Ludwigshafen University
of Business and Society
University of Applied Sciences

Enhancing insights into spending through aggregation with automated document clustering of a large-scale multilingual corpus

Bachelor Thesis

Part of the Examination for the
Bachelor of Science (B.Sc.)
of
International Business Administration and Information Technology
at the University of Business and Society Ludwigshafen

by

Lisa Rebecca Mirjam Schmidt
Sternstraße 93
67063 Ludwigshafen am Rhein

Date of submission:	01. Februar 2018
Company Supervisor:	Dr. Karthik Muthuswamy
Academic Supervisor:	Prof. Dr. Joachim Melcher

Contents

Abkürzungsverzeichnis

AI	Artificial Intelligence
AI BUS	SAP AI Business Services
NLP	Natural Language Processing
CRISP-DM	Cross Industry Standard Process for Data Mining
CoE	Center of Excellence
BoW	Bag of Words
TF-IDF	ferm frequency - inverse document frequency

1 Introduction

1.1 Motivation

1.1.1 Current situation

An essential part of economic counselling is the assessment of allocated spending for different segments of a company. Spending of a firm usually is written down in invoice documents, which have to be grouped for segments to analyze their costs. While the global market is estimated to comprise 550 billion invoices annually, 90% are exchanged paper-based [7]. With modern technology, these paper- or document based (PDF, docx, odt) files can be transformed into a structured or semi-structured format. According to expert estimates, unstructured data makes up for more than 80% of enterprise data [2]. This data is not leverageable with traditional data analysis tools, but its value must be harvested for companies to utilize their full potential. A large share of unstructured data found in companies is textual data. [8]

1.1.2 Importance

1.2 Research Question

1.3 Outline

2 Objectives and Criteria

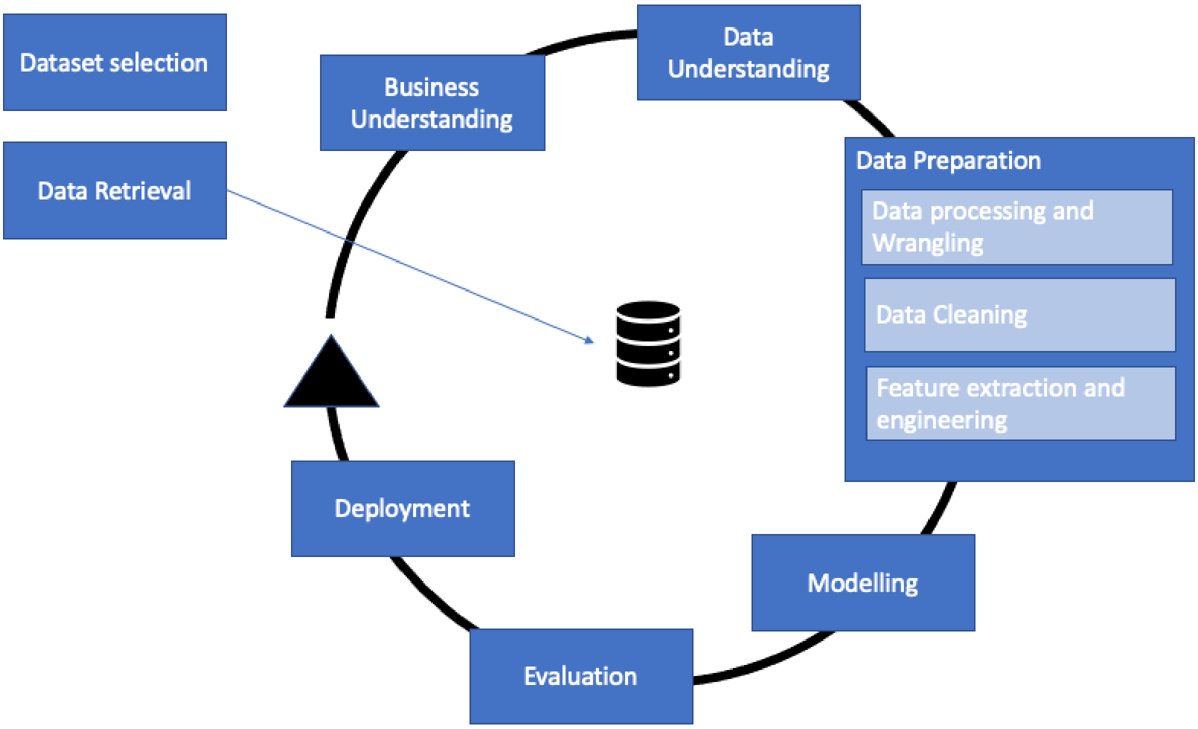
2.1 Detailed Task Description

The goal of the thesis is to add value to real business documents by aggregating expenses into clusters of similar expenses. The supplied document dataset consists of 150.000 invoices. The invoices contain different information, for example the vendor, billing amount or a description of the goods. Valuable information for companies would be insight into the different categories of expenses and the corresponding cost. With traditional data analysis methods, the company's controlling departments cannot identify which expenses are similar in nature (for example logistics costs). The task is to perform a full data analysis on the supplied dataset. The dataset is to be prepared for processing with established methods. An evaluation for different means of feature extraction, machine learning, model evaluation and visualization should be performed. With the evaluation a complete flow for the data processing should be presented. The result should be an added value to the dataset in the form of aggregated expenses.

2.2 Criteria set by SAP SE

2.3 Research Model

To solve the task described in chapter 1.2, this paper employs the Cross Industry Standard Process for Data Mining (CRISP-DM). This process model [2] puts forward a structure for conducting data mining projects. CRISP-DM was developed in 1996 by thee companies, which are now the partners of the CRISP-DM consortium: NCR, DaimlerChrysler AG and SPSS Inc.


Dataset selection
Data Retrieval
Business Understanding
Data Understanding
Data Preparation
Data processing and Wrangling
Data Cleaning
Feature extraction and engineering
Modelling
Evaluation
Deployment

3 Fundamentals

3.1 Glossary of Terms

- SAP SE
- SAP Leonardo MLF
- SAP AI Core
- SAP AI Launchpad
- AI Foundation

3.2 Corporate Environment

3.2.1 Historical

SAP was founded in 1972 by five ex-IBM employees. The original company name was "Systemanalyse Programmentwicklung", which can be translated to "System analysis and program development". In 1976, a second company, the SAP GmbH was founded, where the acronym SAP denoted "Systems, Applications and Products for data processing" [4]. The SAP GmbH is the company, which is today known as SAP SE.

Data processing being part of the company's name shows the importance of this field to SAP since the beginning of the company history. In 2017 SAP entered the AI business with the SAP Leonardo Machine Learning Foundation [3], and challenged the market for hyped products in machine learning, blockchain, big data and design thinking. The name Leonardo refers to Leonardo Da Vinci, who is renowned for his interdisciplinary innovations [1]. SAP has the goal of driving the digital innovation strategies of customers with the help of SAP Leonardo.

With changes in the underlying hyperscalers for Leonardo, and evolving requirements of customers and partners, SAP adjusted their AI strategy. Two new products were introduced: SAP AI Core and SAP AI Launchpad. Both products are united under the collective name of AI Foundation [3]. With the general availability of AI Foundation in late 2021, SAP Leonardo is officially sunsetted.

- What comes in the future?

- how much does SAP earn with AI?

3.2.2 Organizational

SAP SE has an executive board consisting of seven members, each attributed to one area. The Artificial Intelligence (AI) division falls into the responsibility of Jürgen Müller, Chief Technology Officer and leader of the board area for technology and innovation [6]. Members of the organizational unit for AI are divided in different teams concerned with development, product success, operations and specific AI-services. Development Teams are organized into Centers of Excellence (CoEs), in which special expertise for designated areas is united. SAP has a team of researchers around the world concerned with state-of-the-art topics, including few-shot learning, sentiment analysis, privacy and fairness [**AIOverviewResearch**]. The research teams regularly publish articles on their advancements.

3.2.3 Technological

https://www.sap.com/products/artificial-intelligence.html

3.3 Machine Learning

Already Alan Turing understood that for laymen a learning machine can be perceived as a paradox. How can a machine learn, if a human has to define its behavior beforehand? There are three major subfields in the discipline of artificial intelligence that fundamentally explain how a computer can learn how to behave despite predefined behavior.

3.3.1 Supervised Learning

A supervised learning algorithm learns its decision with the help of a data set (input) that also contains the correct decision (output) as information. It is trained with only a part of the entire data set, so that the model can be tested in a later step with the help of unknown data. This way, a statement can be made about the accuracy of the model.

3.3.2 Unupervised Learning

Unsupervised learning is complementary to supervised learning. All algorithms that fall into the category of unsupervised learning are trained with data that does not contain the correct output (label) as information. Here, the categorization is not constrained by the given data, but decided on by the algorithm.

3.3.3 Reinforcement Learning

The third way in which an algorithm can make better decisions as it gains experience is called reinforcement learning. Reinforcement learning is about letting algorithms solve very complex tasks. The special feature is that there is no defined solution path, but the algorithm is rewarded for goal-oriented behavior and punished for wrong decisions. The definition of goal-oriented behavior has to be put into place by the engineers setting up the training of the model. Real-world tasks are extremely complex, so not all possible solution paths can be calculated and compared to find the optimal path. Parking a car is a routine task for a human after a few hours of driving, but a computer sees only an infinite set of possibilities for turning angles. This problem can be solved by reinforcement learning. The algorithm is rewarded for each parking attempt where the car ends up seeing in the parking space. For the remaining attempts, the algorithm is penalized. Over many thousands of attempts, the reinforcement learning model is trained in this way.

The three major ways of learning even with previously defined behavior can now be implemented by specific models. For example, there are several ways to create and train a model using Unsupervised Learning.

3.3.4 Clustering Algorithms

multinomial, one bad example for a clustering would be the closest 5 docs to each one (this is multilabel)

3.3.5 Natural Language Processing

Natural Language Processing (NLP) is often attributed to the computer science, but after closer examination, NLP is a discipline comprised of linguistics, computer science, artificial intelligence an mathematics [5].

3.4 SAP AI Core

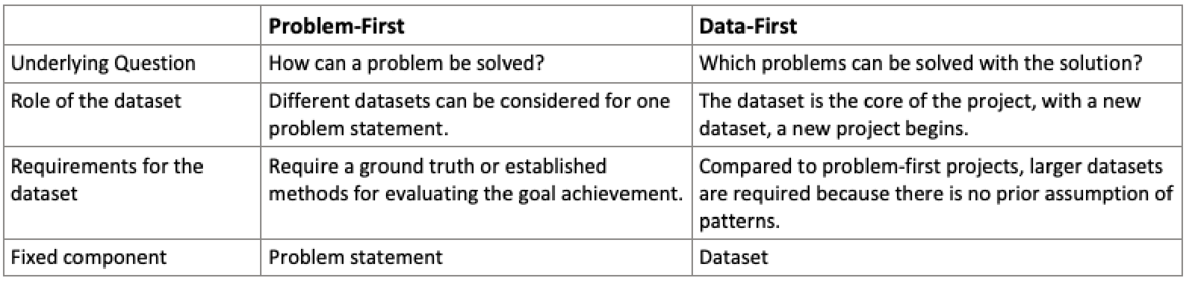
3.4.1 Docker

4 Discussion of Alternatives

For each step in the CRISP-DM process, a multitude of different approaches can be employed. In this chapter, the alternatives are presented and evaluated on the basis of established criteria.

4.1 Dataset selection

The dataset is the fundament of a data-science project. The quality, size and closeness to reality decide the degree to which the findings can be helpful for solving real problems. In this section, a classification for data-science projects is introduced.

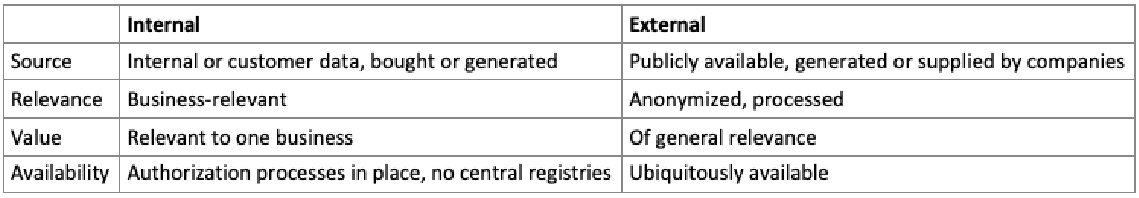


	Problem-First	**Data-First**
Underlying Question	How can a problem be solved?	Which problems can be solved with the solution?
Role of the dataset	Different datasets can be considered for one problem statement.	The dataset is the core of the project, with a new dataset, a new project begins.
Requirements for the dataset	Require a ground truth or established methods for evaluating the goal achievement.	Compared to problem-first projects, larger datasets are required because there is no prior assumption of patterns.
Fixed component	Problem statement	Dataset

The problem-first project type is characterized by a predefined problem statement or research goal. The underlying question is, how a specific or a set of problems can be solved. For the selection of the dataset, this requires that goal achievement can be measured with existing ground truth. In some cases, also other methods such as expert judgements can be sufficient. In a problem-first project, different datasets can and should be considered.

The complementary project type is the data-first project, which describes most data-mining projects. This type is characterized by a more explorative approach, and the goal to find problems and patterns. Here, the dataset is at the core of the project and the fixed aspect of a project. In turn, this means swapping the dataset is the start of a new project.

With the project type as a decisive factor for the dataset explained, the other evaluation criteria for the dataset is described. In the corporate environment two fundamental sources for data exist.



	Internal	External
Source	Internal or customer data, bought or generated	Publicly available, generated or supplied by companies
Relevance	Business-relevant	Anonymized, processed
Value	Relevant to one business	Of general relevance
Availability	Authorization processes in place, no central registries	Ubiquitously available

Firstly, data can be sourced from inside the company. This can include customer data or data generated from observation and monitoring processes inside the company. Data is either directly or very closely related to the company's business. Depending on the solution, it can be of use to customers or it can be utilized inside the company. Internally sourced data is almost exclusively rated confidential, limiting even intra-company access to it. Authorization processes and more than often not existing registries for data may hinder project progress.

Secondly, data can be sourced outside the company. A vast number of online registries for data exist, both with paid and free of charge service offerings. Data sources include real-life data and data generated for educational purposes. Because of its publication, the data is stripped from all parts which could expose confidential information such as corporate secrets. Additionally, data is anonymized and processed to limit the usefulness to potential competitors.

It can be stated, that both sources are suited for different goals.

4.2 Data Retrieval

- how is the dataset stored?

4.3 Business Understanding

- why did i decide for clustering?

4.4 Data Understanding

- different data exploration tools can be used for the data understanding.

4.5 Data preparation

- transforming json documents into dataframe rows
- removing stopwords from several languages
- removing numbers and interpunction
- tokenization

4.5.1 Data processing and data wrangling

4.5.2 Data cleaning

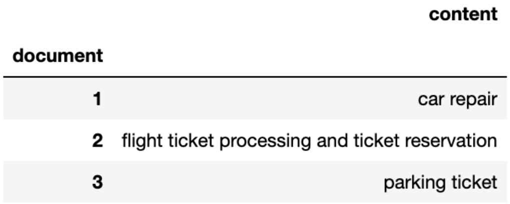
4.5.3 Feature Extraction and Feature Engineering

The majority of pupular ML algorithms require the input of scalar, vector or matrix data. A form of ML models, which work with textual input will be discussed later in this section. But since many algorithms were not designed to work with textual data, a transformation is required before already existing algorithms can be applied. Several methods for representing text as mathematical object will be discussed in the following.

One-hot encoding

Bag of Words Model

Following three documents will be considered to explain the workings of the Bag of Words (BoW) model:

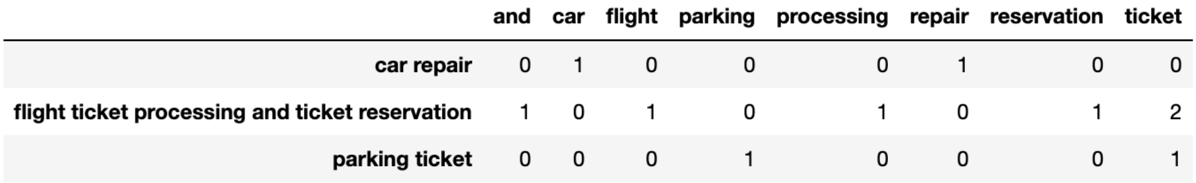


	content
document	
1	car repair
2	flight ticket processing and ticket reservation
3	parking ticket

A corpus is transformed into the BoW representation in two steps.

Firstly, the vocabulary is determined. The vocabulary is a collection of all words occuring in the corpus. Every word is contained exactly once, regardless of the actual number of occurrences. The vector represenation of one document is of the same length as the vocabulary. One document being represented by one vector, a corpus of several documents can be represented as a matrix. The resulting matrix has the size $|D| * |V|$, with $|D|$ being the number of documents in the corpus, and $|V|$ being the size of the vocabulary.

Secondly, for each combination of one document d_i and one word v_j in the vocabulary, the occurrences are counted. The times, how often the specific word is contained in the document is noted in the matrix at locationij.



	and	car	flight	parking	processing	repair	reservation	ticket
car repair	0	1	0	0	0	1	0	0
flight ticket processing and ticket reservation	1	0	1	0	1	0	1	2
parking ticket	0	0	0	1	0	0	0	1

The result of vectorization are three vectors:

$$d_1 = [0, 1, 0, 0, 0, 1, 0, 0]$$

$$d_2 = [1, 0, 1, 0, 1, 0, 1, 2]$$

$$d_3 = [0, 0, 0, 1, 0, 0, 0, 1]$$

This vectorization method can be implemented with ease and in a computationally efficient manner. The BoW model allows for a very intuitive understanding of documents, since texts consisting of the same words are considered topically related.

One of the drawbacks of this method is that no consideration is paid to words being repeated in one document. Additionally, no semantic relationship between words or documents can be inferred. Further, it can be stated, that the BoW representation fails to capture the meaning of synonyms. This becomes obvious with an example: document d_1 and d_3 would be considered to belong into the topic of transportation or automotives. The BoW representation suggests topical proximity between document d_2 and d_3 , through the shared word "ticket". "Ticket" here is used in both the meaning of an entrace pass (d_2) and in the meaning of a note for a traffic offence (d_3).

To conclude, the BoW model is a straightforward text representation method. Still, it fails to capture several aspects of the natural language.

Tf-Idf

The ferm frequency - inverse document frequency (TF-IDF) model aims to capture more meaning from the corpus by considering the composition of the whole corpus for the calculation of individual document vectors.

TF-IDF makes two assumptions about natural language:

1 Term Frequency A word t_i which occurs very frequently in one document is considered to describe a text very well. The occurences of one word in one document is denoted as $\#(t_i)$. One additional consideration needs to be made regarding the document length. In a document of length $|d_2| = 6$ and a document of length $|d_3| = 2$, the word $t_i$occuring once would be considered equally important to each document. Of course, the word should be considered more important to d_3, since it accounts for a larger share of the text. The measure resulting from both ideas is the term frequency:

$$TF(t_i, d_j) = \frac{\#(t_i)}{|d_j|} = \frac{occurences\ of\ word\ t_i\ in\ document\ \ d_j}{legth\ of\ d_j}$$

2 Inverse Document Frequency A word t_i which occurs in a large number of documents does not describe one document well. Words occuring in many documents often are articles or pronouns (stopwords) which do not provide value when inspecting the content of a text. The inverse document frequence is a measure accounting for this fact. The inverse document frequency of a word is the proportion between the number of documents in the corpus and the number of documents containing the word. The logarithm is applied, as the importance of a word does not increase proportionally to the number of occurrences.

$$TF(t_i, d_j) = \log \frac{|D|}{\#(d_{t_i})} = \log \frac{number\ of\ documents\ in\ corpus\ D}{number\ of\ documents\ containing\ word\ t_i}$$

Combining both assumptions, the TF-IDF measure is created:

$$TFIDF(t_i, d_j) \ = \ TF(t_i, d_j) * TF(t_i, d_j)$$

For the corpus displayed in the previous section the document vectors calculated with the TF-IDF measure are:

$$d_1 = \quad [0\ 0.707107\ 0\ 0\ 0\ 0.707107\ 0\ 0 \tag{4.1}$$

$$d_2 = \quad [0.39798\ 0\ 0.39798\ 0\ 0.397980\ 0.397980.605349 \tag{4.2}$$

$$d_1 = \Big[0\ 0.707107\ 0\ 0\ 0\ 0.707107\ 0\ 0\Big]\, d_2 = \ \Big[0.39798\ 0\ 0.39798\ 0\ 0.397980\ 0.397980.605349\Big]\, d_3 = \ \Big[0\ 0\ 0\ 0.7$$

The TF-IDF measure corrects some of the pitfalls of the BoW model. It certainly is less vulnerable to skewing by stopwords as words are ranked by importance to each document and the all over corpus.

Just as the BoW representation, TF-IDF suffers from high-dimensionality. The vectors contain one element for each word in the vocabulary, resulting in vectors which are inefficient to handle. Also, TF-IDF fails to represent the topical relationship between d_1 and d_3.

Word Embeddings

This lack of "understanding" of related words, and the problem of high-dimensionality is corrected with the third presented option: Word Embeddings.

4.6 Modelling

4.6.1 K-Means with Euclidean Distance

4.7 Evaluation

4.7.1 Elbow Diagram

4.7.2 average silhouette method

4.7.3 gap statistic

4.7.4 Visualization

PCA

T-SNE

4.8 Deployment

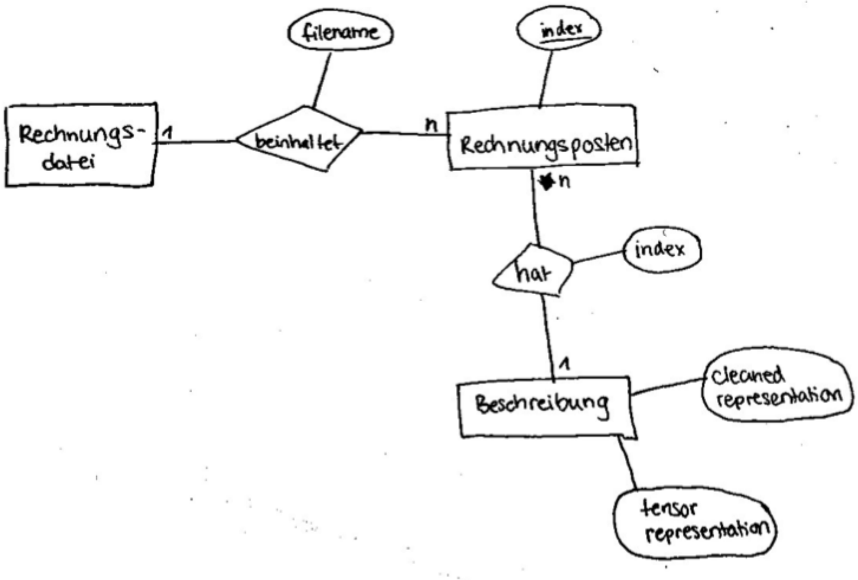
5 Theoretical Implementation

5.1 Dataset selection

5.2 Data Retrieval

5.3 Business Understanding

5.4 Data Understanding



5.5 Data preparation

5.5.1 Data processing and data wrangling

5.5.2 Data cleaning

5.5.3 Feature Extraction and Feature Engineering

5.6 Modelling

5.7 Evaluation

5.7.1 Visualization

5.8 Deployment

6 Practical Implementation

6.1 Dataset selection

6.2 Data Retrieval

6.3 Data preparation

6.3.1 Data processing and data wrangling

6.3.2 Data cleaning

6.3.3 Considerations of Space and Time Complexity

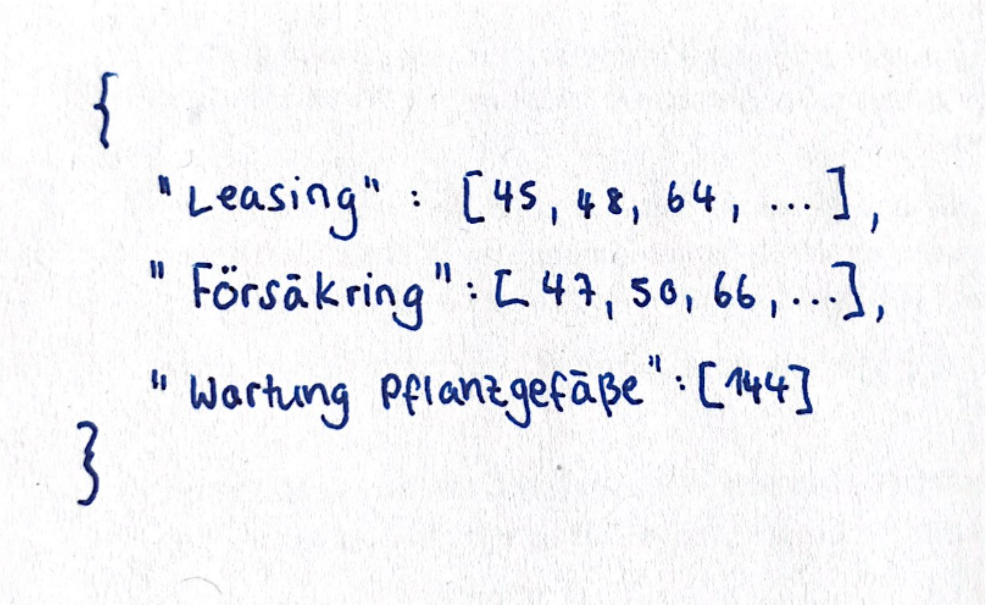
The dataset consists of over 150.000 invoices, and in those invoices, over 350.000 items are listed. With hardware-limitation in place, an optimized approach for storing and processing the data is required. Several considerations for speeding up processing time and reducing storage space can be made. In the following, the observations are explained and approaches for improvement are given.

Duplicates and space complexity

Investigation shows, there are only 79.741 unique descriptions for the listed items. By saving only the unique values, the required space is reduced to less than one fourth compared to before. Additionally, this step is required by most machine learning models, as duplicate input values can skew the outcome. The model is chosen later, so this processing step leaves the model selection more open to different kinds of learning algorithms.

Reconstructing Relationships and time complexity of searching

After the deletion of duplicate descriptions (documents in the corpus), the



```
{
   "Leasing" : [45, 48, 64, ... ],
   " Försäkring": [47, 50, 66, ...],
   " Wartung Pflanzgefäße": [144]
}
```

6.3.4 Feature Extraction and Feature Engineering

6.4 Modelling

6.5 Evaluation

6.5.1 Visualization

6.6 Deployment

7 Evaluation of the result

7.1 Visualization

7.2 Measures

8 Outlook

Literaturverzeichnis

[1] Andreas Schmitz. *Was ist SAP Leonardo?* 07/2017.

[2] Chapman, P. et al. “CRISP-DM 1.0: Step-by-step Data Mining Guide”. In: 2000.

[3] Daniel Rutschmann. *The Journey from SAP Leonardo Machine Learning Foundation to SAP AI Core and SAP AI Launchpad.* https://blogs.sap.com/2021/10/11/the-journey-from-sap-leonardo-machine-learning-foundation-to-sap-ai-core-and-sap-ai-launchpad/. 10/2021.

[4] *Geschichte Der SAP.* https://www.sap.com/germany/about/company/history/1972-1980.html.

[5] Gobinda G. Chowdhury. “Natural Language Processing”. In: *Annual Review of Information Science and Technology* 37.1 (01/2003), pp. 51–89.

[6] *Juergen Mueller Biography.* https://www.sap.com/about/company/leadership/juergen-mueller.html. Company Website.

[7] Koch, B. “The E-Invoicing Journey 2019-2025”. In: (), p. 7.

[8] TURING, A. M. “I.—COMPUTING MACHINERY AND INTELLIGENCE”. In: *Mind* LIX.236 (10/1950), pp. 433–460.