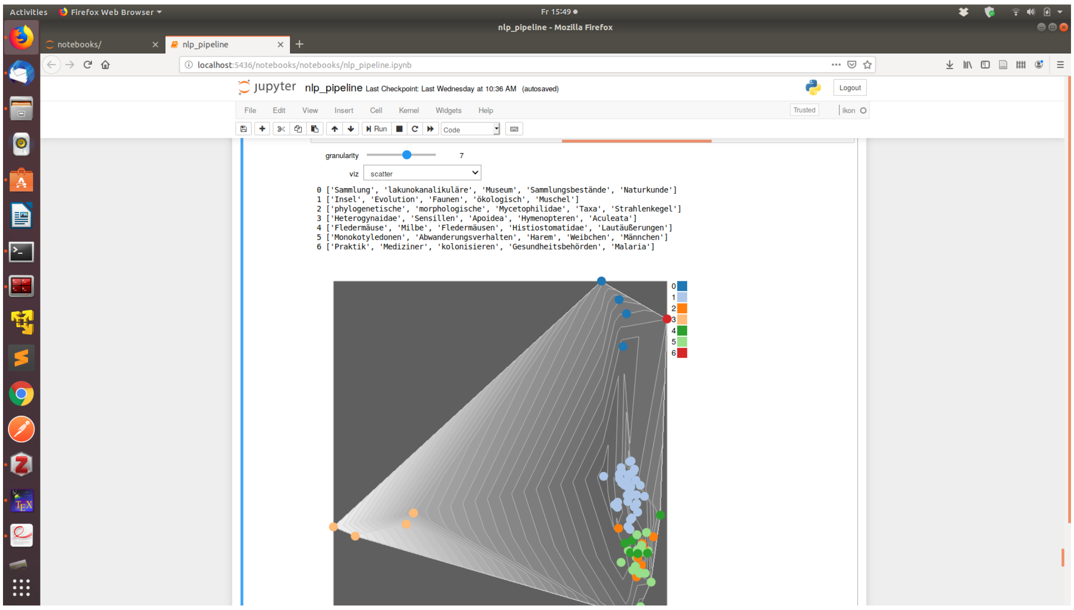
Activities
Firefox Web Browser
Fr 15:48
nlp_pipeline - Mozilla Firefox
notebooks/
nlp_pipeline
localhost:5436/notebooks/notebooks/nlp_pipeline.ipynb
jupyter
nlp_pipeline Last Checkpoint: Last Wednesday at 10:36 AM (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Ikon
Run
Code
0
1
2

granularity 3

viz scatter

```
0 ['Artengruppen', 'Muschel', 'Museum', 'taxonomisch', 'Sammlungsbestände']
1 ['Ökosystem', 'marinen', 'Insel', 'Taxa', 'ökologisch']
2 ['Merkmal', 'Taxa', 'morphologisch', 'Evolution', 'Männchen']
```

Figure 5.2: Cognitive Walkthrough step 1



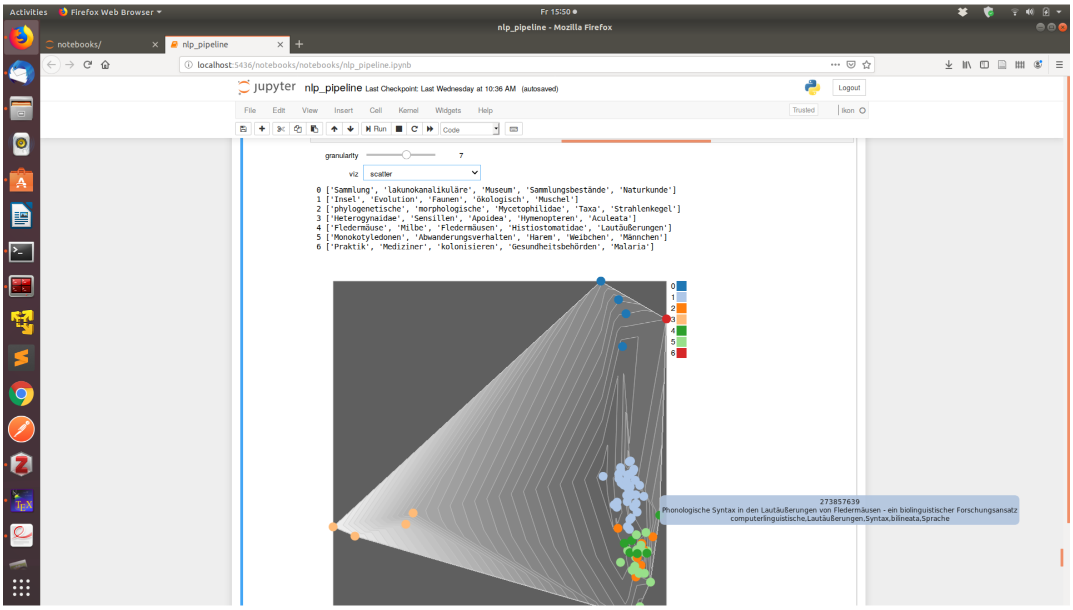
Activities
Firefox Web Browser
Fr 15:49
nlp_pipeline - Mozilla Firefox
notebooks/
nlp_pipeline
localhost:5436/notebooks/notebooks/nlp_pipeline.ipynb
jupyter
nlp_pipeline Last Checkpoint: Last Wednesday at 10:36 AM (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Ikon
Run
Code
0
1
2
3
4
5
6

granularity 7

viz scatter

```
0 ['Sammlung', 'lakunokanalikuläre', 'Museum', 'Sammlungsbestände', 'Naturkunde']
1 ['Insel', 'Evolution', 'Faunen', 'ökologisch', 'Muschel']
2 ['phylogenetische', 'morphologische', 'Mycetophilidae', 'Taxa', 'Strahlenkegel']
3 ['Heterogynaidae', 'Sensillen', 'Apoidea', 'Hymenopteren', 'Aculeata']
4 ['Fledermäuse', 'Milbe', 'Fledermäusen', 'Histiostomatidae', 'Lautäußerungen']
5 ['Monokotyledonen', 'Abwanderungsverhalten', 'Harem', 'Weibchen', 'Männchen']
6 ['Praktik', 'Mediziner', 'kolonisieren', 'Gesundheitsbehörden', 'Malaria']
```

Figure 5.3: Cognitive Walkthrough step 2


Activities
Firefox Web Browser
Fr 15:50
nlp_pipeline - Mozilla Firefox
notebooks/
nlp_pipeline
localhost:5436/notebooks/notebooks/nlp_pipeline.ipynb
jupyter
nlp_pipeline Last Checkpoint: Last Wednesday at 10:36 AM (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
ikon
Run
Code
0
1
2
3
4
5
6

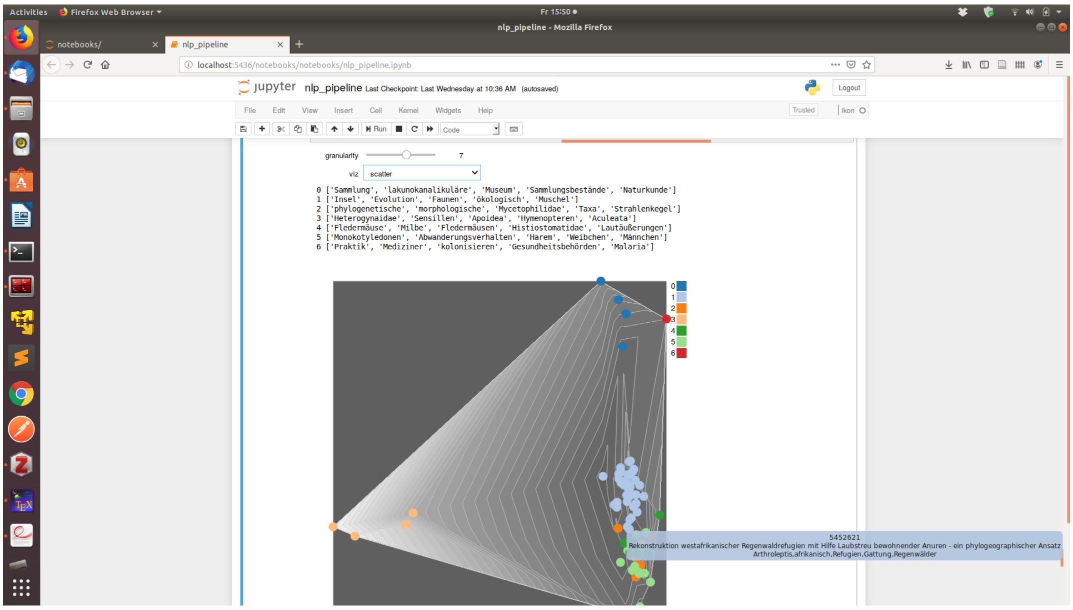
granularity 7

viz scatter

```
0 ['Sammlung', 'lakunokanalikuläre', 'Museum', 'Sammlungsbestände', 'Naturkunde']
1 ['Insel', 'Evolution', 'Faunen', 'ökologisch', 'Muschel']
2 ['phylogenetische', 'morphologische', 'Mycetophilidae', 'Taxa', 'Strahlenkegel']
3 ['Heterogynaidae', 'Sensillen', 'Apoidea', 'Hymenopteren', 'Aculeata']
4 ['Fledermäuse', 'Milbe', 'Fledermäusen', 'Histiostomatidae', 'Lautäußerungen']
5 ['Monokotyledonen', 'Abwanderungsverhalten', 'Harem', 'Weibchen', 'Männchen']
6 ['Praktik', 'Mediziner', 'kolonisieren', 'Gesundheitsbehörden', 'Malaria']
```

273857639
Phonologische Syntax in den Lautäußerungen von Fledermäusen - ein biolinguistischer Forschungsansatz
computerlinguistische,Lautäußerungen,Syntax,bilineata,Sprache

Figure 5.4: Cognitive Walkthrough step 3


Activities
Firefox Web Browser
Fr 15:50
nlp_pipeline - Mozilla Firefox
notebooks/
nlp_pipeline
localhost:5436/notebooks/notebooks/nlp_pipeline.ipynb
jupyter
nlp_pipeline Last Checkpoint: Last Wednesday at 10:36 AM (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
ikon
Run
Code
0
1
2
3
4
5
6

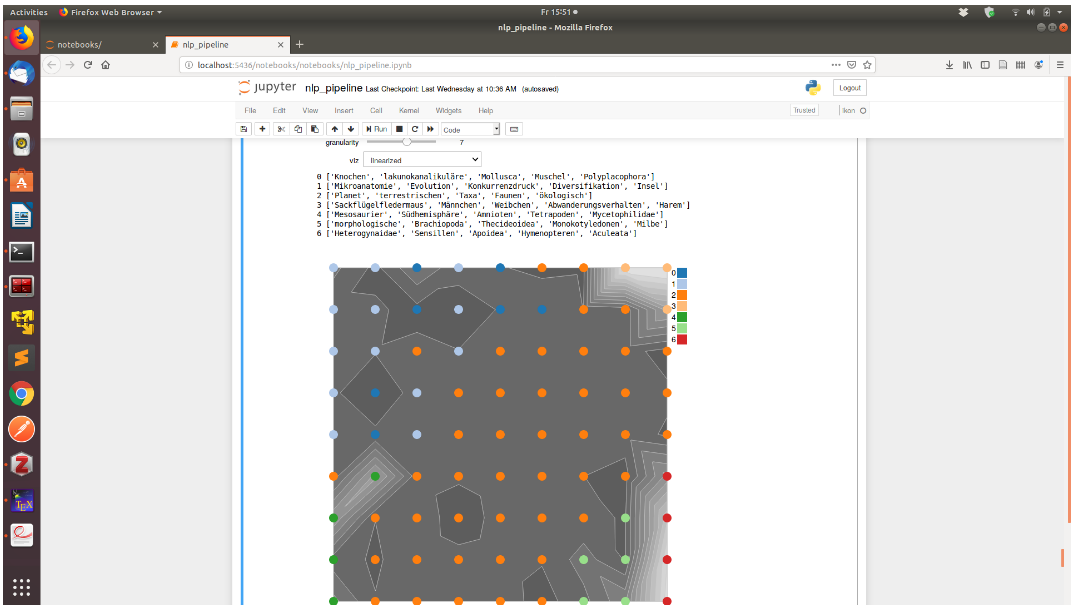
granularity 7

viz scatter

```
0 ['Sammlung', 'lakunokanalikuläre', 'Museum', 'Sammlungsbestände', 'Naturkunde']
1 ['Insel', 'Evolution', 'Faunen', 'ökologisch', 'Muschel']
2 ['phylogenetische', 'morphologische', 'Mycetophilidae', 'Taxa', 'Strahlenkegel']
3 ['Heterogynaidae', 'Sensillen', 'Apoidea', 'Hymenopteren', 'Aculeata']
4 ['Fledermäuse', 'Milbe', 'Fledermäusen', 'Histiostomatidae', 'Lautäußerungen']
5 ['Monokotyledonen', 'Abwanderungsverhalten', 'Harem', 'Weibchen', 'Männchen']
6 ['Praktik', 'Mediziner', 'kolonisieren', 'Gesundheitsbehörden', 'Malaria']
```

5452621
Rekonstruktion westafrikanischer Regenwaldrefugien mit Hilfe Laubstreu bewohnender Anuren - ein phylogeographischer Ansatz
Arthroleptis,afrikanisch,Refugien,Gattung,Regenwälder

Figure 5.5: Cognitive Walkthrough step 4

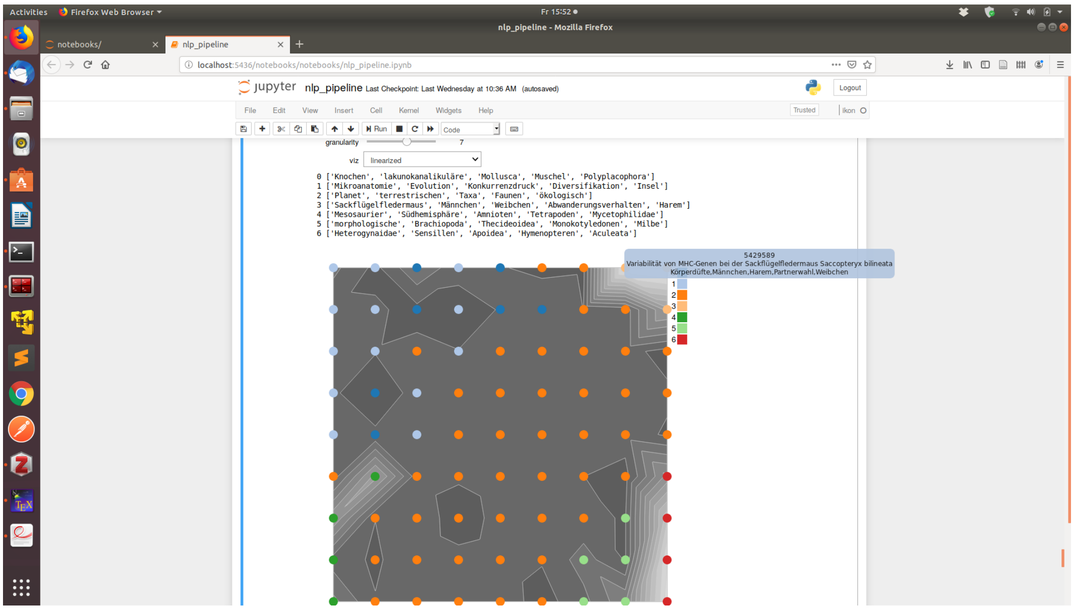

Activities
Firefox Web Browser
Fr 15:51
nlp_pipeline - Mozilla Firefox
notebooks/
nlp_pipeline
localhost:5436/notebooks/notebooks/nlp_pipeline.ipynb
jupyter
nlp_pipeline Last Checkpoint: Last Wednesday at 10:36 AM (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Ikon
Run
Code
0
1
2
3
4
5
6

granularity 7

viz linearized

```
0 ['Knochen', 'lakunokanalikuläre', 'Mollusca', 'Muschel', 'Polyplacophora']
1 ['Mikroanatomie', 'Evolution', 'Konkurrenzdruck', 'Diversifikation', 'Insel']
2 ['Planet', 'terrestrischen', 'Taxa', 'Faunen', 'ökologisch']
3 ['Sackflügelfledermaus', 'Männchen', 'Weibchen', 'Abwanderungsverhalten', 'Harem']
4 ['Mesosaurier', 'Südhemisphäre', 'Amnioten', 'Tetrapoden', 'Mycetophilidae']
5 ['morphologische', 'Brachiopoda', 'Thecideoidea', 'Monokotyledonen', 'Milbe']
6 ['Heterogynaidae', 'Sensillen', 'Apoidea', 'Hymenopteren', 'Aculeata']
```

Figure 5.6: Cognitive Walkthrough step 5


Activities
Firefox Web Browser
Fr 15:52
nlp_pipeline - Mozilla Firefox
notebooks/
nlp_pipeline
localhost:5436/notebooks/notebooks/nlp_pipeline.ipynb
jupyter
nlp_pipeline Last Checkpoint: Last Wednesday at 10:36 AM (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Ikon
Run
Code
1
2
3
4
5
6

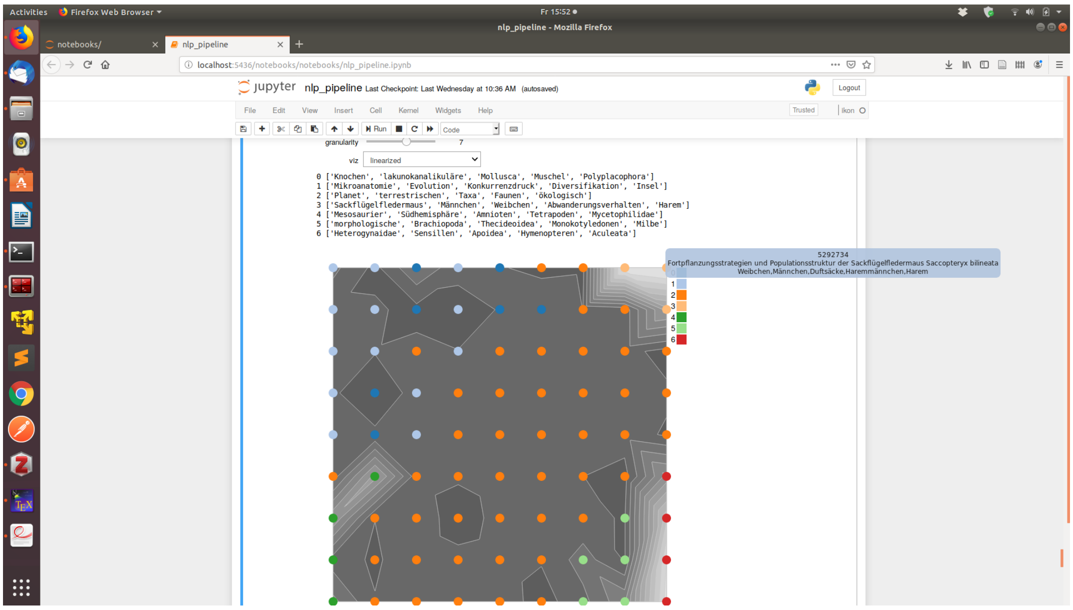
granularity 7

viz linearized

```
0 ['Knochen', 'lakunokanalikuläre', 'Mollusca', 'Muschel', 'Polyplacophora']
1 ['Mikroanatomie', 'Evolution', 'Konkurrenzdruck', 'Diversifikation', 'Insel']
2 ['Planet', 'terrestrischen', 'Taxa', 'Faunen', 'ökologisch']
3 ['Sackflügelfledermaus', 'Männchen', 'Weibchen', 'Abwanderungsverhalten', 'Harem']
4 ['Mesosaurier', 'Südhemisphäre', 'Amnioten', 'Tetrapoden', 'Mycetophilidae']
5 ['morphologische', 'Brachiopoda', 'Thecideoidea', 'Monokotyledonen', 'Milbe']
6 ['Heterogynaidae', 'Sensillen', 'Apoidea', 'Hymenopteren', 'Aculeata']
```

5429589
Variabilität von MHC-Genen bei der Sackflügelfledermaus Saccopteryx bilineata
Körperdüfte,Männchen,Harem,Partnerwahl,Weibchen

Figure 5.7: Cognitive Walkthrough step 6


Activities
Firefox Web Browser
Fr 15:52
nlp_pipeline - Mozilla Firefox
notebooks/
nlp_pipeline
localhost:5436/notebooks/notebooks/nlp_pipeline.ipynb
jupyter nlp_pipeline Last Checkpoint: Last Wednesday at 10:36 AM (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
ikon
Run
Code
0
1
2
3
4
5
6

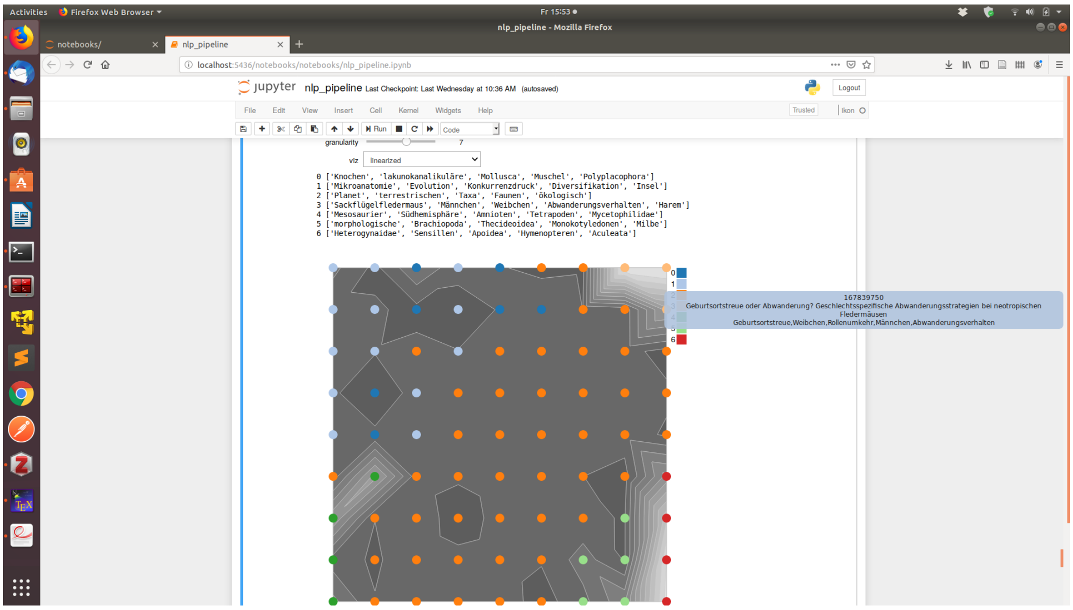
granularity 7

viz linearized

```
0 ['Knochen', 'lakunokanalikuläre', 'Mollusca', 'Muschel', 'Polyplacophora']
1 ['Mikroanatomie', 'Evolution', 'Konkurrenzdruck', 'Diversifikation', 'Insel']
2 ['Planet', 'terrestrischen', 'Taxa', 'Faunen', 'ökologisch']
3 ['Sackflügelfledermaus', 'Männchen', 'Weibchen', 'Abwanderungsverhalten', 'Harem']
4 ['Mesosaurier', 'Südhemisphäre', 'Amnioten', 'Tetrapoden', 'Mycetophilidae']
5 ['morphologische', 'Brachiopoda', 'Thecideoidea', 'Monokotyledonen', 'Milbe']
6 ['Heterogynaidae', 'Sensillen', 'Apoidea', 'Hymenopteren', 'Aculeata']
```

5292734
Fortpflanzungsstrategien und Populationsstruktur der Sackflügelfledermaus Saccopteryx bilineata
Weibchen,Männchen,Duftsäcke,Haremmännchen,Harem

Figure 5.8: Cognitive Walkthrough step 7


Activities
Firefox Web Browser
Fr 15:53
nlp_pipeline - Mozilla Firefox
notebooks/
nlp_pipeline
localhost:5436/notebooks/notebooks/nlp_pipeline.ipynb
jupyter nlp_pipeline Last Checkpoint: Last Wednesday at 10:36 AM (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
ikon
Run
Code
0
1
2
3
4
5
6

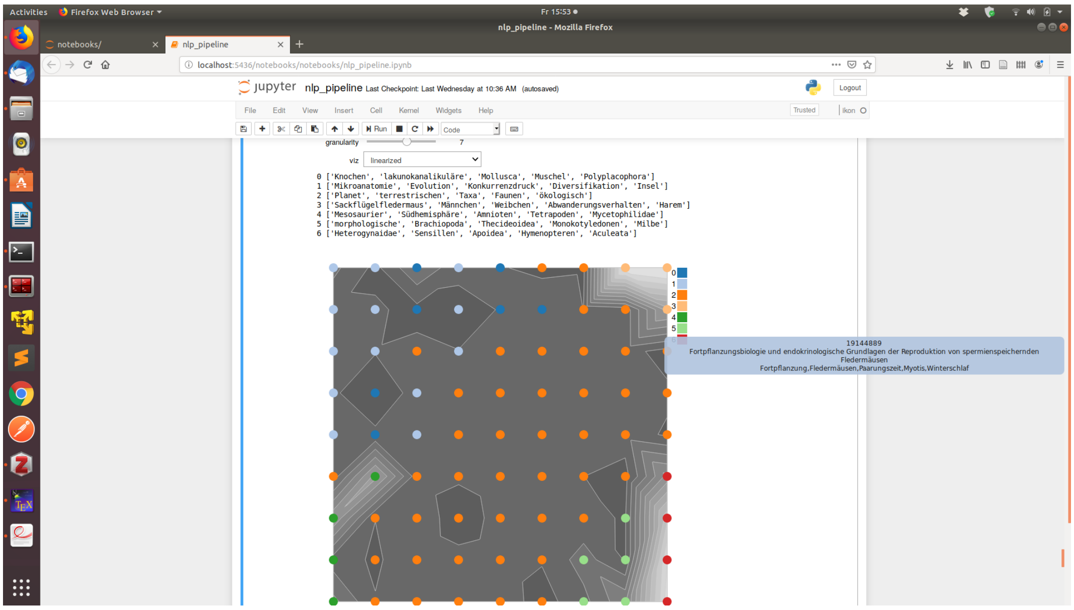
granularity 7

viz linearized

```
0 ['Knochen', 'lakunokanalikuläre', 'Mollusca', 'Muschel', 'Polyplacophora']
1 ['Mikroanatomie', 'Evolution', 'Konkurrenzdruck', 'Diversifikation', 'Insel']
2 ['Planet', 'terrestrischen', 'Taxa', 'Faunen', 'ökologisch']
3 ['Sackflügelfledermaus', 'Männchen', 'Weibchen', 'Abwanderungsverhalten', 'Harem']
4 ['Mesosaurier', 'Südhemisphäre', 'Amnioten', 'Tetrapoden', 'Mycetophilidae']
5 ['morphologische', 'Brachiopoda', 'Thecideoidea', 'Monokotyledonen', 'Milbe']
6 ['Heterogynaidae', 'Sensillen', 'Apoidea', 'Hymenopteren', 'Aculeata']
```

167839750
Geburtsortstreue oder Abwanderung? Geschlechtsspezifische Abwanderungsstrategien bei neotropischen Fledermäusen
Geburtsortstreue,Weibchen,Rollenumkehr,Männchen,Abwanderungsverhalten

Figure 5.9: Cognitive Walkthrough step 8


Activities
Firefox Web Browser
Fr 15:53
nlp_pipeline - Mozilla Firefox
notebooks/
nlp_pipeline
localhost:5436/notebooks/notebooks/nlp_pipeline.ipynb
jupyter
nlp_pipeline Last Checkpoint: Last Wednesday at 10:36 AM (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Ikon
Run
Code
0
1
2
3
4
5
6

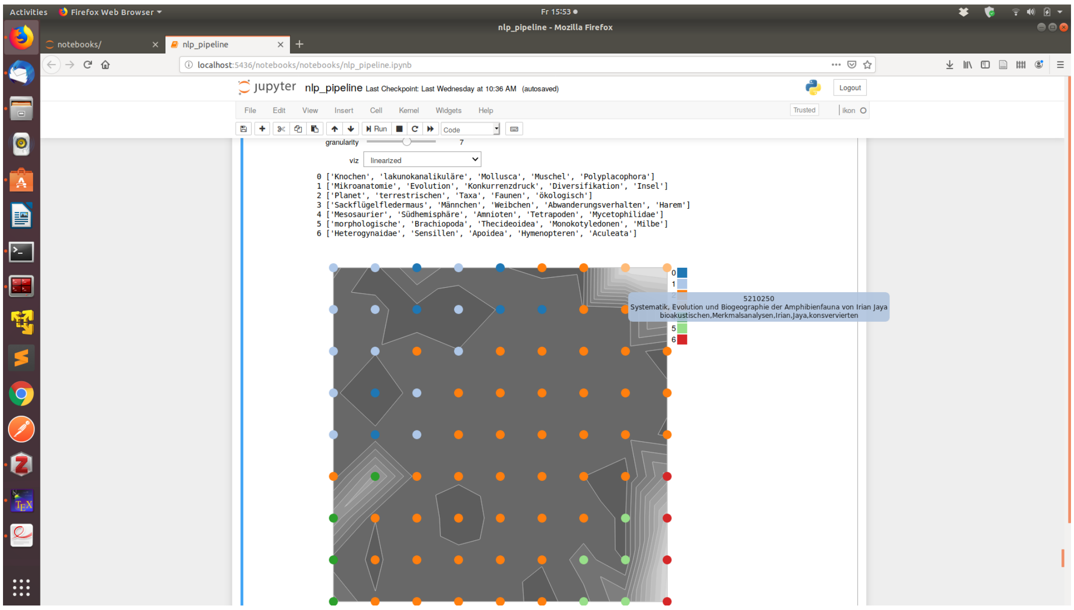
granularity 7

viz linearized

```
0 ['Knochen', 'lakunokanalikuläre', 'Mollusca', 'Muschel', 'Polyplacophora']
1 ['Mikroanatomie', 'Evolution', 'Konkurrenzdruck', 'Diversifikation', 'Insel']
2 ['Planet', 'terrestrischen', 'Taxa', 'Faunen', 'ökologisch']
3 ['Sackflügelfledermaus', 'Männchen', 'Weibchen', 'Abwanderungsverhalten', 'Harem']
4 ['Mesosaurier', 'Südhemisphäre', 'Amnioten', 'Tetrapoden', 'Mycetophilidae']
5 ['morphologische', 'Brachiopoda', 'Thecideoidea', 'Monokotyledonen', 'Milbe']
6 ['Heterogynaidae', 'Sensillen', 'Apoidea', 'Hymenopteren', 'Aculeata']
```

19144889
Fortpflanzungsbiologie und endokrinologische Grundlagen der Reproduktion von spermienspeichernden Fledermäusen
Fortpflanzung,Fledermäusen,Paarungszeit,Myotis,Winterschlaf

Figure 5.10: Cognitive Walkthrough step 9


Activities
Firefox Web Browser
Fr 15:53
nlp_pipeline - Mozilla Firefox
notebooks/
nlp_pipeline
localhost:5436/notebooks/notebooks/nlp_pipeline.ipynb
jupyter
nlp_pipeline Last Checkpoint: Last Wednesday at 10:36 AM (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Ikon
Run
Code
0
1
5
6

granularity 7

viz linearized

```
0 ['Knochen', 'lakunokanalikuläre', 'Mollusca', 'Muschel', 'Polyplacophora']
1 ['Mikroanatomie', 'Evolution', 'Konkurrenzdruck', 'Diversifikation', 'Insel']
2 ['Planet', 'terrestrischen', 'Taxa', 'Faunen', 'ökologisch']
3 ['Sackflügelfledermaus', 'Männchen', 'Weibchen', 'Abwanderungsverhalten', 'Harem']
4 ['Mesosaurier', 'Südhemisphäre', 'Amnioten', 'Tetrapoden', 'Mycetophilidae']
5 ['morphologische', 'Brachiopoda', 'Thecideoidea', 'Monokotyledonen', 'Milbe']
6 ['Heterogynaidae', 'Sensillen', 'Apoidea', 'Hymenopteren', 'Aculeata']
```

5210250
Systematik, Evolution und Biogeographie der Amphibienfauna von Irian Jaya
bioakustischen,Merkmalsanalysen,Irian,Jaya,konsvervierten

Figure 5.11: Cognitive Walkthrough step 10

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.

[BNJ03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[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.

[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.

[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.

[KLHK19] D. Kim, W. Lim, M. Hong, and H. Kim. The Structure of Deep Neural Network for Interpretable Transfer Learning. In *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 1–4, February 2019.

[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 &#38; 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/.

[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.

[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.

[ZYL+18] Quanshi Zhang, Yu Yang, Yuchen Liu, Ying Nian Wu, and Song-Chun Zhu. Unsupervised Learning of Neural Networks to Explain Neural Networks. *arXiv:1805.07468 [cs]*, May 2018.