UNIKASSEL VERSITÄT
Department of Electrical Engineering and Computer Science
Chair Intelligent Embedded Systems

Identification of key information with topic analysis on large unstructured text data

BACHELORTHESIS

Department of Electrical Engineering and Computer Science
University of Kassel

Author Name:	Klara Maximiliane Gutekunst
Address:	*** REMOVED ***
	34125 Kassel
Matriculation number:	*** REMOVED ***
E-Mail:	klara.gutekunst@student.uni-kassel.de
Department:	Chair Intelligent Embedded Systems
Examining board 1:	Prof. Dr. rer. nat. Bernhard Sick
Examining board 2:	Prof. Dr. Gerd Stumme
Supervisor:	Dr. Christian Gruhl
Date:	October 4, 2023

Abstract

Finding relevant documents and connections between multiple ones becomes significantly more difficult due to the sheer amount of documents available. Institutes, such as the (German) tax offices have access to leak data, e.g., the Bahama leak, containing huge amounts of documents and valuable information yet to be extracted. However, these institutes, companies and individuals do not have sufficient resources to explore individual documents in order to find a specific one or to identify the key topics of them. Hence, computational means, such as text mining, may facilitate the situation. This thesis proposes an approach to find relevant documents and identify topics from a large text corpus.

Contents

Abkürzungsverzeichnis

CSS	Cascading Style Sheet
RQ	Research Question
LDA	Latent Dirichlet Allocation
TF-IDF	Term Frequency - Inverse Document Frequency
TF	Term Frequency
IDF	Inverse Document Frequency
BERT	Bidirectional Encoder Representations from Transformers
BERTopic	BERT Topic Model
Doc2Vec	Document to Vector
Word2Vec	Word to Vector
CBOW	Continuous-Bag-of-Words
GloVe	Global Vectors
USE	Universal Sentence Encoder
PCA	Principal Component Analysis
kNN	k-nearest neighbor
API	Application Programming Interface
HTML	Hypertext Markup Language
CSS	Cascading Style Sheet
JSON	JavaScript Object Notation
PKL	Pickle
HNSW	Hiercharical Navigable Small World
OPTICS	Ordering Points To Identify the Clustering Structure
AE	Autoencoder
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
KL	Karhonen-Loéve
SVD	singular value decomposition
HTTP	Hypertext Transfer Protocol
URL	Uniform Resource Locator
SQL	Structured Query Language
NoSQL	Not only SQL
ACID	Atomicity, Consistency, Isolation, Durability
VSM	Vector Space Model
NN	Neural Network
DNN	Deep Neural Network
RNN	Recurrent Neural Network

RMSE	Root Mean Square Error
ML	Machine Learning
NLP	Natural Language Processing
PVDM	Paragraph Vector Distributed Memory
PV-DBOW	Distributed Bag of Words
SNLI	Stanford Natural Language Inference
NLI	Natural Language Inference
BiLSTM	bi-directional Long Short-Term Memory
LSTM	Long Short-Term Memory
DAN	Deep Averaging Network
BERT	Bidirectional Encoder Representations from Transformers
SBERT	Sentence-BERT
GloVe	Global Vectors for Word Representation
PDF	Portable Document Format

1 Introduction

According to [24], the Bahamas leak is roughly 38 GB collection of documents, which were leaked in 2016. The data is used by (German) tax offices to identify tax evasion. However, it has proven to be challenging to identify the relevant documents and connections between documents due to the amount of documents in the leak.

Therefore, the goal of this thesis is to suggest approaches to support the investigators of the tax offices. Text exploration methods include topic modelling.

The topics to be identified can be groups of words which appear more often than the average or groups of similar documents. Hence, a topic is not always the defined topic in terms of content, but sometimes a statistical phenomenon. Since different methods define different topics, as they work and define the meaning of 'topic' differently, their results are compared and evaluated on the dataset.

Besides literature research, application and evaluation of the methods identified, certain preprocessing methods have proven to be eminent to successful work with unstructured text data. These methods include chunking/ tokenization (separating texts into equally sized segments), lemmatization (e.g., faster to fast), conversion to small letters and stop-word-lists.

1.1 Motivation/ Objective

Assumption: similarities between documents (in terms of appearance and content-wise) On a broader scope, this thesis aims to provide computational means to facilitate the work with large unstructured text data for individuals. In the following, certain goals are defined, which are to be achieved in this thesis.

Motivation/ problem: actively use machine learning techniques to analyse large text corpus and thus, reduce the amount of manual (human) work. This includes analysis in terms of textual (content) and visual (appearance/ layout) information like a human would do. The goal is to identify similarities between documents and group (cluster) them together - the topic of the cluster does not have to be labelled specifically. This serves as a first

step/ preprocessing, e.g., a human finds a document of interest (for instance from random sampling) and wants to find similar documents to it.

Usability. The methods should be bundled in an application, which is easy to use and does not require any programming skills.

Semantic similarity. The documents grouped together should be semantically similar.

Topic identification. The topics identified should be meaningful to the task at hand.

Offline Calculation. The database should be calculated offline so that the queries can be executed with little latency.

1.2 Research Questions

The following research questions build the guidelines for this thesis.

1.2.1 Research Question (RQ)1: Effect of different preprocessing pipelines on performance?

In terms of RQ1, one could compare different types of stemmers (i.e. algorithmic vs. dictionary-based).

1.2.2 RQ2: Effect of different similarity measurement types on performance?

In terms of RQ2, one could compare different types of similarity measurement types (i.e. cosine similarity vs. soft cosine similarity).

1.2.3 RQ3: Which type of database is best suited for this task?

In terms of RQ3, one could compare different types of databases (i.e. object-orientated, relational, document).

1.2.4 RQ4: Effect of different embeddings on performance?

In terms of RQ4, one could compare different types of embeddings (i.e. Doc2Vec, Bag-of-words, LDA, BERTopic).

1.3 Structure of the Thesis

The rest of this thesis is structured as follows. Chapter 2 provides background information on the topic of this thesis. Chapter 3 describes the implementation of the methods. Chapter 4 evaluates the methods. Chapter 5 discusses the results. Chapter 6 concludes this thesis and Chapter 7 gives an outlook on future work.

1.4 Related work

2 Fundamentals/ State of the art

[23] [4]

Basic concepts, methods used, etc.

2.1 Preprocessing

Similar to other Machine Learning (ML) domains, Natural Language Processing (NLP) requires preprocessing of the data. Usually, textual data contains irrelevant information and noise. Hence, preprocessing improves the performance and results [28]. The next sections describe a selection of the preprocessing steps applied in this work.

2.1.1 Tokenization/ Chunking

Tokenization is the process of splitting a text into smaller pieces, so-called *tokens*. Tokens can be words and punctuation marks [4]. However, the definition of a token depends on the application. For instance, certain tokenization implementations may identify tokens as subsequent series of non-whitespace characters omitting all numbers and punctuation marks [32].

Chunking is the process of splitting a text into smaller pieces, so-called *chunks*. A chunk is a sequence of tokens, e.g. words, in a text [4]. Chunks do not overlap. According to Bird et al., chunkers produce their pieces by following a set of rules, e.g., grammar rules.

2.1.2 Stemming

In order to avoid language inflections, i.e. treating words with similar meanings differently, stemming is applied [28]. According to Bird et al., *stemming* is the process of striping off any affixes, i.e. prefixes and suffixes [32], from a word and returning the stem. Different types of stemmers are better suited for certain applications than others. Hence, the choice of the stemmer depends on the application.

For instance, the *Porter Stemmer* performs well for English texts [4, 28]. It is predominantly used for the normalization of inflected forms. The *Porter Stemmer* is an algorithmic stemmer, i.e. it applies a set of rules to a word to produce the stem and thus, does not use a dictionary [32, 28].

2.1.3 Lemmatization

Stemming or lemmatization is used to reduce the vocabulary size [28]. By ensuring the resulting stem is a valid word, the process of stemming is called *lemmatization* [4]. Some implementations of lemmatizers only stem words if the result is in its dictionary. Since lemmatizers validate the result prior to returning it, they are usually slower than stemmers [4].

The *WordNetLemmatizer* from the `nltk` package requires a vocabulary [28]. According to Radu et al., it is frequently used for English texts. It manages the meaning of words and considers the order of the words [28].

2.1.4 Stop-Word-Removal

Omitting words that are not relevant to the context of the text is called *stop-word-removal*. Stop words not only depend on the domain but also on the language [32]. Possibly, domain-specific stop-word lists are used to remove words that are not relevant to the context of the text [32].

2.1.5 Lower case

Words with capital letters are converted to lowercase.

2.2 Similarity Measurement

Since embeddings represent texts as vectors, they not only facilitate human interpretability of relationships between texts using the text's respective point in a N-dimensional space, but also enable the use of similarity measures to quantify the similarity between texts. A similarity measure defines a metric to quantify the similarity between two texts [32, 15].

There are several similarity measures, such as the dot product quantifying the number of shared tokens of two texts, the (soft) cosine similarity, which is the normalized dot product

and calculates the angle between two vectors, and many more [32, 15, 29]. The following section describes a few of the metrics usable for similarity measurement.

2.2.1 Euclidian distance

The *euclidian distance* is a distance measure. In order to measure the distance between two points in a N-dimensional space, the root of the sum of squared distances between the respective values of every dimension is calculated. The distance function Euclidean (L2) norm is given in Equation 2.1 from [15]. The points x_1, x_2 correspond to objects d_1, d_2.

$$d_E(x_1, x_2) = \sqrt{\sum_{i=1}^{N} (x_1[i] - x_2[i])^2} \tag{2.1}$$

2.2.2 Cosine Similarity

In the traditional bag-of-words approach the texts are represented as vectors of TF-IDF coefficients [6]. Without further processing, the vector is of size N, N being the number of different words of the texts [6]. Hence, a vector represents its corresponding text in a N-dimensional space. This space is called Vector Space Model (VSM) [31].

The similarity between two texts is measured by the cosine of the angle between their respective vectors [31]. The cosine similarity is defined in Equation 2.2 from [31]. $a \cdot b = \sum_{i=1}^{N} a_i b_i$ is the dot-product. The dot-product is normalized with $\|x\| = \sqrt{x \cdot x}$ to unit Euclidean length [31]. The cosine similarity is a value between 0 and 1 for positive values [31]. According to Sidorov et al., the formula has a time and space complexity of $O(N)$ for a pair of N-dimensional vectors.

$$cosine(a, b) = \frac{a \cdot b}{\|a\| \times \|b\|} = \frac{\sum_{i=1}^{N} a_i b_i}{\sqrt{\sum_{i=1}^{N} a_i^2} \sqrt{\sum_{i=1}^{N} b_i^2}} \tag{2.2}$$

The formula Equation 2.2 assumes that the vectors, which span the VSM are orthogonal and thus, completely independent [31]. However, in practical applications, this often is not the case [31].

2.2.3 Soft Cosine Similarity

This similarity measure not only evaluates whether two texts consist of the same words but also takes into account the semantic (word-level) similarity or lexical relation of different words of the texts [6]. Hence, it improves the shortcomings of the traditional cosine similarity measure, which assumes the tokens of the vocabulary are completely independent of each other [31].

According to Sidorov et al., in order to model this additional information, more dimensions are added to the VSM. These dimensions can be obtained, for instance, by multiplying the mean of two features of one vector with the similarity between them [31]. The similarity can be calculated by using Levenshtein distance for e.g., n-grams, i.e. the number of operations necessary to convert one string into another, or using a dictionary of synonyms [31].

Since this approach no longer assumes that different words are independent of each other, the basis vectors which span the VSM are no longer orthogonal [31]. The formula for the soft cosine similarity is defined in Equation 2.3 from [31]. The similarity s_{ij} between the i-th and j-th basis vector is obtained using a similarity measure, such as synonymy [31].

$$soft\_cosine(a, b) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} s_{ij} a_i b_j}{\sqrt{\sum_{i=1}^{N} \sum_{j=1}^{N} s_{ij} a_i a_j} \sqrt{\sum_{i=1}^{N} \sum_{j=1}^{N} s_{ij} b_i b_j}} \tag{2.3}$$

According to Charlet and Damnati, the similarity between two texts is non-zero as soon as they share related words [6]. If there is no similarity between different features, the soft cosine similarity from Equation 2.3 is equal to the cosine similarity from Equation 2.2. The time and space complexity of the soft cosine similarity is $O(N^2)$ [31].

In order to reduce the complexity, Sidorov et al. propose to use a sparse similarity matrix which only stores $s_{ij} > t$, t being a threshold [31].

2.3 Embeddings

Usually, ML techniques embeddings, such as K-Means, require the text input data to be converted to embeddings [19]. Embeddings are numerical representations of words, sentences or texts. They can be used to present the textual data as vectors in a VSM. VSMs are commonly used due to their conceptual simplicity and because spatial proximity serves as a metaphor for semantic proximity [56, 5, 29]. Representations in a vector space can improve the performance in NLP tasks [21]. According to Zhang et al., when representing

text the first step is indexing, i.e. assigning indexing terms to the document. The second task is to assign weights to the terms which correspond to the importance of the term in the document. The weights assigned depend on the method and the assumptions of the model chosen to carry out the assignments.

The following section outlines a selection of embeddings. Let a corpus of documents be denoted $D = \{d_1, d_2, ..., d_M\}$, the number of documents in the dataset $M = \|D\|$, a sequence of terms w_{ij} or so-called document $d_i = \{w_{i1}, w_{i2}, ..., d_{iV}\}$, V being the length of the vocabulary, i.e. set of distinct words, of the corpus of documents [28].

2.3.1 TF-IDF

TF-IDF provides a numerical representation of a word in a document [28]. It considers the frequency of a word in a document and the frequency of a word in the whole corpus.

TF-IDF is calculated as displayed in Equation 2.4 from [28] and exemplatory in Figure 2.1. Term Frequency (TF) is computed using $TF(w_{ij}, d_i) = f_{w_{ij},d_i}$, whereas the Inverse Document Frequency (IDF) is computed using $IDF(w_{ij}, D) = \log_2 \frac{M}{M_{ij}}$, M_{ij} being the number of documents the term w_{ij} appears in. IDF measures the importance of a term w_{ij} in the corpus of documents D. The underlying assumption of IDF is that a term's importance to the data corpus is inversely proportional to its occurrence frequency [56]. In other words: Terms which appear in many documents are not as important and thus, weighted less than document-specific terms.

$$TFIDF(w_{ij}, d_i, D) = TF(w_{ij}, d_i) \cdot IDF(w_{ij}, D) \tag{2.4}$$

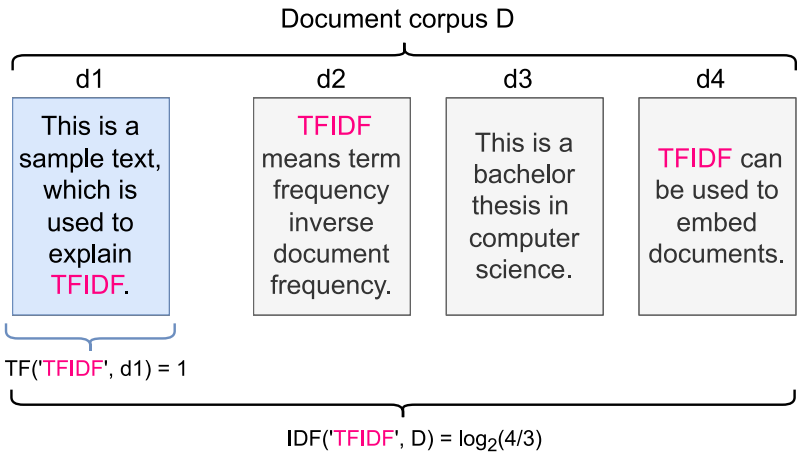


Figure 2.1: Example of calculation of TF-IDF parts: TF only considers the documents of interest while IDF incorporates the importance of the word with respect to the whole dataset.

According to Zhang et al., the computation complexity of TF-IDF embeddings is $O(V \cdot M)$ The TF-IDF more has several drawbacks [28, 56]:

- TF-IDF does not consider semantic similarities between words.
- TF-IDF does not consider the order of words in a document.
- TF-IDF often produces high dimensional representations which have to be post-processed to reduce their dimensionality, e.g., using Principal Component Analysis (PCA).
- The embeddings are not derived from a mathematical model of term distribution and may be criticised as not well reasoned.

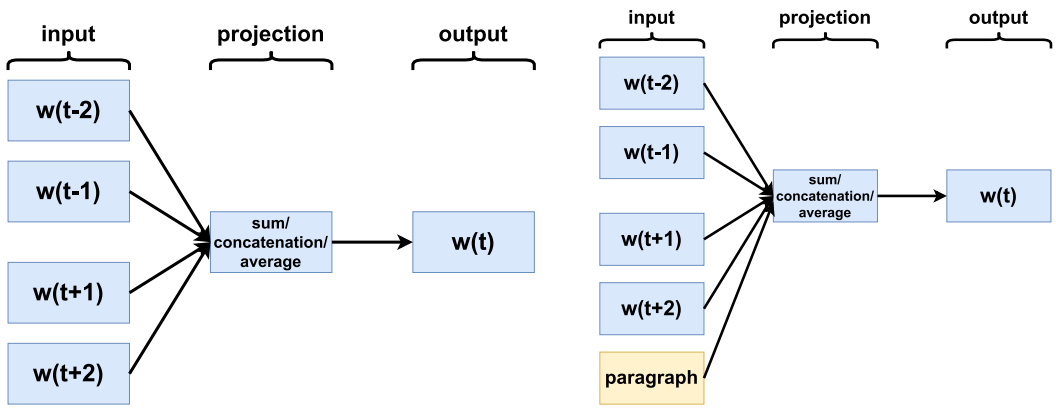
2.3.2 Doc2Vec

Another term used for Doc2Vec is *Paragraph Vector* [28, 19]. Doc2Vec addresses the problems of TF-IDF by encoding texts as $n-$dimensional vectors learnt using the words' context [28]. Hence, it preserves semantic similarities between words and encodes many linguistic regularities and patterns [21]. According to Radu et al. and Mikolov and Le, Doc2Vec learns continuous distributed vector representations for pieces of the text. The model handles inputs of different dimensions and thus, tokens are sentences, paragraphs or documents.

Doc2Vec is an adaption of the Word to Vector (Word2Vec) model, which maps words into a VSM under consideration of their semantic similarities [28]. The underlying hypothesis of both approaches is that words appearing in similar contexts are semantically similar [28]. The Word2Vec embedding is obtained using a Neural Network (NN) [28]. The NN is shallow, i.e. has only one hidden layer. This hidden layer creates the embedding of input data. There are two approaches as to how to design the architecture of the NN:

- Paragraph Vector Distributed Memory (PVDM): Predicts a word given a context [19, 20].
- Distributed Bag of Words (PV-DBOW): Predicts the context given a word [16, 21, 19].

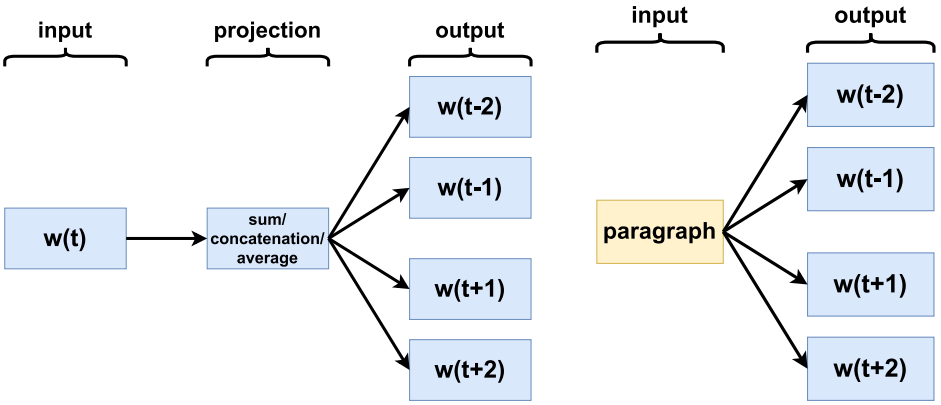
The PVDM extends the CBOW to work on a corpus of documents instead of on a set of words [28]: As usual, vectors representing the words are obtained using the CBOW model. The word vectors can be concatenated or averaged/ summed up [19]. Each document is mapped to a vector using an additional document-to-vector matrix. Both document and



(a) Continuous-Bag-of-Words (CBOW) architecture cf. [20].

(b) PVDM architecture cf. [19].

Figure 2.2: Both approaches predict the centre word using the context. PVDM is an adaption of CBOW to work on a set of documents or paragraphs instead of words.



(a) Skip-gram architecture cf. [20].

(b) PV-DBOW architecture cf. [19].

Figure 2.3: Both approaches predict the context. PV-DBOW is an adaption of Skip-gram to work on a set of documents or paragraphs instead of words.

word vectors are initialized randomly, but trained to convey meaning in terms of semantic differentiation. In order to train the model, centre words are predicted using the context, i.e. words within a sliding window, and their document, respectively represented as vectors [19]. The document vector is added to incorporate the document's topic and thus, acts like a memory. In the work of [19], the document vector is concatenated to the word vectors. The resulting vector is the prediction of the central word. The approaches are displayed in Figure 2.2. prediction, loss function? B. in paper

According to [19], both vector(s) generating models are trained using stochastic gradient descent and backpropagation. The authors of [19] also state that the document vectors are unique, while the word vectors are shared across the whole corpus.

The PV-DBOW approach is the adaption of the Word2Vec algorithm Skip-Gram and predicts the context given (a centre word and) the paragraph/ document [19]. The approaches are displayed in Figure 2.3.

2.3.3 USE

Cer et al. have published their USE models on TF Hub. They propose two models, one based on a Transformer architecture and one based on a Deep Averaging Network (DAN). Both models' input is a lowercase tokenized string. Their output is a 512-dimensional vector.

The transformer model is more accurate and more complex than the DAN model [5]. The transformer's attention is used to compute context aware word embeddings, which consider both the word order and their semantic identity. Since the sequence of word embeddings of a sentence would produce embeddings of different dimensions, the approach postprocesses the word embeddings. A sentence vector is obtained by computing the element-wise sum of the word embeddings and normalizing the result by dividing by the square root of the sentence length.

The DAN model receives embeddings of words and bi-grams as input Woher embeddings?. The embeddings are averaged and subsequently passed to a feedforward Deep Neural Network (DNN).

The models are trained on both unsupervised training data, e.g., Wikipedia, and the supervised training dataset Stanford Natural Language Inference (SNLI) [5, 29].

The transformer model is more complex than the DAN model. More specifically, the transformer model complexity is $O(n^2)$, whereas the DAN model complexity is $O(n)$, n being the number of words in the sentence [5].

The memory usage of both models is equivalent to their complexity. Cer et al. state that DAN's memory usage is dominated by the parameters used to store the embeddings of the uni- and bi-grams. Moreover, the transformer model only stores the uni-gram embeddings and thus, can require less memory than DAN for short sentences [5].

According to Reimers and Gurevych, the transformer model is used.

2.3.4 InferSent

InferSent is a sentence embedding method trained in a supervised manner on the SNLI dataset [7, 29]. The trained model is transferable to other tasks.

The SNLI dataset contains a huge data corpus of English sentence pairs. The sentence pairs are labelled with one of three categories: *entailment*, *contradiction* or *neutral*. This dataset is used because it captures Natural Language Inference (NLI) and thus, enables learning sentence semantics. To train the model, a shared sentence encoder encodes both the premise and the hypothesis to their vector representations u and v. In order to extract information about the relation of u and v, three matching methods are applied:

- (u, v): Concatenation of the two vectors.
- $u \cdot v$: Element-wise product.
- $|u - v|$: Element-wise difference of the two vectors.

The results of the matching methods are concatenated (cf. [29]). The resulting vector is then fed into a three-class classifier. The classifier consists of multiple fully connected layers and a softmax layer [7].

Conneau et al. have compared multiple architectures in their work. The bi-directional Long Short-Term Memory (BiLSTM) architecture with max pooling has been found the best option for the sentence encoder [7]. According to Reimers and Gurevych, it is a single siamese BiLSTM layer [29]. Given a sentence $(w_1, w_2, ..., w_T)$ of T words, the BiLSTM architecture computes the hidden representations h_t for each word w_t. The hidden representation h_t is the concatenation of the forward and backward hidden vectors $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$. $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ are produced by a forward and backward Long Short-Term Memory (LSTM) respectively. Hence, the sentence is read from both directions and thus, considers past and future context.

A LSTM is a Recurrent Neural Network (RNN) that is able to learn long-term dependencies. In other words: A LSTM is able to remember information as a so-called *state*. Certain LSTM mechanisms control whether the current state is deleted, whether new data is saved and to what degree the current state contributes to the current input processed in the node. Hence, LSTM nodes are not only influenced by the former output but also by their state.

Since the LSTM computes different numbers of hidden vectors h_t depending on the length of a sentence, a max pooling layer is applied to the hidden vectors. The max pooling layer selects the maximum value for each dimension of the hidden vectors.

2.3.5 Hugging face's SBERT

SBERT is an enhancement of Bidirectional Encoder Representations from Transformers (BERT). BERT is a pre-trained transformer network. It predicts a target value, for i.e. classification or regression tasks, based on two input sentences [29]. The input sentences are separated by a special token `[SEP]`. The base-model applies multi-head attention over 12 transformer layers, whereas the large model applies multi-head attention over 24 transformer layers. The final label is derived from a regression function, which receives the output of the 12th or 24th layer, respectively. Reimers and Gurevych state that BERT is not suitable for specific pair regression tasks, since the number of input sentence combinations is too big. Another shortcoming of BERT is that it does not produce independent embeddings for single sentences. More Reimers and Gurevych found that common similarity measurements, for instance, the ones discussed in section 2.2, do not perform well on the representations of sentences in a VSM produced by BERT [29].

SBERT is a modification of BERT that provides fixed-sized embeddings for single sentences [29]. It consists of a siamese and triplet network architecture. It differs from BERT in terms of architecture, since it adds a pooling layer after the BERT model. The pooling strategies compared by Reimers and Gurevych are using the output of the first/ `CLS` token, mean pooling and max pooling.

SBERT is trained on the SNLI dataset.

According to Reimers and Gurevych, SBERT outperforms InferSent and USE.

2.4 Topic Modelling

2.4.1 BERTopic

2.4.2 LDA

[28]

2.4.3 Word Clouds

frequency of words in a document

2.5 Compression of data

According to Radu et al., a decomposition of data which preserves the inner structure in clusters, improves the quality of clusters obtained by cluster methods. However, not onyl clustering improves when being applied on reasonably low dimensional data, but also other techniques to analyse data. Moreover, decompressed data is less memory consuming and often less difficult to interprete by humans since there are more methods to visualize low-dimensional data.

2.5.1 AE

The idea of this approach is to use a low-dimensional representation of the input, which is generated by an Autoencoder (AE), to reduce complexity. The high-dimensional data is encoded into a low-dimensional representation using the encoder of an undercomplete AE [22]. The low-dimensional representation can be decoded into an approximation of the high-dimensional original using the decoder of the AE.

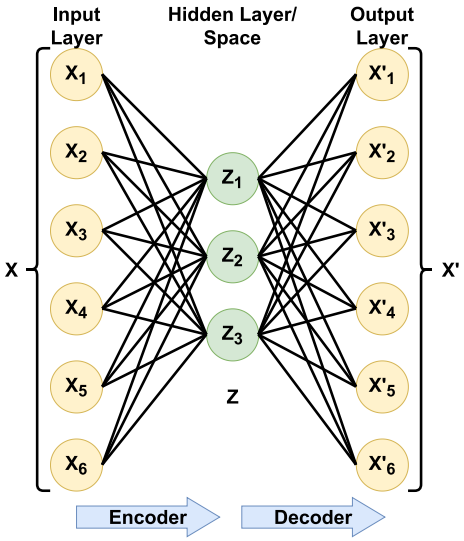


Figure 2.4: Structure of an AE cf. [22]

An undercomplete AE is a feed-forward NN, which consists of an encoder and a decoder. It learns efficient (non-correlated) encodings of the input data [22]. It is *undercomplete* because the dimensionality of the hidden layer, or so-called hidden space, is lower than the dimensionality of the input layer [13]. *Feed-forward* means that the information flows from the input layer to the output layer [13]. However, while training, the network employs backpropagation to update the parameters of the network [13].

The AE's goal is to approximate the identity function $f_\theta(X) = X$ (trivial solution eliminated) for input X and function parameters to be learned θ [13]. The input and output layers have the same dimensionality.

$$Z = f_E(W_\theta X + B_\theta) \tag{2.5}$$

The formulae for the encoder and the decoder are given in Equation 2.5 and in Equation 2.6. The parameter W_θ is the weight, whereas B_θ is the bias. The activation functions f_E and f_D are possibly non-linear and thus, the NN is capable of more than linear regression. Z is the low-dimensional representation of the input X and X' is the reconstructed version of Z.

$$X' = f_D(W_\theta Z + B_\theta) \tag{2.6}$$

The loss function L is defined as the reconstruction error between the input X and the output X' [13]. In order to train the AE, the loss function is minimized [17].

2.5.2 Eigenfaces

According to Turk and Pentland, the idea of Eigenfaces is inspired by information theory. Opposed to former approaches in the domain of face recognition which relied on the classification of images based on a set of predefined facial features, such as distance between eyes, Eigenfaces does not use predefined features [50]. The goal of this approach is to represent images using a smaller set of image features, i.e. compression to a lower-dimensional feature space, such that it is possible to distinguish between the images [50, 52]. Similar pictures, i.e. of the same person, should lie on a manifold in the lower-dimensional feature space [34]. These features do not necessarily correspond to human facial features [50]. The decomposition of input images not only reduces the complexity but also facilitates modelling probability density of a face image [34].

The input greyscale images are two-dimensional arrays of numbers: $\mathbf{x} = \{x_i, i \in \mathbf{S}\}$, $\mathbf{S}$ being a square lattice [55, 50]. The images are reshaped to an one-dimensional array $\mathbf{x} = [x_1, x_2, ..., x_n]^T \in \mathbb{R}^n$, where $n = \|\mathbf{S}\|$ and $\mathbb{R}^n$ is the n-dimensional euclidean space [55]. Some authors stress that the background is removed to omit values outside the face area [50]. In literature, typically, the original images' dimension is 512x512 [50]/ 64x64 [9], whereas the projected images' dimension is 16x16 [50]/ 250 [9]. Turk and Pentland stress that the data should be normalized, i.e. centred: $\Phi_k = \mathbf{x}_k - \psi$. Φ_k being the difference of the k-th training image and the average image $\psi = \frac{1}{N}\sum_{k=1}^{N} \mathbf{x}_k$, N being the number of training images. Some implementations of Eigenfaces, for instance, the one from sklearn, provide the normalization described above and thus, do not require the user to manually preprocess the data.

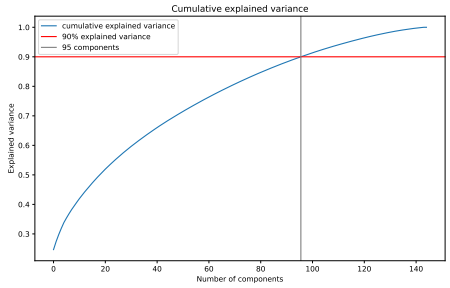
The next step is to find an alternative lower-dimensional representation of the images, which preserves most of the information of the original image. In mathematical terms, this decomposition can be expressed as $\mathbf{x} = \sum_{i=1}^{n} \hat{x}_i \mathbf{e}_i$, $\hat{x}_i$ being inner product of $\mathbf{x}$ and $\mathbf{e}_i$, $\mathbf{e}$ being an orthogonal basis [55]. If all basis vectors are used, the original image can be reconstructed using a linear combination of the basis vectors [50, 9]. The number of basis vectors is limited by the minimum of the training set size N [50] and the number of pixels n [9]. In order to compress the input from a $n-$ to a m-dimensional space, given $m \ll n$, only the first m basis vectors are used. The parameter m is chosen such that $\hat{x}_i$ is small for $i \geq m$ [55]. The compressed version of the image is denoted $\mathbf{x} \simeq \hat{\mathbf{x}} = [\hat{x}_1, \hat{x}_2, ..., \hat{x}_m]^T$. In other words: The compressed image is a vector of the first m weights of the linear combination of weight and basis vectors used to transform the image back to the original space [50]. The weights denote the position of the projection of the face images in the feature space or so-called face space spanned by the first m basis vectors [50].

In the context of Eigenfaces one basis used for decomposition is the Karhonen-Loéve (KL) basis, i.e. PCA [55, 50]. According to Zhang et al., the KL representation is optimal in the sense that it minimizes the Root Mean Square Error (RMSE) between the original image and the compressed image calculated using $m < n$ orthogonal vectors. The KL basis consists of the eigenvectors of covariance matrix $\mathbf{C} = E\left[\mathbf{x}\mathbf{x}^T\right]$ of the input images $\mathbf{x}$ [55]. Since these eigenvectors can have facial features, they are called *Eigenfaces*. The are two approaches in the literature to determine the number of Eigenfaces m used to compress the input images:

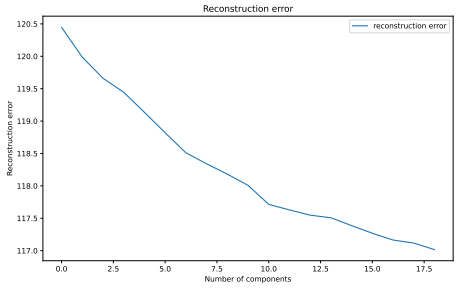
(a) The cumulative explained variance of the first $i \leq n$ eigenvectors (sorted by eigenvalues λ_i) is calculated [55, 9, 33]. The eigenvalues λ_i can be interpreted as the amount of variance explained by the corresponding eigenvector $\mathbf{e}_i$, which is equivalent to information or entropy. The user can choose how much variance, i.e. information, should be preserved, by choosing m such that the explained variance is greater than the chosen threshold. Sudiana et al. use a threshold of 90%. A plot displaying the cumulative explained variance and a threshold of 90% is shown in Figure 2.5 (a).

(b) The number of Eigenfaces m is chosen using the reconstruction error-complexity trade-off. The reconstruction error, i.e. the RMSE of the original image X and the inverse transformed image X' calculated in Equation 2.7 for different values of m. A "knee" marks the point where the reconstruction error decreases only slightly for increasing m and thus, is an indicator for the optimal m. A visualization of this approach is shown in Figure 2.5 (b).

$$\text{RSME} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - x_i')^2}{N}} \tag{2.7}$$



(a) The cumulative explained variance of the first $i \leq n$ eigenvectors (sorted by eigenvalues λ_i).



(b) The reconstruction error RMSE calculated for different values of m. Around 13 is a "knee".

Figure 2.5: Two approaches in the literature to determine the number of Eigenfaces m used to compress the input images.

In order to reduce calculation complexity, C is approximated. Zhang et al. propose the approximation $\mathbf{C} \simeq \frac{1}{N}\sum_{k=1}^{N} \mathbf{x}_k\mathbf{x}_k^T = \frac{1}{N}\mathbf{X}\mathbf{X}^T$, with $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]$, $\mathbf{x}_i \in \mathbb{R}^n$ [55].

Finding the eigenvectors of $\mathbf{X}\mathbf{X}^T$ is still computationally expensive, since $\mathbf{X}\mathbf{X}^T$ is a n by n matrix. According to Zhang et al., the eigenvectors of $\mathbf{X}\mathbf{X}^T$ can be calculated by using the eigenvectors of $\mathbf{X}^T\mathbf{X}$. The eigenvalues $\mathbf{e}_i \in \mathbb{R}^n$ of $\mathbf{X}\mathbf{X}^T$ can be derived from the eigenvectors $\mathbf{v}_i \in \mathbb{R}^N$ of $\mathbf{X}^T\mathbf{X}$ by $\mathbf{e}_i = \frac{1}{\sqrt{\lambda_i}}\mathbf{X}\mathbf{v}_i$ as discussed in more detail in [55]. Hence, the problem is reduced to a N by N matrix, which is computationally less expensive to solve, since $N \ll n$. Eigenvectors can be calculated using singular value decomposition (SVD) [55].

In the literature, face images are classified by comparing their position in the face space with those of already known faces [50]. According to [50], this approach performs well on datasets with little variation in pose, lighting and facial expression. However, Zhang et al. state, that the performance deteriorates if the variations increase since the changes introduce a bias that makes the distance function used to make classifications a no longer reliable measure.

2.6 Clustering

Clustering is used in a variety of domains to group data into meaningful subclasses, i.e. clusters [26, 10, 14]. According to Patwary et al. and Radu et al., common domains include anomaly detection, noise filtering, document clustering and image segmentation. The objective is to find clusters, which have a low inter-class similarity and a high intra-class similarity [26]. The similarity is measured by a distance function, which is dependent on the data type. Common distance functions are the Euclidean distance, the Manhattan distance and the Minkowski distance [14].

There are multiple clustering techniques, which can be divided into four categories [1]:

- **Hierarchical clustering**: Algorithms, that create spherical or convex-shaped clusters, possibly naturally occurring. A terminal condition has to be defined beforehand. Examples include CLINK, SLINK [10] and Ordering Points To Identify the Clustering Structure (OPTICS) [26].

- **Partitional based clustering**: Algorithms, that partition the data into k clusters, whereas k is given apriori. Clusters are shaped in a spherical manner, are similar in size and not necessarily naturally occurring. KMeans is a popular example of a partitional-based clustering algorithm.

- **Density based clustering**: Density is defined as the number of objects within a certain distance of each other [14]. The resulting clusters can be of arbitrary shape and size. The algorithm usually chooses the optimal number of clusters given the input data. However, some algorithms are sensitive to input parameters, such as radius, minimum number of points and threshold. Popular examples are Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and OPTICS.

- **Grid based clustering**: Similar to density-based clustering, but according to Agrawal et al. better than density-based clustering. Examples include flexible grid-based clustering [10].

Multiple approaches listed below use the term *ε-neighbourhood*, which is defined as the set of all objects within a certain distance ε of a given object [26]. In other words: $N_\varepsilon(x) = \{y \in X | dist(x, y) \leq \varepsilon, y \neq x\}$, ε being the so-called generating distance.

2.6.1 KMeans

KMeans partitions the data into $k \in \mathbb{N}$ clusters, k is given apriori [14, 28]. First, k centroids, i.e. cluster centres, are randomly initialized. Then, the objects are assigned to the closest centroid. Afterwards, the centroids are updated by calculating the mean of the assigned objects. The process is repeated until the terminating condition, for instance, no more change in the clusters, is met [14]. By iteratively reassinging the objects to the closest centroid and updating the centroids, the algorithm minimizes the within-cluster sum of squared errors E, i.e. the sum of squared (Euclidean) distances between objects in a cluster and their centroid μ_i, calculated in Equation 2.8 from [14], where C_i is the i-th cluster.

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \tag{2.8}$$

Kanagala and Krishnaiah claim, that KMeans does not identify outliers.

2.6.2 DBSCAN

The clusters identified by DBSCAN have a high density and are separated by low-density regions [14]. In order to create clusters of minimum size and density, DBSCAN distinguishes between three types of objects [14]:

- **Core objects**: An object x with at least $minPts \in \mathbb{N}$ objects in its ε-neighbourhood $N_\varepsilon(x)$, i.e. $|N_\varepsilon(x)| \geq minPts$ is true [26].
- **Border objects**: An object with less than $minPts$ objects in its ε-neighbourhood, which is in the ε-neighbourhood of a core object.
- **Noise objects**: An object, which is neither a core object nor a border object.

Kanagala and Krishnaiah define $y \in X$ as *directly density reachable* from $x \in X$, if y is in the ε-neighbourhood of core object x [14]. Moreover, a point $y \in X$ is *density reachable* from $x \in X$, if there is a chain of objects $x_1, ..., x_n$ with $x_1 = x$ and $x_n = y$, which are directly density reachable from each other as displayed in Figure 2.6 [14].

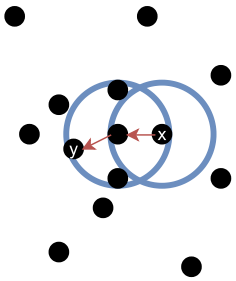


Figure 2.6: Density reachability cf. [2]. The point $y \in X$ is density reachable from $x \in X$, since there is a chain of directly density reachable objects x, o, y.

The points $x \in X$ and $y \in X$ are said to be *density connected*, if there is an object o, from which both x and y are density reachable [14]. Density connectivity is visualized in Figure 2.7.

The DBSCAN algorithm starts by labelling all objects as core, border or noise points. Then, it eliminates noise points and links all core points, which are within each other's neighbourhood [14]. Groups of connected core points form a cluster [14]. In the end, every

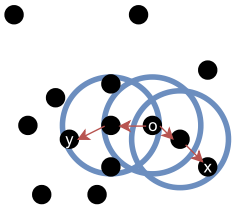


Figure 2.7: Density connectivity cf. [2]. The objects x and y are density connected since there is an object o, from which both x and y are density reachable.

border point is assigned to a cluster [14]. The non-core point cluster assigning is non-deterministic [26]. This algorithm creates clusters as a maximal set of density-connected points [14].

According to Kanagala and Krishnaiah, DBSCAN can identify outliers or noise. However, the algorithm is sensitive to the input parameters $minPts$ and ε and has difficulties distinguishing closely located clusters [14]. Moreover, if one wants to obtain hierarchical clustering, one has to run the algorithm multiple times with different ε, which is expensive in terms of memory usage [26]. According to Radu et al., DBSCAN is affected by the curse of dimensionality. Since DBSCAN relies on nearest neighbour queries and these become less meaningful in high dimensions, i.e. distances become difficult to interprete, the quality and accuracy of the results declines with increasing dimensionality [28]. Radu et al. found that their DBSCAN model assigned most documents noise, when the dimensionality was sufficiently large.

2.6.3 OPTICS

OPTICS does not return an explicit clustering, but rather a density-based clustering structure of the data, which is equivalent to clustering results of a broad range of parameters [2]. Ankerst et al. claim that real-world datasets cannot be described by a single global density, since they often consist of different local densities, as displayed in Figure 2.8.

Opposed to DBSCAN, OPTICS is able to detect clusters of varying densities [10]. OPTICS produces an order of the elements according to the distance to the already added elements [10, 26]: The first element added to the order list is arbitrary. The order list is iteratively expanded by adding the element of the ε-neighbourhood to the order list, which has the smallest distance to any of the elements already in the order list. Hence, clusters with higher density, i.e. lower ε, are added first (prioritized) [14, 2]. When there are no more elements in the ε-neighbourhood to add, the process is repeated for the other clusters. The non-core point cluster assigning is non-deterministic [26].

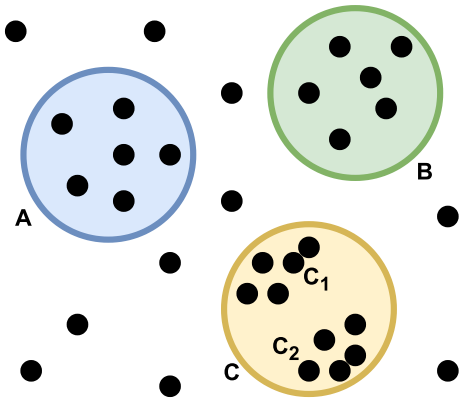


Figure 2.8: Clusters of different densities cf. [2]. Since C_1 and C_2 have different densities than A and B, a clustering algorithm using one global density parameter would detect the clusters A, B and C, rather than A, B, C_1 and C_2 .

$$RD(y) = \begin{cases} \text{NULL} & \text{if } |N_\varepsilon(x)| < minPts \\ max(core\_dist(x), dist(x,y)) & \text{otherwise} \end{cases} \tag{2.9}$$

OPTICS saves the reachability distance $RD(y)$, as calculated in Equation 2.9 from [26], with core distance core_dist being the minimal distance ε^{min} such that $|N_{\varepsilon^{min}}(x)| \geq minPts$ (the distance to the $minPts^{th}$ point in N_ε) or NULL else, of each element to its predecessor in the order list and thus, a representation of the density necessary to keep two consecutive objects in the same cluster [26]. If $\varepsilon < RD(y)$, then y is not density reachable from any of its predecessors and thus, one can determine whether two points are in the same cluster for given information saved by OPTICS [26, 2]. If the core distance of an element is not NULL, i.e. it is a core object, and it is not density reachable from its predecessors, it is the start of a new cluster [2]. Otherwise, the element is a noise point [2]. According to Patwary et al., the algorithm builds a spanning tree, which enables obtaining the clusters for a given ε by returning the connected components of the spanning tree after omitting all edges with $\varepsilon < RD(y)$ [26]. The relationship between ε, cluster density and nested density-based clusters is displayed in Figure 2.9.

Hence, this procedure enables the extraction of clusters for arbitrary $0 \leq \varepsilon_i \leq \varepsilon$ [14, 2]. According to Patwary et al.'s work, even though the clustering algorithm is expensive the extraction only needs linear time. According to [2], the algorithm yields good results if the input parameters $minPts$ and ε are "large enough" and thus, the algorithm is rather insensitive to the input parameters.

The smaller ε is chosen, the more objects will be identified as noise and thus, the algorithm will not identify clusters with low density, since some objects only become core objects for a larger ε [2]. According to Ankerst et al., the optimal value for ε creates one cluster for most of the objects with respect to a constant $minPts$, since information about all

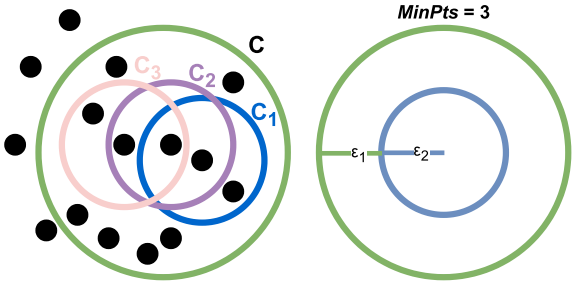


Figure 2.9: The relationship between ε, cluster density and nested density-based clusters cf. [2]. For a constant $minPts$, clusters with higher density such as C_1, C_2 and C_3, i.e. a low ε_2 value, are completely contained in lower density clusters such as C given $\varepsilon_1 > \varepsilon_2$. This idea forms the basis of OPTICS of expanding clusters iteratively and thus, enables the detection of clusters for a broad range of neighbourhood radii $0 \leq \varepsilon_i \leq \varepsilon$.

density-based clusters for $\varepsilon_i < \varepsilon$ is preserved. A heuristic for choosing ε based on the expected k-nearest neighbour distance is presented in [2].

High values for $minPts$ smoothen the reachability curve, even though the overall shape stays roughly the same [2]. According to Ankerst et al., the optimal value for $minPts$ is between 10 and 20.

2.7 Database Elasticsearch

Elasticsearch is a widely used non-relational database, which was designed to store and perform full-text search on a large corpus of unstructured data [51]. This open-source distributed document-driven database system is built in Java and is based on the Apache Lucene (Java) library for high-speed full-text search [51, 54]. According to Zamfir et al., Elasticsearch provides Wikipedia's full-text search and suggestions as well as Github's code search and Stack Overflow's geolocation queries and related questions. It enables near real-time search by index refreshing periods of one second. Needless to say, Elasticsearch is qualified to handle Big Data.

Elasticsearch is a document store, which stores schemaless key-value pairs called documents [11]. Elasticsearch's entries, i.e. documents, are stored in logical units, so-called indices. As stated by Zamfir et al. and Voit et al., the indices are structured similarly to Apache Lucene's inverted index format. An index can be spread into multiple nodes. A node is a single running instance of Elasticsearch [54]. An index is divided into one or more shards, which can be stored on different servers and enable parallelization [54].

Elasticsearch indices' entries are documents, which are saved in a JavaScript Object Notation (JSON) format [51]. A document's fields and field types are defined by the user

when initializing the database index. By default, every field of a document is indexed and searchable [54].

Replicas are copies of shards, which create redundancy and thus, ensure availability [54].

By specifying the unique `_id` of a document and the database `index`, it is possible to retrieve a specific document from the database using the `GET Application Programming Interface (API)`. The query is real-time by default. The parameters `_source_excludes` or `_source_includes` may be used to exclude or include specific fields of the document in the response [36].

The keyword used when performing a full-text search is `match`. To query for a specific value, one has to specify the `<field>` of interest and the query value.

Elasticsearch preprocesses the query value before starting the search [42]. The default preprocessing steps of the so-called default analyzer include tokenization and lowercasing [42]. Omitting stop words is disabled by default, but it is possible to provide custom stop words or use the English stop word list [42]. It is possible to create custom tokenizers, which split the query value into tokens of a certain maximum length.

Another useful feature of Elasticsearch is the multi-term synonym expansion. When the user queries a specific phrase Elasticsearch expands the query to include synonyms of the query terms [41]. The maximum number of expansion terms is set to 50 by default but can be configured by the user [40]. By default, the multi-terms synonym expansion option is enabled [40].

Elasticsearch also provides the option to perform fuzzy matching instead of exact search. By enabling the fuzzy matching option, a Elasticsearch query consisting of for instance, *Bahama* returns documents that have the word *Bahamas*. By default, this option is not enabled but can be enabled and configured individually by the user [40].

Another search option of Elasticsearch is the k-nearest neighbor (kNN) search. The return value of a kNN search is the `k` nearest neighbors in terms of a certain distance function of a query vector [18]. The query is a dense vector of the same dimension as the (dense) vectors stored in the database. According to [38], the kNN either returns the exact brute-force nearest neighbors or approximate nearest neighbors calculated by the Hiercharical Navigable Small World (HNSW) algorithm [18, 38]. HNSW is a graph-based algorithm [18]. The term `navigable` refers to the graphs used, which are graphs with (poly-)logarithmic scaling of links traversed during greedy traversal concerning the network size [18]. The idea of a `hiercharical` algorithm is to create a multilayer graph, grouping links according to their link length, as displayed in Figure 2.10. The search starts on the uppermost layer, i.e. the layer containing the longest links, greedily traversing the layer until reaching the

local minimum. It uses this local minimum as the starting point at the next lower layer and the process is repeated until the lowest layer is reached [18]. The layers of the graph are built incrementally, and a neighbour selection heuristic, as depicted in Figure 2.11, not only creates links between close elements, but also between isolated clusters to ensure global connectivity [18].

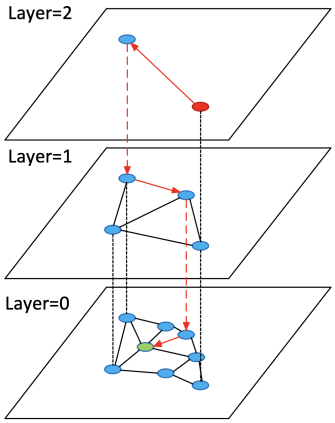


Figure 2.10: Structure of HNSW layer from [18]. The search starts on the uppermost layer, i.e. the layer containing the longest links, greedily traversing the layer until reaching the local minimum. The local minimum is used as the starting point at the next lower layer and the process is repeated until the lowest layer is reached.

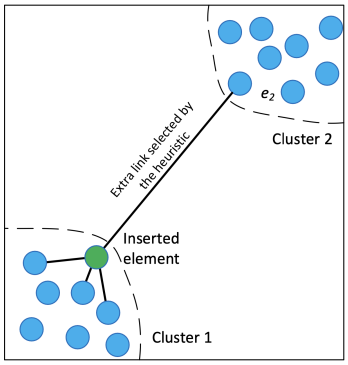


Figure 2.11: Neighbour selection heuristic of HNSW from [18]. The heuristic creates diverse links, i.e. links between close elements (e.g., green circle and elements in cluster 1) and between isolated clusters (e.g., green circle and e_2) to ensure global connectivity.

In order to perform the kNN search on a `<field>` it has to be of type `dense_vector`, indexed and a `similarity` measure has to be defined when initializing the database [38].

Elasticsearch's kNN implementation not only allows literal matching on search terms but also semantic search [38].

Besides Elasticsearch, the elastic stack offers other tools, for instance, Kibana, which provides a user interface to manage different models. After saving a model in Kibana, it is possible to create a text embedding ingest pipeline, which embeds new documents or reindexes existing documents [39].

2.8 Flask

Flask is open source and written in Python by Armin Ronancher in 2004 [3, 25]. According to Copperwaite and Leifer and Mufid et al., Flask is one of the most popular Python web frameworks. It provides powerful libraries for core functionality such as routing, templating, and Hypertext Transfer Protocol (HTTP) request parsing [8]. It is extensible and thus, can be extended with additional plugins without affecting the internal structure of the existing system [3].

Flask uses the Jinja Template Engine for template files including Hypertext Markup Language (HTML) pages, whereas static files such as Cascading Style Sheet (CSS) files are handeled using the Werkzeug WSGI toolkit [3]. According to Aslam et al., Jinja is modeled after the Django template system. Werkzeug implements, for instance, requests and response objects [25].

All requests received from clients are passed to an instance of the Flask application [12]. Hence, the first step is to create an instance of the Flask class, such as done in Listing 2.1.

```
1 app = Flask(__name__)
```

Listing 2.1: Initialization of Flask application instance.

Clients send requests to the web server, which passes them to the Flask application instance. The queries are then routed to the corresponding functions. Routing is the process of mapping Uniform Resource Locator (URL) paths to functions [12]. To define a route, the `route` decorator is used as displayed in Listing 2.2.

```
1 @api.route('/documents/<id>', endpoint='document')
2 class Document(Resource):
3     def get(self, id):
4         elastic_search_client = Elasticsearch(CLIENT_ADDR)
5         return query_database.get_doc_meta_data(elastic_search_client,
6             doc_id=id)
```

Listing 2.2: Exemplartary definition of a function to display routing with Flask. The `route` decorator is used to define the URL path.

URL can contain dynamic components, which are enclosed in <> angle brackets. The values of these components are passed to the function as arguments [12]. By default, dynamic components are of type `string`. However, other types including `int` and `float` are supported [12].

During development, the Flask application can be run using `flask run` to start the built-in development web server [12]. By enabling debug mode, the server automatically reloads the application when changes are detected [12].

An endpoint is a class with certain methods, which can be accessed using HTTP requests. Every endpoint can have a `GET`, `PUT` and a `DELETE` method [11]. The `GET` method is used to retrieve data from the server, whereas the other methods are used to either insert or delete data.

2.9 Angular

Angular is a framework for building web applications. It uses Node.js and TypeScript. Usually, the source code is structured into different modules, including components and services. Components are used to define the appearance of the application, while services are used to define the logic of the application and communicate with the backend.

Angular applications are created using the `ng new NAME` command line interface [30]. This command creates a skeleton, which can be customized to the needs of the application. By running `ng serve` the application can be served locally.

3 Implementation

3.1 Slurm

Slurm is an open-source management tool for Linux clusters [47]. It allocates resources, i.e. compute nodes, and provides the means to start, execute and monitor jobs [47, 53].

The so-called slurm daemons control nodes, partitions, jobs and job steps [47]. According to TODO, a partition is a group of nodes and a job is the allocation of resources, i.e. compute nodes, to a user for a limited period of time. A basic visualization of the architecture is given in Figure 3.1.

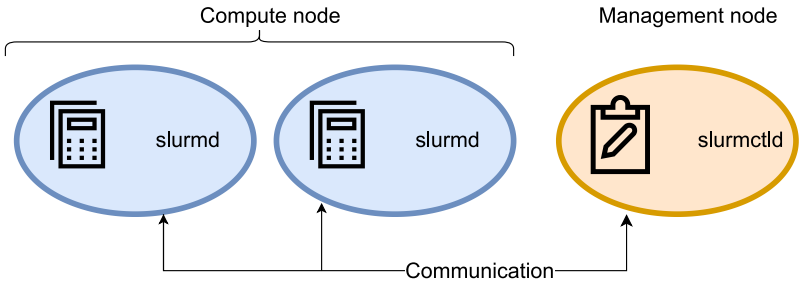


Figure 3.1: Slurm architecture. The management node has a `slurmctld` daemon, while every compute node has a `slurmd` daemon. The nodes communicate. The user can use certain commands, for instance `srun` and `squeue`, anywhere on the cluster.

3.2 Elasticsearch

In this work, the database is filled once with data from a large unstructured corpus of Portable Document Format (PDF) files. After the initialization of the database, it is used for queries. Therefore, the workflow is completely offline.

The index *Bahamas* stores different embeddings of the text layer information and metadata of the documents. As depicted in Figure 3.2, not only textual information is stored in the database, but also the images of the first page of the PDFs. The structure of the index is presented in Table 3.1.

Table 3.1: Fields in Elasticsearch database in index *Bahamas*.

field name	field description
_id	Unique identifier of document `i`. The identifier is generated by the sha256 hash algorithm from hashlib.
doc2vec	55 dimensional doc2vec embedding of `i`.
sim_docs_tfidf	sim_docs_tfidf embedding + all-zero flag of `i`. The all-zero flag is one if the TF-IDF embedding consists of only zeros, zero else.
google_univ_sent_encoding	512 dimensional google_univ_sent_encoding embedding of `i`.
huggingface_sent_transformer	384 dimensional huggingface_sent_transformer embedding of `i`.
inferSent_AE	inferSent_AE embedding of `i`. Since the pretrained infersent model embedding's dimension is 4096, the encoder of a trained AE is added to reduce the dimension to 2048.
pca_image	Two dimensional PCA version of first page image of `i`.
pca_kmeans_cluster	Cluster of `i` identified by KMeans on PCA version of image.
text	Text of `i`.
path	Path on local maschine to `i`.
image	Base64 encoded image of first page of `i`.

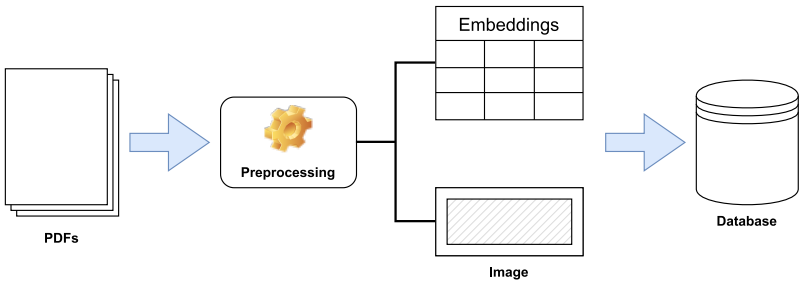


Figure 3.2: PDFs to Database. First, the data is preprocessed: The first page of a PDF file is converted to an image and the complete text is extracted. The images are stored in the database as well as the text and different embeddings of the text.

The default analyzer is used for the full-text search, since for instance configuring a maximum token length did not seem necessary or likely to improve the results.

```
results = elastic_search_client.search(
    index='bahamas',
    size=count,
    from_=(page*count),
    query= {'match' : {
                'text': {   'query':text,
                            'fuzziness': 'AUTO',}
            },
        }, source_includes=SRC_INCLUDES)
```

Listing 3.1: Exemplartary query to an Elasticsearchdatabase index. The number of results to return `size` and the start index of the results `from_` is defined. To enable fuzzy search a value for `fuzziness` has to be defined.

Moreover, the fuzzy matching option is set to `AUTO`, which means in terms of keyword or text fields that the allowed Levenshtein Edit Distance, i.e. number of characters changed to create an exact match between two terms, to be considered a match, is correlated to the length of the term [35]. By default, terms of length up to two characters must match exactly, terms of length three to five characters must have an edit distance of one and terms of length six or more characters must have an edit distance of two [35]. An exemplartary query, which uses fuzzy search is given in Listing 3.1.

According to Malkov and Yashunin, one of kNN search's use cases is semantic document retrieval, which makes it a good fit for this task. In this work, the approximate nearest neighbors search is used, since it is faster and the results are good enough for the use case of this work. The similarity measure used in this work is the cosine similarity, which calculates the `_score` of a document according to Equation 3.1 from [37], where `query` is the query vector and `vector` is the vector representation of the document in the database. The other similarity measures provided by Elasticsearch are `l2_norm` or so-called Euclidian distance and `dot_product` which is the non-auto-normalized version of the `cosine` option. Since cosine is not defined on vectors with zero magnitude, embeddings that can return all zero vector representations, such as sim_docs_tfidf, are enhanced with an all-zero flag before inserting them into the database.

$$\frac{1 + \text{cosine}(\text{query}, \text{vector})}{2} \tag{3.1}$$

In this work, the only tool from the elastic stack used is Elasticsearch. Without Kibana, the used models are saved on disk as Pickle (PKL) files. Consequently, instead of using the kNN query structure for semantic search on embeddings provided by Elasticsearch, the normal kNN search on a field that contains an embedding is used.

3.3 Eigendocs

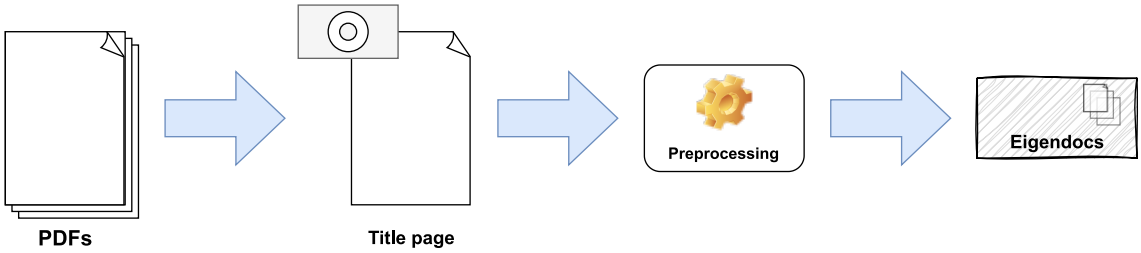


Figure 3.3: From PDFs to Eigendocs. Firstly, the first page of a document is converted to an image. Then the image is preprocessed: It is placed on a white canvas, to ensure all images have the same dimensions. Moreover, it is converted to greyscale and normalized to values between zero and one. Afterwards, the 2d image is reshaped to a 1d array. Lastly, the image is compressed using Eigenfaces.

In this work, the Eigenfaces approach from subsection 2.5.2 is used to compress the images of the first page of documents. The idea is that documents not only hold textual information but also visual information, such as layout, company logo or signature. By mapping those images on a subspace, they ought to be grouped by visual similarity. The procedure of the eigenface adaption *eigendocs* is displayed in Figure 3.3.

The documents are first read from a directory. Subsequently, their first page is converted to an image and saved. When initially filling the database, these images are read from their directory. Firstly, the maximum height and width of all images in the corpus is calculated. These dimension are used to create a white canvas for each image which forms the background. Every image is placed in the upper left corner. Hence, scaling is not necessary and thus, the portion of white pixels on the right and bottom side encodes the dimension of the former image. Therefore, the relative size of images in the corpus is incooperated in the resulting representation fo the input images.

```
1 C = np.ones((max_w,max_h))
2 C[:doc.shape[0],:doc.shape[1]] = rgb2gray(doc)
3 documents.append(C.ravel())
```

Listing 3.2: Preprocessing of the input images from Dr. Christian Gruhl. The background is a white canvas. The images are converted to one-dimensional greyscale values.

Afterwards, they are converted to greyscale images using 3.3. Before returning the image, the two-dimensional image vectors are converted to one-dimensional ones as displayed in the last line of 3.2. The decomposition is transformed using PCA as displayed in 3.4.

```
0.299*img[:,:,0] + 0.587*img[:,:,1] + 0.114*img[:,:,2]
```

Listing 3.3: Conversion of RGB pixel values to greyscale from a script by Dr. Christian Gruhl.

```
pca = decomposition.PCA(n_components=n_components, whiten=True,
    svd_solver="randomized")
```

Listing 3.4: Initialization of the PCA instace used to compress the image data. In order to work according to the Eigenfaces approach a svd_solver has to be used.

3.4 Autoencoder

In this work, the AE is used to reduce the dimensionality of the InferSent embedding. Since the InferSent model is pretrained, it is not possible to change the dimensionality of the embedding without a considerably big effort, i.e. retraining the model on a sufficiently large data corpus and reconfiguring the model's parameters. Moreover, retraining the model would destroy the purpose of its presence in this work, which is to provide a pretrained model and thus, reducing the complexity of training an own model. Therefore, it is not feasible to change the dimensionality of the InferSent embedding, but rather adding a supplementary layer after the model produce the final embedding. Hence, the idea is to use the encoder of an AE to reduce the dimensionality of the InferSent embedding.

The implementation was provided by a blog post using keras and adapted to fulfil the needs of the specific context. Hence, The dimensionality of the layers was increased since the AE of the blog post has a smaller dimensionality in the latent space. More layers were added to the encoder and decoder to improve the ability of the model to reconstruct. Figure 3.4 Figure 3.5

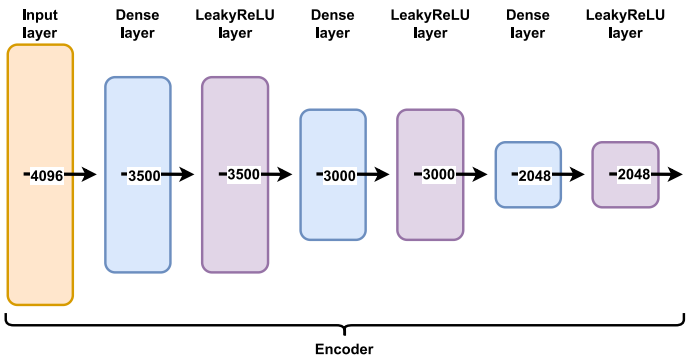


Figure 3.4: Architecture of the encoder of the AE

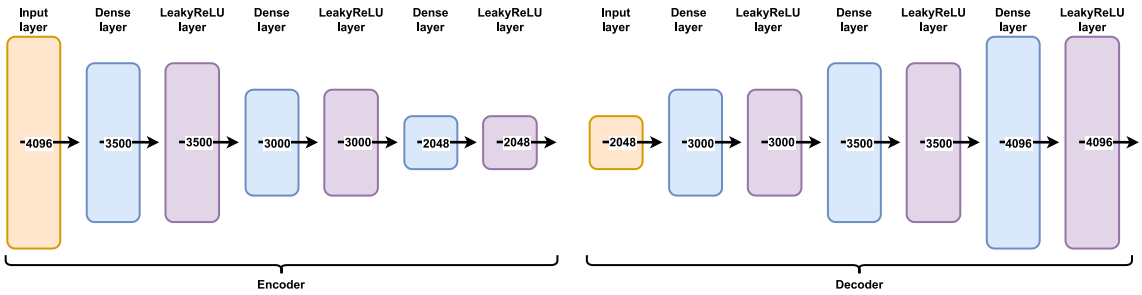


Figure 3.5: Architecture of the AE

3.5 TF-IDF

The TF-IDF model has to be initialized and trained on the data corpus to build the data-specific vocabulary. An exemplary implementation is given in Listing 3.5. The `TfidfVectorizer` is provided by the `scikit-learn` package. When initializing the model, the parameters define not only the input type but also the way the data is preprocessed. The `input` parameter defines the input type, i.e., `content` means that the input is a list of strings or bytes, whereas `file` assumes the input has a `read` method and `filename` denotes a list of filenames as input [49].

The `preprocessor` parameter defines the preprocessing (string transformation) stage. It is possible to override the default with a custom preprocessing function. The parameters `min_df` and `max_df` define the minimum and maximum document frequency of a word in the corpus. The default values are 1, i.e. a term has to appear at least once, and 1.0, i.e. a term appears at most in all documents, respectively [49].

The implementation of TF-IDF in `scikit-learn` is different from the original TF-IDF definition. By default, the `scikit-learn` implementation uses the `norm='l2'` parameter, i.e. the Euclidean norm [48]. The difference is the calculation of the IDF part, which is given in Equation 3.2 from [48]. The one is added to M_{ij} due to the parameter `smooth_id=True` by default to prevent zero divisions. After calculating the TF-IDF values, they are normalized by the Euclidean norm $v_{norm} = \frac{v}{\|v\|_2} = \frac{v}{\sqrt{v_1^2+v_2^2+...+v_M^2}}$.

$$\text{idf}(w_{ij}) = \log \frac{1+M}{1+M_{ij}} + 1 \tag{3.2}$$

In this work, the text of the PDFs is first extracted, then preprocessed using a custom preprocessor and afterwards embedded using the `TfidfVectorizer`, which returns the TF-IDF weights as embedding. Before storing the TF-IDF weights in the database, they are enhanced with an all-zero flag. The all-zero flag ensures that no all-zero vectors are stored in the database by enhancing those that have a zero magnitude with a "1" entry and "0" otherwise. All-zero TF-IDF weights indicate that a document does not have any

```
tfidf = TfidfVectorizer(input='content',
            preprocessor=TfidfTextPreprocessor().transform, min_df=3,
            max_df=int(len(docs)*0.07))
tfidf.fit(documents)
```

Listing 3.5: Initialization of the TF-IDF model. Firstly, an instance of the `TfidfVectorizer` class is created. Secondly, the `fit` method is called to fit the model to the documents.

terms with the vocabulary in common. Since the vocabulary is kept relatively small with respect to the number of different words in the data corpus to reduce the dimensionality of the embeddings, it is not unlikely that a document does not contain any of the vocabulary terms. The all-zero flag is necessary because the cosine similarity used to query for similar documents in the database cannot handle vectors of zero magnitude. The pipeline in Figure 3.6 visualizes the steps.

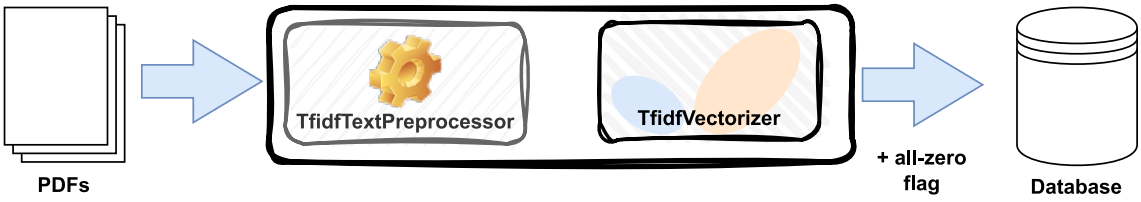


Figure 3.6: TF-IDF pipeline. Firstly, the text extracted from the documents is preprocessed using a custom preprocessor. Then the TF-IDF are obtained from the `TfidfVectorizer`. Lastly, the all-zero flag is added to the TF-IDF weights and they are stored in the database.

In this work, a custom preprocessing function is used. The preprocessing steps are visualized in Figure 3.7 on an exemplary text. Firstly, the accents are stripped from the text. Then, all new line symbols are replaced with a whitespace. Afterwards, the text is converted to lowercase. Then the numbers are discretized, i.e. all numbers between 0 and 99999 are replaced with the string `SMALLNUMBER`, numbers bigger than 99999 are replaced with the string `BIGNUMBER` and floats are replaced with the string `FLOAT`. The next step is to remove all punctuation symbols. After that, the symbols for numbers are enclosed with pointed brackets, e.g. `<SMALLNUMBER>`. Then, the text is tokenized, i.e. splitting at whitespaces, and stop words are omitted. The stop word corpus used is provided by the `nltk` package and consists of common English stop words. Afterwards, the tokens are lemmatized. The lemmatizer used is `WordNetLemmatizer` from the `nltk` package. In the end, the tokens are joined to a string again.

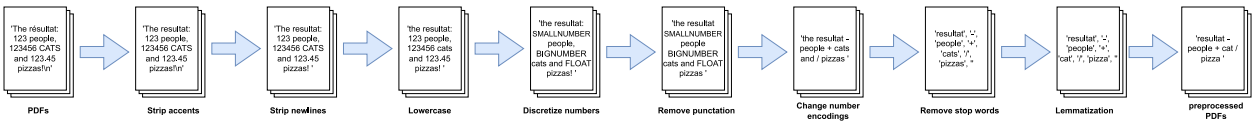


Figure 3.7: TF-IDF Preprocessing visualized using a example text. TODO: nicht aktuell bzgl. small number - etc.

3.6 Doc2Vec

The library `gensim` provides the Doc2Vec model used in this work. The input data to initialize the model has to be of type `tagged Documents`, which are documents with (numerical) tags. The parameter `dm` determines the training algorithm used. The value `dm=1` specifies the PVDM algorithm, while `dm=0` specifies the PV-DBOW algorithm [44]. The default algorithm, i.e. PVDM, is used in this work [46]. The parameters `vector_size` and `window` define the dimensionality of the embeddings and the size of the window, i.e. the maximum distance between the current and the predicted word, respectively. The default value for `vector_size` is 100, whereas the default window size is 8 [46, 45]. The `min_count` parameter specifies a threshold below which words will be ignored. Its default value is 5. The `workers` parameter specifies the number of threads to be used for training. The default value is 1 [46]. The `epochs` parameter specifies the number of iterations over the corpus. The default value is 10. By default, the hierarchical softmax algorithm, i.e. `hs=1`, is used for training [44]. Many Doc2Vec default values are adopted from Word2Vec since the `gensim` Doc2Vec implementation inherits from the Word2Vec implementation.

3.7 InferSent

The InferSent model is based on PyTorch [29]. The parameters used to initialize the model are presented in Listing 3.6. The parameter `version` indicates whether the model was trained with GloVe or fastText for 1 or 2 respectively. Since the model is precomputed, it is not possible to change certain parameters, such as the word embedding dimension `word_emb_dim` or the dimension of the output vectors `enc_lstm_dim`.

```
1 {'bsize': 64, 'word_emb_dim': 300, 'enc_lstm_dim': 2048, 'pool_type': 'max',
2 'dpout_model': 0.0, 'version': 1}
```

Listing 3.6: Parameters of the InferSent model.

The steps necessary to create a working instance of the InferSent model are presented in Listing 3.7. After the InferSent model is initialized, the `state_dict` of the model is loaded. This dictionary consists of learnable parameters, i.e. weights and bias, of the model. The next step is to set the path to the word embeddings. Finally, the vocabulary of the model is built or more precisely, only those embeddings needed are kept while the rest is discarded.

W2V path The InferSent model is based on GloVe word embeddings. GloVe is an unsupervised learning algorithm for obtaining vector representations of words. It is possible to download embeddings computed by GloVe, instead of using the algorithm to generate

```
infersent = InferSent(params_model)
infersent.load_state_dict(torch.load(model_path))
infersent.set_w2v_path(w2v_path)
infersent.build_vocab(docs, tokenize=True)
```

Listing 3.7: Initializing the InferSent model.

them. The precomputed word embeddings are stored in a 5.65 GB text file. The file contains 840 B tokens and a vocabulary of 2.2 M cased 300-dimensional vector representations of words each [43]. TODO: glove info [27] GloVe introduces bias in terms of ageism, racism and sexism into the model [5].

In this work, a custom set of vector representations of words is used. The custom word embeddings are computed by a Word2Vec model trained on a selection of 195 documents from the Bahamas dataset. The only parameter which differs from the default settings is the `vector_size` which is set to 300. After the Word2Vec model is trained, the word embeddings are saved in a file. The file is post-processed to be compatible with the InferSent model. To be more precise, only lines that consist of at least two whitespace-separated char sequences are kept. Usually, word embeddings stored in a text file are structured in a way that the first char sequence is the word and the following numbers are the vector representation of the word.

3.8 USE

The USE model is based on TensorFlow [29].

3.9 SBERT

The SBERT model is based on PyTorch [29].

3.10 Clustering using OPTICS

Similar to the approach from [2], OPTICS was used to cluster the images of the first page of documents in this work. The procedure is displayed in Figure 3.8. There were two different preprocessing approaches:

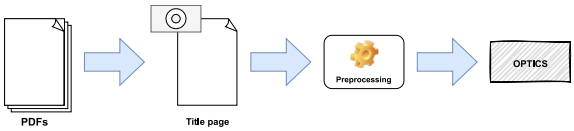


Figure 3.8: First, the first page of each document is converted to an image. Then the image is preprocessed, i.e. conversion to greyscale and resizing.

1. The images were first preprocessed to 32x32 normalized greyscale pixels (cf. [2]) as visualized in Figure 3.9 and afterwards compressed to 13-dimensional vectors using PCA.

2. The technique Eigendocs from subsection 2.5.2 was used to compress the images to 13-dimensional normalized greyscale images as displayed in Figure 4.1.

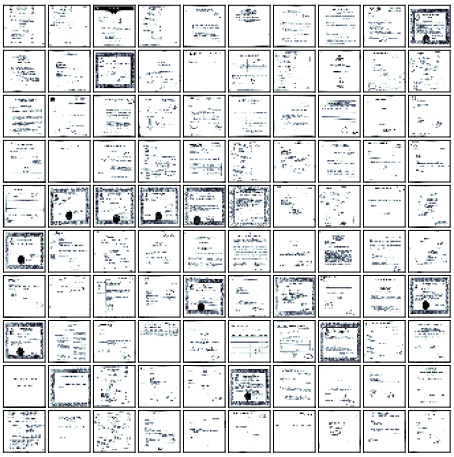
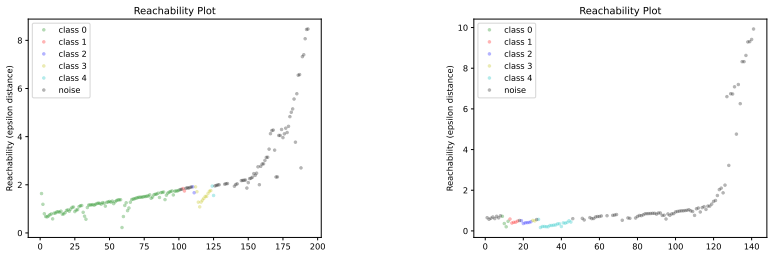


Figure 3.9: The first 100 documents of the dataset compressed to 32x32 greyscale pixels.

The reachability distance ordered by OPTICS is displayed in Figure 3.10. The resulting clusters are displayed in Figure 4.2.



(a) The reachability plot of the documents preprocessed according to item 1.

(b) The reachability plot of the documents preprocessed according to item 2.

Figure 3.10: The plot was created using the OPTICS algorithm from the Python library scikit-learn. It shows the reachability distance of each document to its predecessor in the order list.

The configurations used when initializing an OPTICS model greatly influence the clusters returned. The user has to consider the increase of complexity when choosing a high `max_eps` value. The way the reachability plot is used to extract cluster is dependent on the `cluster_method`, whereas `min_smaples` and `eps` influence the cluster sizes and number of cluster found for a given clustering approach. The code to initialize an exemplartary OPTICS model is displayed in Listing 3.8.

```
optics_model = OPTICS(cluster_method='dbscan', min_samples=2, max_eps=10,
    eps=0.5)
```

Listing 3.8: Initialization of the OPTICS model. One can choose either `dbscan` or `xi` as clustering method. The number of minimum samples in a cluster corresponds to *minPts*. The parameter `max_eps` is infinity as default, but can be specified by the user to reduce complexity and runtime. According to literature, `max_eps` should be big enough to include almost all points in a cluster. The value of `eps` define the distance between two points to still be considered neighbors and can be chosen consulting the reachability plot.

3.11 User Interface

Since this work should be valuable to (German) tax offices, a basic user interface is provided. However, the focus of this work is on the methods and not on the user interface. The user interface is divided into two parts, the frontend and the backend.

3.11.1 Backend

The framework used for the backend is Flask. In this work, only the `GET` method is used. There are multiple endpoints, which are used to retrieve data from the server:

- Documents: Returns a list of documents, which best match the query. The query can be of type `match_all`, which returns all documents in the database, or a fuzzy full-text query, or a kNN query on a certain field of the database. Moreover, the number and start index of the results returned can be specified.

- Document: Returns the document with the specified `id`.

- PDF: Returns the path to a PDF file. In order to access the path information a query for a document with the specified `id` is performed.

- WordCloud: Returns the bytes of a WordCloud image. Depending on additional parameters, the WordCloud is either generated from one document or the most similar documents to the query field, identified by kNN.

- Term Frequency: Returns the term frequency calculated for the specified document.

In order to test the endpoints during development, swagger documentation for every endpoint is provided.

3.11.2 Frontend

The framework used for the frontend is Angular. There are two main components, which are used to display the data:

- Home: The home component is used to display the results of a query. It consists of a search bar, which is used to enter the text query, and a list of results. If no text query is entered the first documents of the database, i.e. the result of a `match_all` query, are displayed. The search component is shown in Figure 3.11.

- Detail: The detail component is used to display the details of a document. The document name and ID are located on the left side of the screen. Beneath the document name and ID, a button which opens the term frequency image on a new page upon pressing is located. Moreover, the WordCloud of the document is displayed. The WordCloud is generated from the text of the document. On the right side of the screen, there is a PDF viewer which displays the pages of the document. Beneath the PDF viewer, the names and WordClouds of the most similar documents are displayed after a query for them is initiated by the user. The detail component is shown in Figure 3.12.

To change between the components, the routes have to be defined. The routes are defined in the `app-routing.module.ts` file, as shown in 3.9.

```
1 const routes: Routes = [
2     { path: ':id', component: DocumentDetailComponent},
3     { path: '', component: HomeComponent},
4   ];
```

Listing 3.9: Definition of routes in Angular in the app-routing.module.ts.

3.12 Trade-off between memory and query time

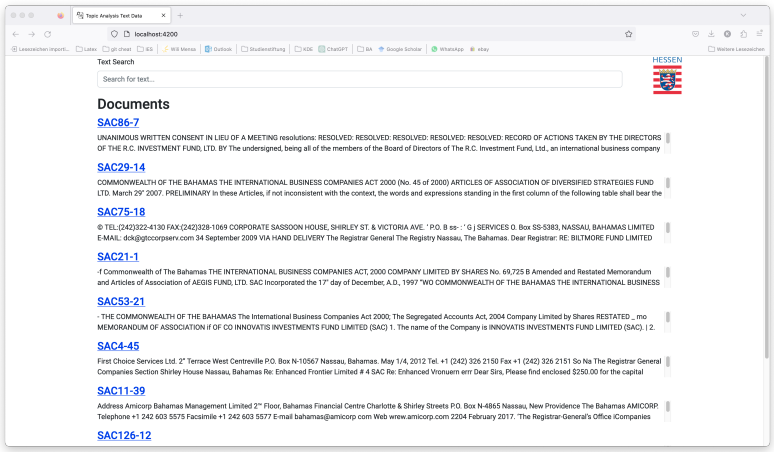


Figure 3.11: Home component of the frontend. The search bar is used to enter the text query. The results of the query are displayed below the search bar.

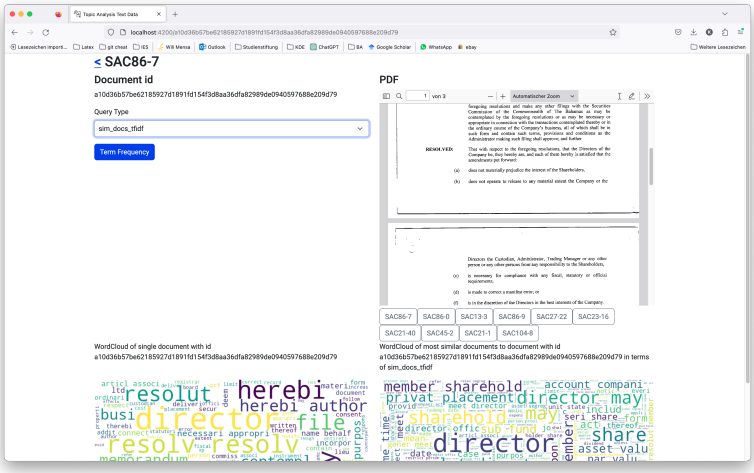


Figure 3.12: Detail component of the frontend. The chosen document is displayed, as well as its most similar documents in the database. WordClouds of the document and the most similar documents are displayed.

4 Evaluation

Parameters of models

4.1 Similarity measurements

According to Reimers and Gurevych, the similarity measurements discussed above obtained roughly the same results in their experiments [29].

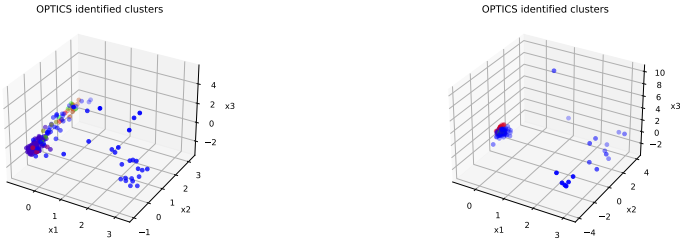
4.2 Eigendocs

In order to determine the optimal number of components used for Eigendocs the cumulative explained variance and the reconstruction error were plotted as displayed in Figure 2.5 from subsection 2.5.2. The first plot indicated that 90% of the variance is explained by 95 components. Usually, that would have been the number of dimensions of the subspace onto which the documents would have been projected. However, when working with cluster algorithms like OPTICS prior to this step, the number of dimensions should be reduced even further to achieve valid clusters. Therefore the second approach was used. The second plot showed the reconstruction error with respect to different numbers of components. "knees" were visible at 10 and 13. Since visual inspection accounted for the fact that the decline of the reconstruction error after 13 declined more than after 10, the number of components chosen is 13.

The results of the Eigendocs algorithm are displayed in Figure 4.1.

documents saved as images in .png format, bad quality to minimize the size of the database when querying db, top image results looked similar, which is how the idea of this section arose

Figure 4.1: The first 10 preprocessed documents of the dataset. The original images are displayed in the first row. The second row shows the reconstructed images using the compressed images from the fourth row. The third row shows the reconstruction error, i.e. the difference between the reconstructed and the original image. The last row presents the greyscale values of the compressed 13-dimensional image as a line.



(a) The clusters identified by OPTICS of the documents preprocessed according to item 1.

(b) The clusters identified by OPTICS of the documents preprocessed according to item 2.

Figure 4.2: The clusters were extracted from the respective reachability plots in Figure 3.10. The blue points are noise points, whereas any other colour denotes a cluster.

4.3 Evaluation of OPTICS

The algorithm OPTICS was applied to data, which was preprocessed according to item 1 and item 2. The clusters from Figure 4.2 were extracted from the respective reachability plots in Figure 3.10. The three-dimensional plots visualize the first three dimensions of the data and thus, the weights of the first three principal components assigned by the Eigendocs algorithm. By visual inspection and comparison of both plots, it can be seen that the projection by the combination of resizing and PCA of item 1 scatters the objects further along the x_2 axis. Hence, the distance between the objects is larger and more clusters are identified. One could argue that the narrow distribution of the objects in the Eigendocs plot is due to the fact, that the input data encodes not only the visual appearance in terms of page layout but also the size of the document. Possibly, this could explain why the objects are less scattered along this dimension.

To analyse the results of the clustering, the content of the clusters was examined. Since the documents are not labelled, the content of the clusters was analysed by visual inspection. The content of the clusters is displayed in Figure 4.3 and Figure 4.4. The yellow images belong to the group identified as noise. The images preprocessed according to item 1 were partitioned into multiple small and one big cluster. The Eigendocs images' clusters have similar sizes. The row of noise images is thus, way longer than the other rows in Figure 4.4. Most of the certificates are classified as noise for both approaches.

Figure 4.3: The yellow images belong to the group denoted noise. Most certificates are classified as noise. There is one big cluster and multiple small clusters. The images were preprocessed as discussed in item 1 to 32x32 greyscale pixels.

Figure 4.4: Most certificates are classified as noise. The rest of the clusters have similar sizes. The images were preprocessed as discussed in item 2 to 13-dimensional greyscale pixels.

The preprocessing approach used to create the OPTICS input for the Elasticsearch database index is Eigendocs since it also encodes information about the document size.

According to Deng et al., OPTICS was developed to improve DBSCAN flaws. Therefore, DBSCAN is chosen for the cluster method in Listing 3.8, since the literature consulted works with DBSCAN as a basis. In order to reduce calculation complexity the maximum ε is 10. The distance between to two points to still be considered neighbours is defined after visual inspection of the reachability plot. Considering the intrinsic structure of the data it is set to 0.5 to return meaningful clusters.

4.4 Evaluation of database

According to [12], Structured Query Language (SQL) databases are a good choice for efficiently storing structured data. This is because their paradigm ACID, i.e. Atomicity, Consistency, Isolation, Durability, provides high reliability. Not only SQL (NoSQL) databases, on the other hand, are more flexible and can be used to store unstructured data [12]. They do not require a predefined schema and can therefore accept documents of arbitrary structure [11].

Usually, NoSQL databases do not offer services such as `JOIN` [11]. According to Gaspar and Stouffer, NoSQL databases make a tradeoff between storage and speed, as well as a tradeoff between consistency and availability. NoSQL databases are said to outperform out-of-the-box SQL databases [11].

Since the similarity between vectors is usually calculated using some form of cosine similarity, rather than Euclidean distance in literature, cosine is preferred over Euclidean distance. Since the models may produce embeddings, which are not normalized, the cosine similarity is used instead of the dot product. Soft cosine similarity is not used, since it is not available in Elasticsearch. soft cosine would be better!

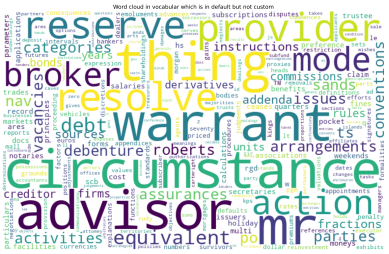
A document store database can be used if the primary goal is to write fast rather than write save [11].

4.5 Evaluation of TF-IDF

The main obstacle to overcome was the high dimensionality of the TF-IDF embeddings. Hence, the goal of the parameter selection was to find a way to reduce the dimensionality of the vocabulary to the maximum vector dimensionality of Elasticsearch. However, the quality of the embeddings should not decline too much.

The choice of the preprocessor was investigated with regard to the goal of minimizing the vocabulary size. Both the default and custom preprocessor were tested on a data corpus of 195 documents with regard to the vocabulary (size) and the result of preprocessing. A visualization obtained from the comparison is given in Figure 4.5. While the default preprocessor had a vocabulary size of 1641, the custom preprocessor had a size of 1906. The custom preprocessor was chosen because it had a smaller vocabulary size. The differences between both vocabularies were compared and visualized. The custom vocabulary has some bigrams, which are not present in the vocabulary produced by the default preprocessor. Initially, the idea was to have a vocabulary that consists only of unigrams.

Initially, there should have been two different TF-IDF models. The first one should have been used to obtain documents which are similar to the query document. Therefore, terms which occur only once in the corpus should have been removed from the vocabulary. The second approach should have been used to obtain specific documents from the corpus. Hence, the vocabulary should consist of very document-specific terms and thus, `max_df` would have been relatively low, to omit terms that occur in many documents. However, the restrictions imposed by the database implementation made it impossible to explore many parameter ranges. Therefore, only one TF-IDF model was used in the end, whose parameters `min_df` and `max_df` were set to values which kept the vocabulary and thus, the dimensionality of the embeddings reasonably small.



(a) The terms only present in the vocabulary produced by the default preprocessor.

(b) The terms only present in the vocabulary obtained from the custom preprocessor.

Figure 4.5: The WordClouds visualize which words are not shared by both vocabularies.

4.6 Evaluation of Doc2Vec

Since no labelled data is available, the evaluation of the Doc2Vec embeddings is limited. Therefore, the Doc2Vec embeddings are evaluated by comparing them to other embeddings. The Doc2Vec model is not tuned in terms of hyperparameter selection, but the default settings are used since there is no way to evaluate the resulting embeddings.

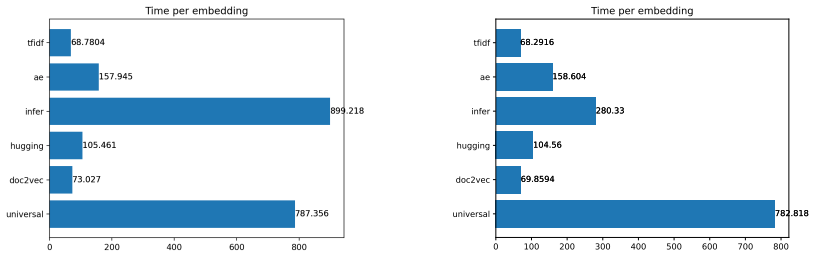
4.7 InferSent

The `max` pooling type is used for the InferSent model, since the Conneau et al. found from conducting experiments using different pooling techniques that it was the best option.

Initially, in this work, the Global Vectors for Word Representation (GloVe) word embeddings were used for the InferSent model. However, since the file of precomputed GloVe word embeddings has a size of 5.65 GB and thus, slows down the model, ultimately another word embedding was used. The time necessary to initialize the database, compute and insert 195 documents for specific embeddings is displayed in Figure 4.6. The custom word embedding used in this work is a Word2Vec model trained on a selection of 195 documents from the Bahamas dataset.

4.8 analysis/ comparison of models

difference query responses for different models? any images which produce unusual results?



(a) Precomputed embeddings of GloVe.

(b) Custom Word2Vec embeddings.

Figure 4.6: Time (seconds) necessary to initialize the database, compute and insert 195 documents for specific embeddings.

4.9 Evaluation of the performance

4.9.1 Fahnder clustern

4.9.2 Fahnder bewerten Resultate (image matrix)

4.10 Evaluation of the usability

4.10.1 Metrics

5 Results

Evaluate the results from the previous chapter.

5.1 Fulfilment of objective

5.2 Research results

5.2.1 RQ1: Question 1?

6 Conclusion

7 Outlook

7.1 Future Work

Bibliography

[1] K.P. Agrawal, Sanjay Garg, Shashikant Sharma, and Pinkal Patel. Development and validation of optics based spatio-temporal clustering technique. *Information Sciences*, 369:388–401, 2016. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2016.06.048. URL `https://www.sciencedirect.com/science/article/pii/S0020025516304765`.

[2] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: Ordering points to identify the clustering structure. *SIGMOD Rec.*, 28(2):49–60, jun 1999. ISSN 0163-5808. doi: 10.1145/304181.304187. URL `https://doi.org/10.1145/304181.304187`.

[3] Fankar Armash Aslam, Hawa Nabeel Mohammed, Jummal Musab Mohd. Munir, and Murade Aaraf Gulamgaus. Efficient way of web development using python and flask. *International Journal of Advanced Research in Computer Science*, 6:54–57, 2015.

[4] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, Sebastopol, CA, USA, 1st edition, 2009.

[5] Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Universal sentence encoder, 2018.

[6] Delphine Charlet and Géraldine Damnati. Simbow at semeval-2017 task 3: Soft-cosine semantic similarity between questions for community question answering, 2017.

[7] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data, 2018.

[8] Matt Copperwaite and Charles Leifer. *Learning Flask Framework*. Packt Publishing, 2015.

[9] Laura Dayton, Dante Rousseve, Neil Sehgal, and Sindura Sriram. Csci 1430 final project report: Methods of facial recognition, 2020.

[10] Z. Deng, Y. Hu, M. Zhu, and et al. A scalable and fast optics for clustering trajectory big data. 18:549–562, 2014. doi: 10.1145/304181.304187. URL https://doi.org/10.1007/s10586-014-0413-9.

[11] Daniel Gaspar and Jack Stouffer. *Mastering Flask Web Development: Build Enterprise-Grade, Scalable Python Web Applications,*, volume 2. Packt Publishing, 2018.

[12] Miguel Grinberg. *Flask Web Development: Developing Web Applications with Python.* O'Reilly Media, Inc., 2018.

[13] Klara M. Gutekunst. Identifying fiscal fraud with anomaly detection techniques. Technical report, University of Kassel, 2023. URL https://github.com/KlaraGtknst/identifying-fiscal-fraud.

[14] Hari Krishna Kanagala and V.V. Jaya Rama Krishnaiah. A comparative study of k-means, dbscan and optics. In *2016 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–6, 2016. doi: 10.1109/ICCCI.2016.7479923.

[15] Gunjan Khosla, Navin Rajpal, and Jasvinder Singh. Evaluation of euclidean and manhanttan metrics in content based image retrieval system. In *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 12–18, 2015.

[16] Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. From word embeddings to document distances, 2014.

[17] Tzu-Hsuan Lin and Jehn-Ruey Jiang. Credit card fraud detection with autoencoder and probabilistic random forest. *Mathematics*, 9(21), 2021. doi: 10.3390/math9212683. URL https://www.mdpi.com/2227-7390/9/21/2683.

[18] Yu. A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs, 2018.

[19] Tomas Mikolov and Quoc Le. Distributed representations of sentences and documents, 2014.

[20] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[21] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality, 2013.

[22] Sumit Misra, Soumyadeep Thakur, Manosij Ghosh, and Sanjoy Kumar Saha. An autoencoder based model for detecting fraudulent credit card transaction. *Procedia Computer Science*, 167:254–262, 2020. doi: https://doi.org/10.1016/j.procs.2020.03.219. International Conference on Computational Intelligence and Data Science.

[23] M. Mitra and B. B. Chaudhuri. Information retrieval from documents: A survey, 1999.

[24] Mauritius Much, Frederik Obermaier, Bastian Obermayer, and Vanessa Wormer. So funktioniert das system bahamas. URL `https://www.sueddeutsche.de/wirtschaft/bahamas-leaks-so-funktioniert-das-system-bahamas-1.3172913`. [Accessed 08.08.2023].

[25] Mohammad Robihul Mufid, Arif Basofi, M. Udin Harun Al Rasyid, Indhi Farhandika Rochimansyah, and Abdul rokhim. Design an mvc model using python for flask framework development. In *2019 International Electronics Symposium (IES)*, pages 214–219, 2019. doi: 10.1109/ELECSYM.2019.8901656.

[26] Mostofa Ali Patwary, Diana Palsetia, Ankit Agrawal, Wei-keng Liao, Fredrik Manne, and Alok Choudhary. Scalable parallel optics data clustering using graph algorithmic techniques. In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, SC '13, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450323789. doi: 10.1145/2503210.2503255. URL `https://doi.org/10.1145/2503210.2503255`.

[27] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation, 2014.

[28] Robert-George Radu, Iulia-Maria Rădulescu, Ciprian-Octavian Truică, Elena-Simona Apostol, and Mariana Mocanu. Clustering documents using the document to vector model for dimensionality reduction, 2020.

[29] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[30] Shyam Seshadri. *Angular Up and Running: Learning Angular, Step by Step*. O'Reilly Media, Inc., 2018.

[31] Grigori Sidorov, Alexander Gelbukh, Helena Gómez-Adorno, and David Pinto. Soft similarity and soft cosine measure: Similarity of features in vector space model, 2014.

[32] Gerd Stumme, Robert Jäschke, and Christoph Scholz. Internet-suchmaschinen, 2011.

[33] Dodi Sudiana, Mia Rizkinia, and Fahri Alamsyah. Performance evaluation of machine learning classifiers for face recognition. In *2021 17th International Conference on Quality in Research (QIR): International Symposium on Electrical and Computer Engineering*, pages 71–75, 2021. doi: 10.1109/QIR54354.2021.9716171.

[34] Yichuan Tang and Xuan Choo. Intrinsic divergence for facial recognition, 2008.

[35] TODO. Fuzziness, . URL `https://www.elastic.co/guide/en/elasticsearch/reference/current/common-options.html#fuzziness`. [Accessed 15.09.2023].

[36] TODO. Get api, . URL `https://www.elastic.co/guide/en/elasticsearch/reference/current/docs-get.html`. [Accessed 15.09.2023].

[37] TODO. Dense vector field type, . URL `https://www.elastic.co/guide/en/elasticsearch/reference/current/dense-vector.html#dense-vector-similarity`. [Accessed 15.09.2023].

[38] TODO. k-nearest neighbor search, . URL `https://www.elastic.co/guide/en/elasticsearch/reference/current/knn-search.html`. [Accessed 15.09.2023].

[39] TODO. How to deploy a text embedding model and use it for semantic search, . URL `https://www.elastic.co/guide/en/machine-learning/8.10/ml-nlp-text-emb-vector-search-example.html`. [Accessed 15.09.2023].

[40] TODO. Match query, . URL `https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-match-query.html`. [Accessed 15.09.2023].

[41] TODO. Synonyms, . URL `https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-match-query.html#query-dsl-match-query-synonyms`. [Accessed 15.09.2023].

[42] TODO. Text analysis overview, . URL `https://www.elastic.co/guide/en/elasticsearch/reference/current/analysis-overview.html`. [Accessed 15.09.2023].

[43] TODO. Glove: Global vectors for word representation, . URL `https://nlp.stanford.edu/projects/glove/`. [Accessed 04.10.2023].

[44] TODO. Doc2vec paragraph embeddings, . URL `https://radimrehurek.com/gensim/models/doc2vec.html`. [Accessed 01.10.2023].

[45] TODO. gensim.models.doc2vec, . URL https://tedboy.github.io/nlps/generated/generated/gensim.models.Doc2Vec.__init__.html. [Accessed 01.10.2023].

[46] TODO. Word2vec embeddings, . URL https://radimrehurek.com/gensim/models/word2vec.html#gensim.models.word2vec.Word2Vec. [Accessed 01.10.2023].

[47] TODO. Quick start user guide, . URL https://slurm.schedmd.com/quickstart.html. [Accessed 16.09.2023].

[48] TODO. Tf–idf term weighting, . URL https://scikit-learn.org/stable/modules/feature_extraction.html#text-feature-extraction. [Accessed 29.09.2023].

[49] TODO. Tf–idf term weighting, . URL https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html. [Accessed 29.09.2023].

[50] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 586–591, 1991.

[51] A. Voit, A. Stankus, S. Magomedov, and I. Ivanova. Big data processing for full-text search and visualization with elasticsearch, 2017.

[52] Chang Wang and Sridhar Mahadevan. Multiscale manifold learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 27:912–918, Jun. 2013. doi: 10.1609/aaai.v27i1.8633. URL https://ojs.aaai.org/index.php/AAAI/article/view/8633.

[53] Andy B. Yoo, Morris A. Jette, and Mark Grondona. Slurm: Simple linux utility for resource management. In Dror Feitelson, Larry Rudolph, and Uwe Schwiegelshohn, editors, *Job Scheduling Strategies for Parallel Processing*, pages 44–60, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg. ISBN 978-3-540-39727-4.

[54] V. Zamfir, M. Carabas, C. Carabas, and N. Tapus. Systems monitoring and big data analysis using the elasticsearch system, 2019.

[55] Jun Zhang, Yong Yan, and M. Lades. Face recognition: eigenface, elastic matching, and neural nets. *Proceedings of the IEEE*, 85(9):1423–1435, 1997. doi: 10.1109/5.628712.

[56] Wen Zhang, Taketoshi Yoshida, and Xijin Tang. Tfidf, lsi and multi-word in information retrieval and text categorization, 2008.

List of Figures

List of Tables

Listing-Verzeichnis

A Anhang

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und nur mit den nach der Prüfungsordnung der Universität Kassel zulässigen Hilfsmitteln angefertigt habe. Die verwendete Literatur ist im Literaturverzeichnis angegeben. Wörtlich oder sinngemäß übernommene Inhalte habe ich als solche kenntlich gemacht.

Kassel, October 4, 2023

Klara Maximiliane Gutekunst