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

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 [**koch_e-invoicing_nodate**]. 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. [**turing_icomputing_1950**]

1.1.2 Importance

1.2 Research Question

1.3 Outline

2 Objectives and Criteria

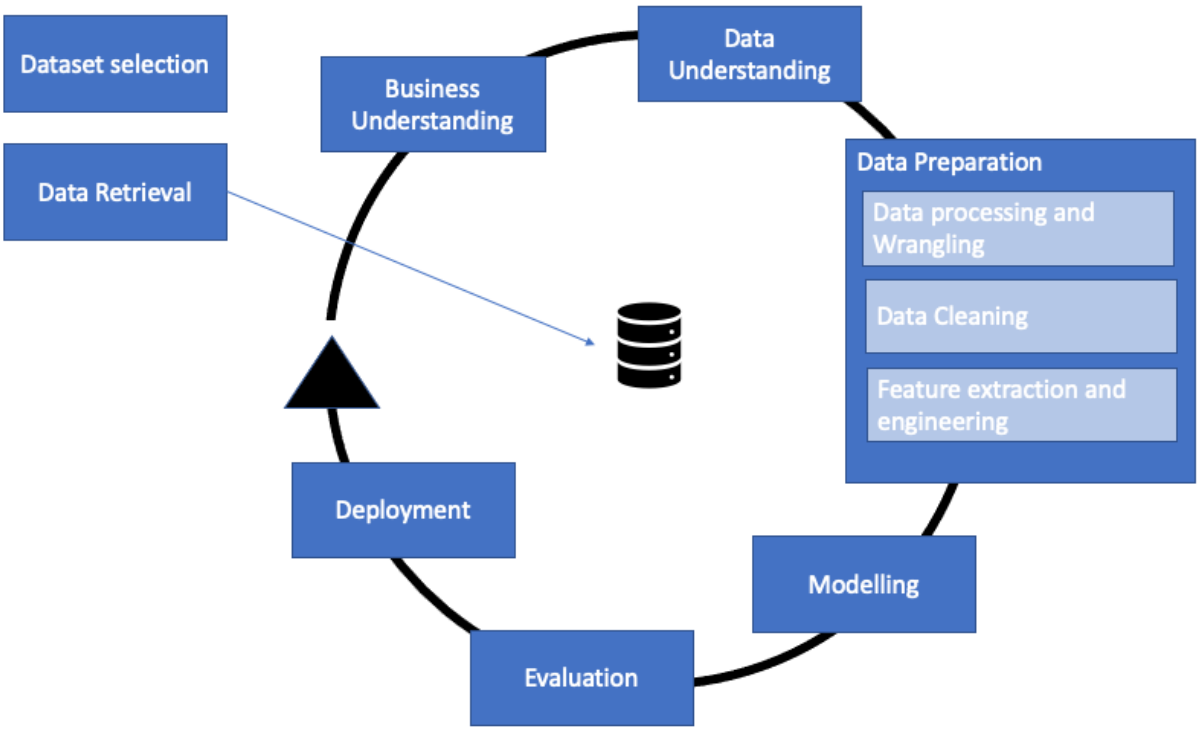
2.1 Detailed Task Description

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 [**chapman_crisp-dm_2000**] 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

3.2 Corporate Environment

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 [**gobinda_g_chowdhury_natural_2003**].

3.4 SAP AI Core

3.4.1 Docker

4 Discussion of Alternatives

4.1 Criteria

4.2 Dataset selection

4.3 Data Retrieval

4.4 Business Understanding

4.5 Data Understanding

4.6 Data preparation

- transforming json documents into dataframe rows

- removing stopwords from several languages

- removing numbers and interpunction

- tokenization

4.6.1 Data processing and data wrangling

4.6.2 Data cleaning

4.6.3 Feature Extraction and Feature Engineering

Tf-Idf

- bag of words model

Word2Vec, Embeddings

- explain tf-idf

HashVectorizer

n-Grams and Tf-Idf

4.7 Modelling

4.7.1 K-Means with Euclidean Distance

4.8 Evaluation

4.8.1 Elbow Diagram

4.8.2 Visualization

PCA

T-SNE

4.9 Deployment

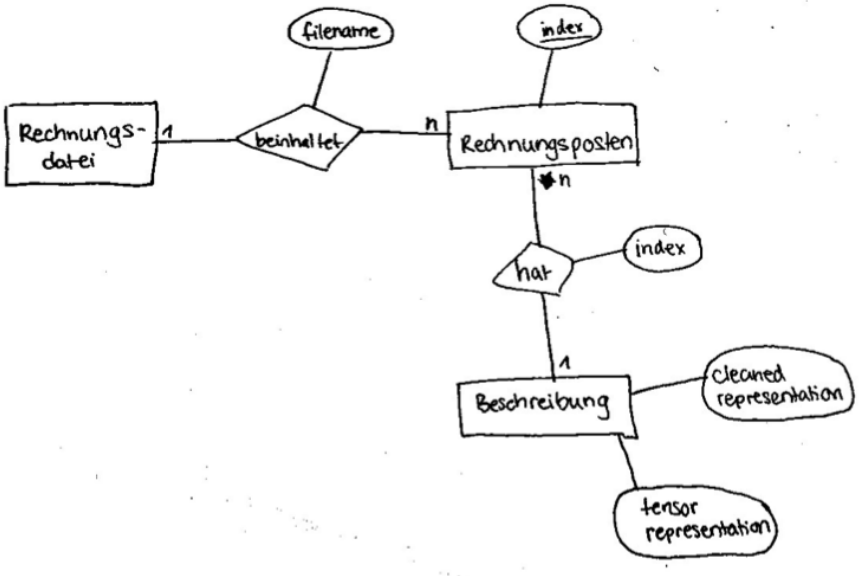
5 Theoretical Implementation

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

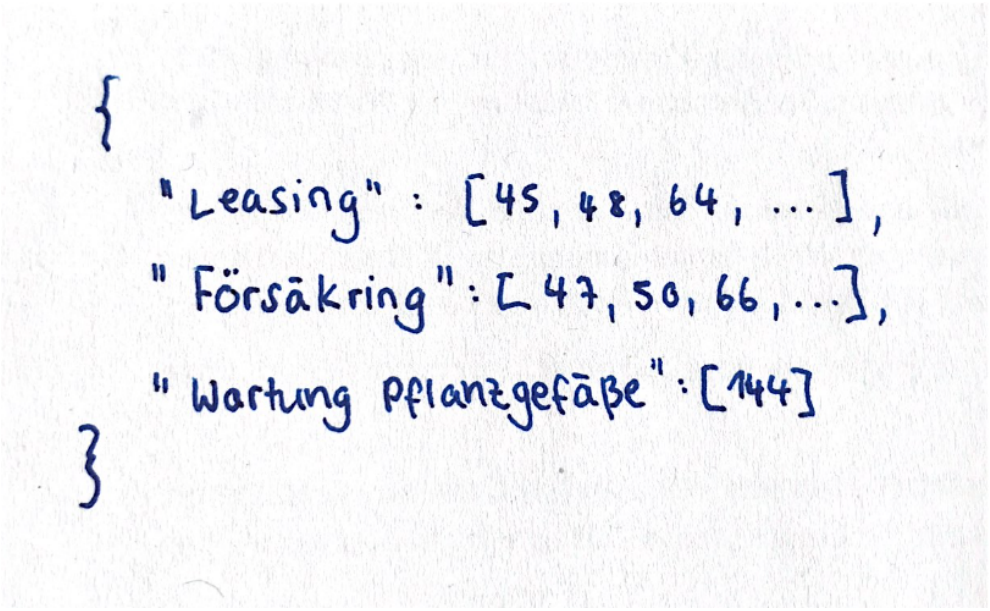
The dataset consists of over 150.000 invoices, and in those invoices, over 350.000 items are listed. With hardware-limitation in place, an optimized approach for storing and processing the data is required. Several considerations for speeding up processing time and reducing storage space can be made.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



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