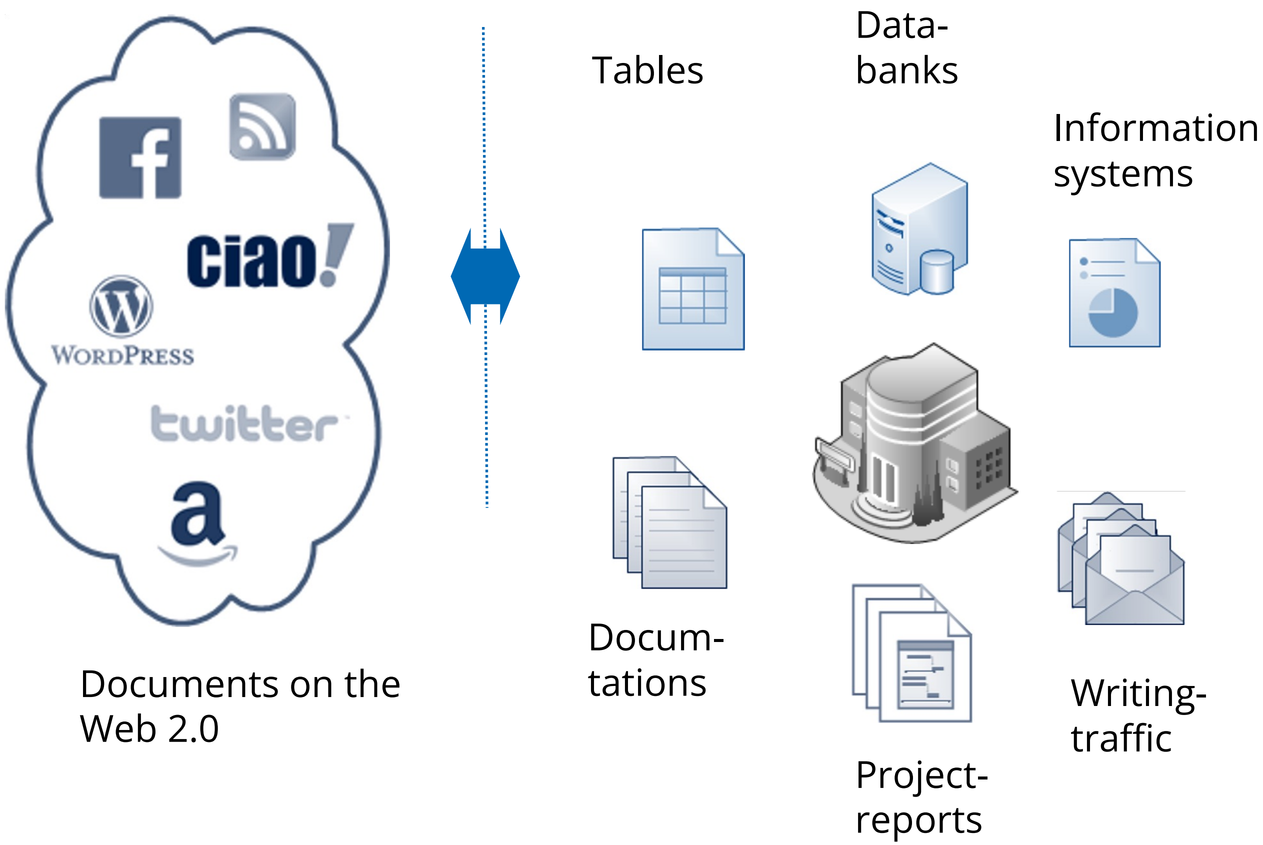
Prof. Dr. Alfred Benedikt Brendel
Chair of Business Information Systems, esp. Intelligent Systems and Services

Data Science: Advanced Analytics
Text Mining: An Introduction

Dresden // 24.05.2023
Sommersemester 2023

Text mining in the BI environment



External data

Internal data

- **Structured** data
- Scope: approx. **20% of** all data in the company

Business Intelligence

Integration potential

- **Semi-structured and unstructured** data
- Scope: approx. **80% of** all data in the company

Text Mining in the Big Data Environment

- **The amount of available texts increases extremely in Web 2.0 (keyword: user-generated content)**
- **Social networks, forums and blogs can be interesting sources for product improvements, marketing activities or opinion analysis**
- **Text processing and analysis is very computationally intensive even for a few documents**
- **Big Data technologies enable high-performance and scalable analysis systems**
- **Ex: Hadoop-based systems**
  - Distributed File Systems (Hadoop Distributed File System, HDFS).
  - Distributed execution framework MapReduce
  - Implementation e.g. in Apache Tika

http:

Relevance of text mining

...in science

Handling unstructured data

Need for evaluations based on unstructured data

Integration of unstructured data in the context of decision support

Potentials through the expansion of business understanding

Research branch Social Business Intelligence deals with the evaluation of social media

...in practice

- **Importance of** unstructured data **for companies** increases
  - Internal documents
  - External documents
- Structuring information for **knowledge management**
- Analysis of large amounts of text provides implications for the **further development of products and services**

Definition of Text Mining

Understanding of the term in the literature

Information extraction	Analysis method	Knowledge generation process

• Extraction of passages from texts
• Annotation of these passages with attributes
• Ex:
  • Recognition of companies and people
  • Assignment of functions of a person in the company
• Source: Hotho et al. (2005), p. 45ff.
	• Methods for computer-aided text analysis
	• (semi-)automatic structuring of texts
	• Ex:
	  • Text Clustering
	  • Text Categorization
	• Sources: He (2013), p. 501; Heyer et al. (2006), p. 3.
		• Analogous to the KDD process for data mining according to Fayyad (1996), p. 9.
		• Knowledge Discovery in Textual Databases (KDT)
		• Process includes, among other things, phases for the selection, preparation and analysis of text data
		• Sources: Feldman & Dagan (1995), p. 112; Hippner & Rentzmann (2006), p. 287.

Definition of Text Mining

Understanding of terms in the context of the lecture

Analysis techniques are used to evaluate and generate the text mining results.

From a **process perspective**, text mining includes both the collection and preparation of texts and the exploitation of the results.

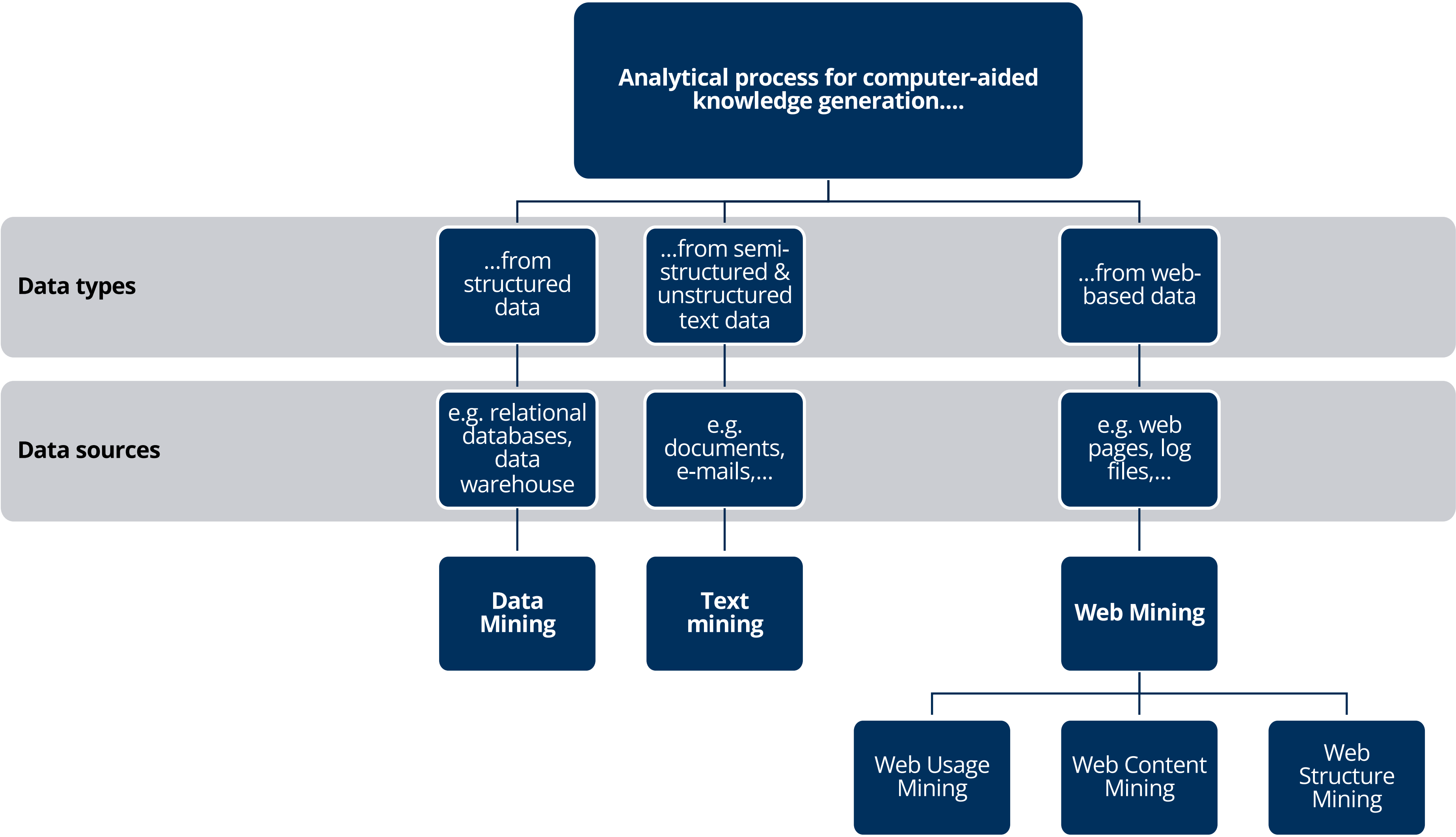
The term **text mining** refers to the analytical process for computer-assisted knowledge generation from text data.

The processing and analysis steps are integrated as functions in **application systems**.

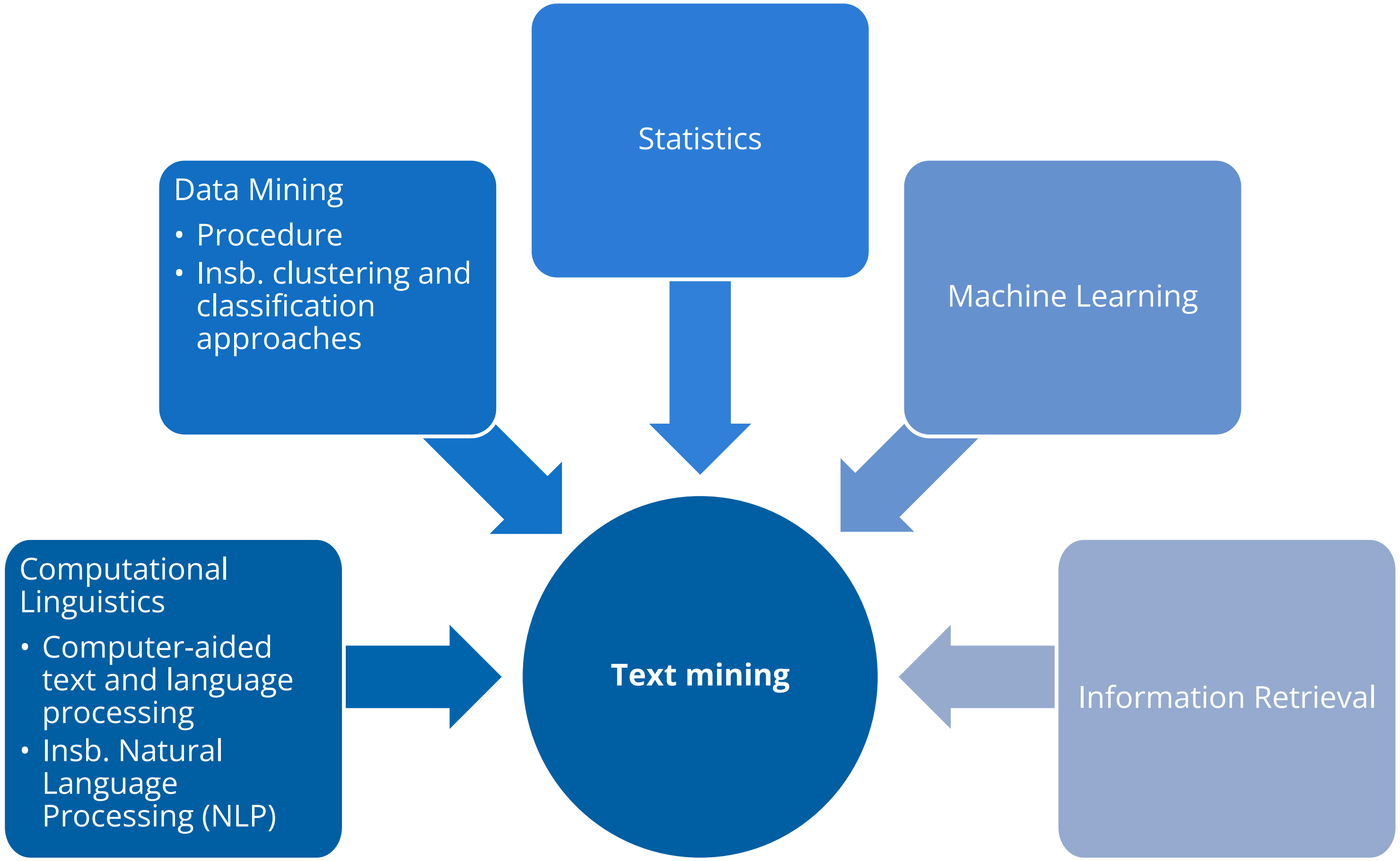
The goal of text mining is to generate new knowledge to be translated into **action**.

Text data can be both internal and external to the company; in the context of data processing, one also speaks of **unstructured data**.

Differentiation from data and web mining



Interdisciplinary interfaces



Text mining challenges

The challenges of data mining also apply to text mining:

Very large amounts of (text) data

High dimensionality

"**Noise**" in the data

Understanding the results

However, additional difficulties arise:

Texts are usually **not intended for computer-based processing**

Reason: texts have a **complex and little standardized structure**

Language, **morphology**, syntax and **semantics** strongly depend on the author himself and the target audience of the text

TECHNISCHE UNIVERSITÄT DRESDEN

CHAIR OF BUSINESS INFORMATION SYSTEMS
ESP. INTELLIGENT SYSTEMS AND SERVICES

DRESDEN concept

Prof. Dr. Alfred Benedikt Brendel
Chair of Business Information Systems, esp. Intelligent Systems and Services

Text Mining

Procedure model

Process models in business informatics

Process models specify the sequence in which certain activities should be carried out or procedures should be used in order to successfully complete (IT) projects; this serves on the one hand to structure tasks and on the other hand to reduce complexity within these projects.

Process models such as the KDD, SEMMA or CRISP-DM process structure phases and tasks of the data mining process

Procedure model for text mining

According to this understanding of the term and in connection with the definition of text mining, a procedure model for text mining thus considers the flow of the analytical process for the (partially) automated extraction of knowledge from unstructured data.

The process model divides this process into several phases, in which, in turn, different activities must be carried out in order to convert the raw data into useful knowledge over several steps. To perform the activities, procedures, methods and techniques must be applied to achieve the desired results.

Process models and procedures

Literature Review: Implications

Findings

Variety of activities and procedures

Holistic, cross-case view of the process is missing

Previous generic models show **only rough flow**

Structural requirements for a generic process model

Design of the procedure model

Phase arrangement and structuring

Feedback loops

Functional requirements for a generic process model

Task definition

Document preprocessing

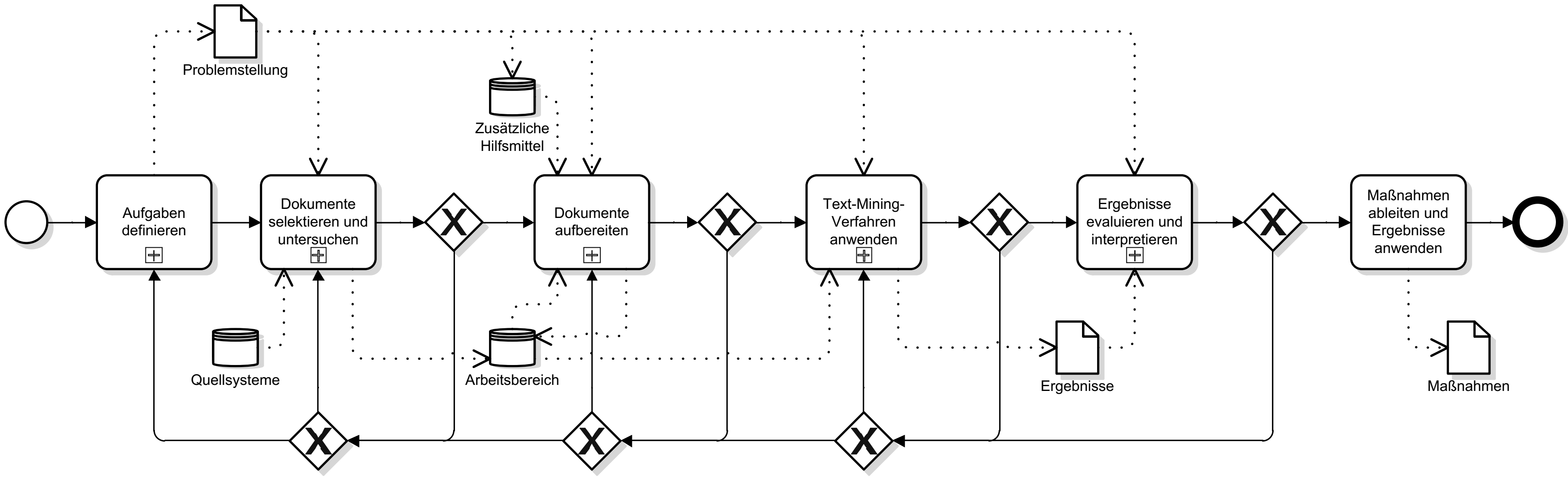
Data Analysis

Procedure for the implementation of the activities

Procedure model according to Schieber & Hilbert

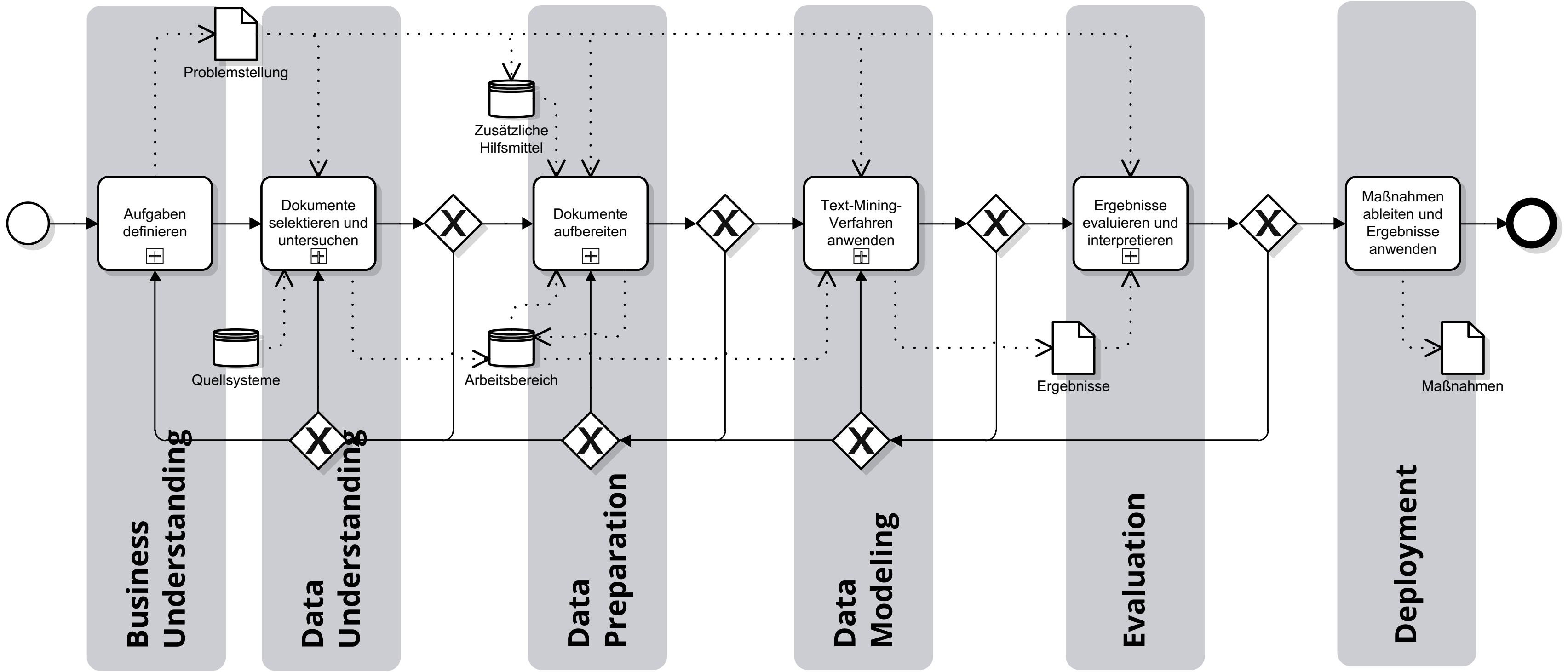
Process models such as the KDD, SEMMA or CRISP-DM process structure phases and tasks of the data mining process

Analogously, the procedure model according to Schieber & Hilbert structures the phases and tasks of the text mining process



Comparison of the processes

The phases and tasks are similar to the CRISP-DM process,...
...differ, however, due to the data basis, especially in the phases of document preparation and analysis



The text mining process at a glance

Process activity	Description
Define task	The goal of the text mining project and the task must be clearly defined and described. This has an **impact on the concrete design of the process** as well as the choice of procedures.
Select documents	Source systems and target data are determined and **extracted** for analysis.
Prepare documents	The extracted data is processed; a distinction is made between **linguistic and technical processing. This** phase is particularly different from data mining processes and has a **strong influence on the achievable results.**
Apply text mining techniques	Text mining methods are applied to the processed data; for example, **documents** are **classified or grouped.**
Evaluate and interpret results	The results are reviewed and evaluated; specific, statistical **key figures** can be **evaluated** for this purpose, depending on the procedure.
Derive and apply measures	If the results are satisfactory, measures can be derived in line with the terms of reference.

Prof. Dr. Alfred Benedikt Brendel
Chair of Business Information Systems, esp. Intelligent Systems and Services

Text Mining

Procedure model: task definition

Task definition

Activities

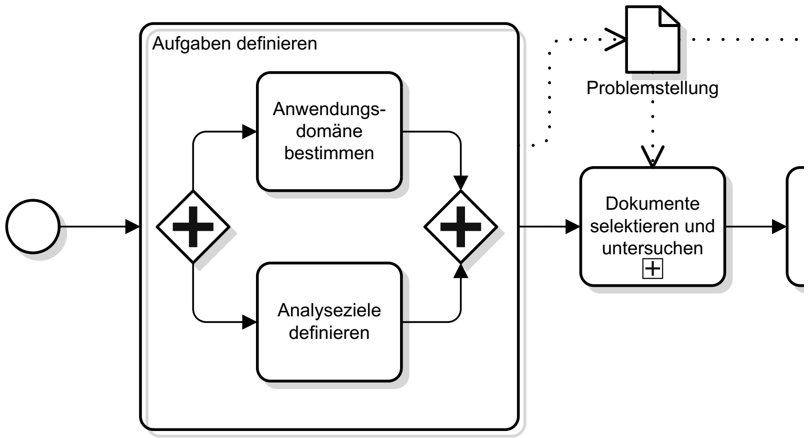
Determine application domain
— (Business) **context**
— Backgrounds
— Important **keywords** in the domain

Define analysis goals
What is to be found out or achieved by the process?

Output
Problem definition

Contains information about context and goals

Influences the process in later steps, e.g. during document pre-processing or processing or in the selection of analysis procedures



Task definition

Use case: Text mining in customer support

Text mining to categorize customer queries

Manufacturers receive very many support requests by mail

Text mining to **channel and distribute the flood of requests**

Support requests by mail or ticket

Goal: Automatic **presorting of mails or tickets**

Examination of the content, what words are included, what topics do they indicate

Assignment of mails to support topics

Forwarding of the mails to the appropriate processor

Task definition

Use case: Product development

Text mining for product review analysis

Many reviews and testimonials about products are available on Web 2.0

Text mining to **uncover implications for product development**

Collection of customer reviews

Goal: Detection of **weak points of the product**

Identification of product features in the texts (e.g. battery life)

Identification of opinion-forming adjectives that are related to the product property (e.g., low).

Aggregation and ranking of ratings

Investigation of particularly critically evaluated product properties

Prof. Dr. Alfred Benedikt Brendel
Chair of Business Information Systems, esp. Intelligent Systems and Services

Text Mining

Procedure model: document selection

Document selection and examination

Input

Problem

Source systems

Activities

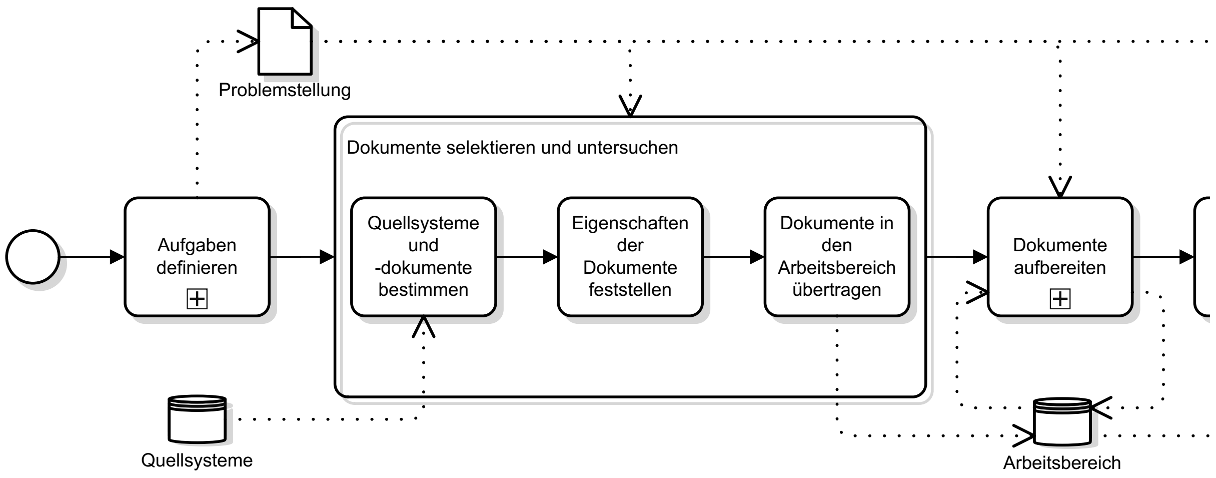
Determine source systems and documents

Determine properties of the documents

Transfer documents to the workspace

Output

Data in the workspace



Document selection and investigation

Source systems and documents

Source systems

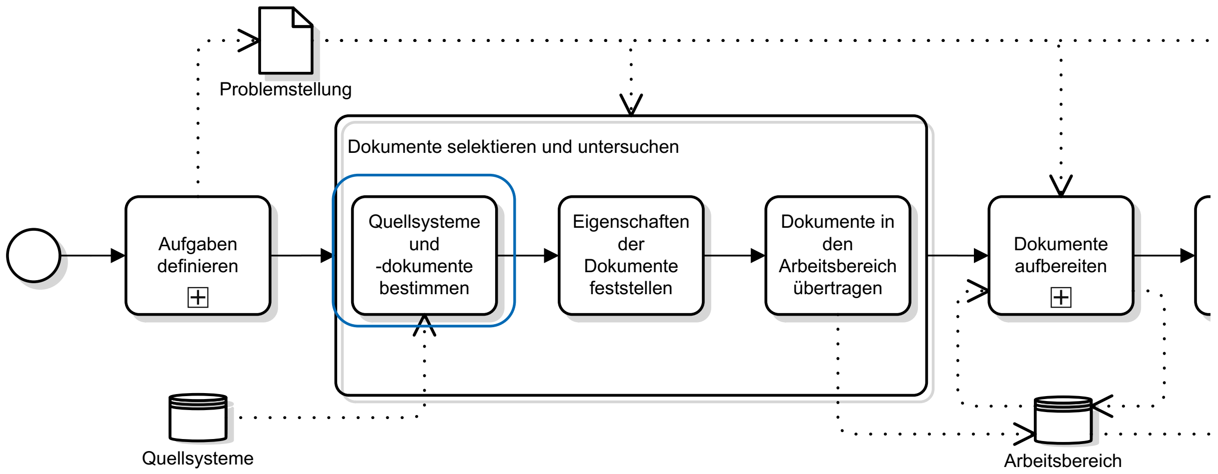
Online store, web platform

E-mail archive

Collection of electronic documents

Source documents (Corpus)

- Customer reviews from online stores or rating portals
- Emails to the central address of a company
- Electronic copies of handwritten or printed essays



Document selection and investigation
Use case: Extraction of texts from weblogs

Weblogs are a popular tool for publishing content on the Web, and are gaining new importance for analysis under the term Social BI

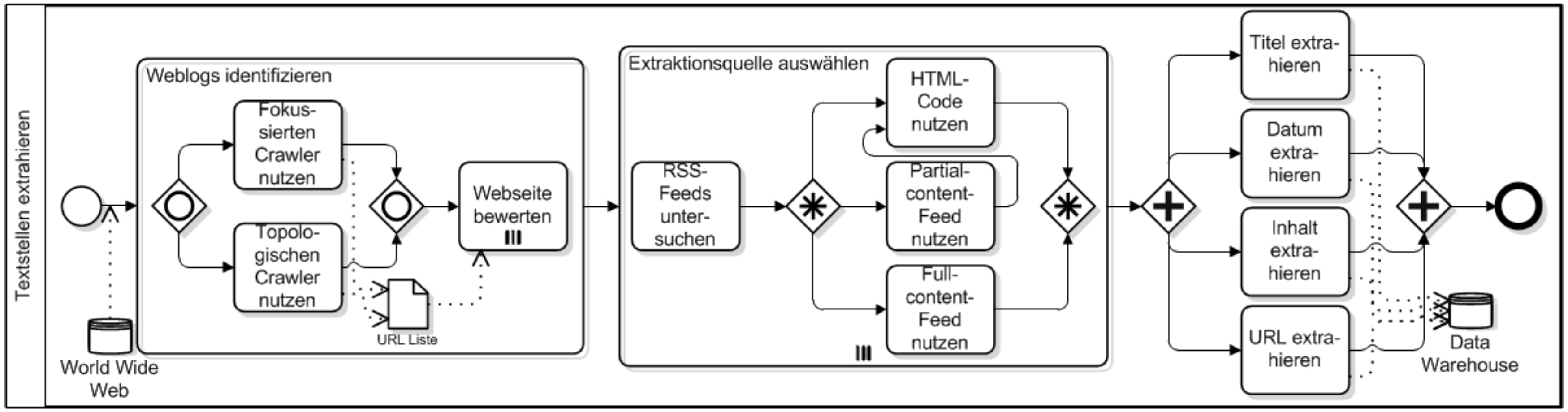
Documents must be extracted, transformed and loaded in traditional BI systems

Tasks in the context of extraction from weblogs:

Identification of relevant websites

Selection of the best structured extraction source

Extraction of the desired data



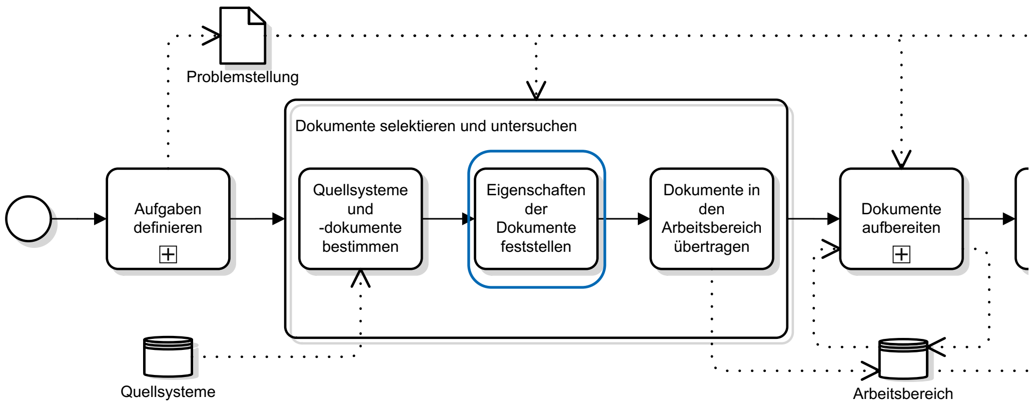
Document selection and examination
Document properties

Metadata, e.g.

- Language of the texts
- Availability of structured data besides text data, e.g. timestamps, ratings, etc.
- Expressions

Division into **training and test data set**

- depending on the distribution of target categories
- esp. for classification tasks



CHAIR OF BUSINESS INFORMATION SYSTEMS
ESP. INTELLIGENT SYSTEMS AND SERVICES

Prof. Dr. Alfred Benedikt Brendel
Chair of Business Information Systems, esp. Intelligent Systems and Services

Text Mining

Procedure model: document preparation

Document preparation

Input

Problem

Data from the workspace

Additional tools

Activities

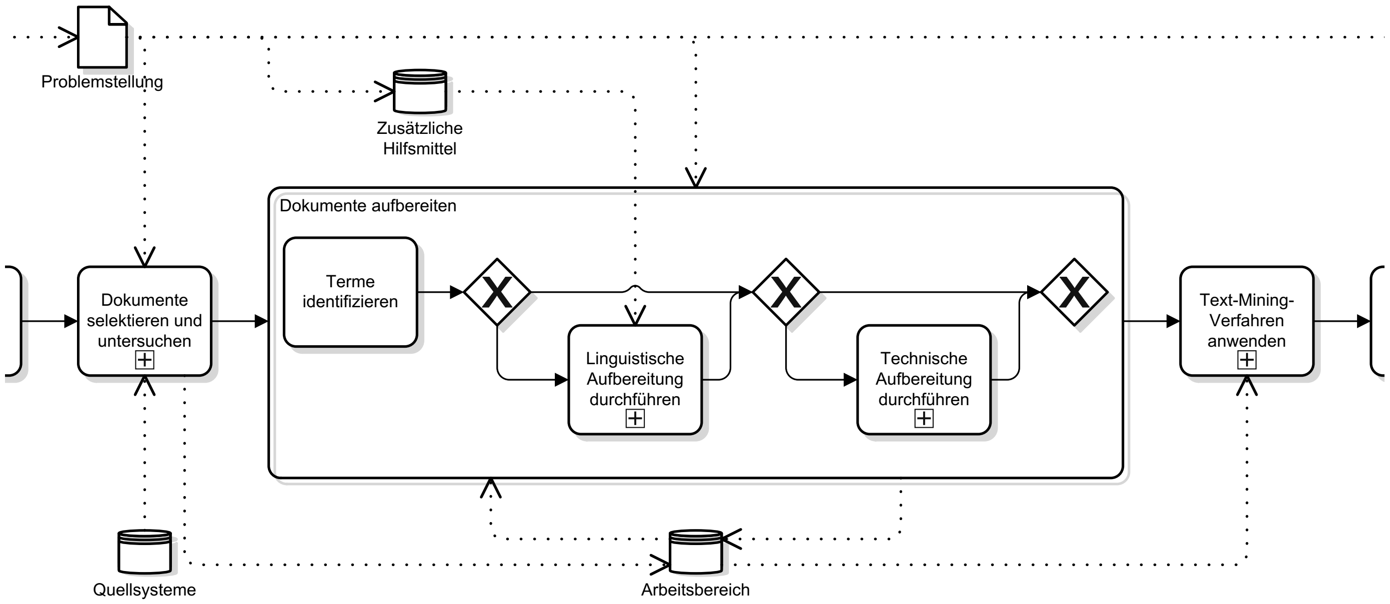
Term identification

Linguistic Processing

Technical Preparation

Output

Processed data In the workspace



Document preparation
Term identification

Terms (or tokens) are components of a text

mostly **words,** but also

Telephone numbers or

E-mails etc.

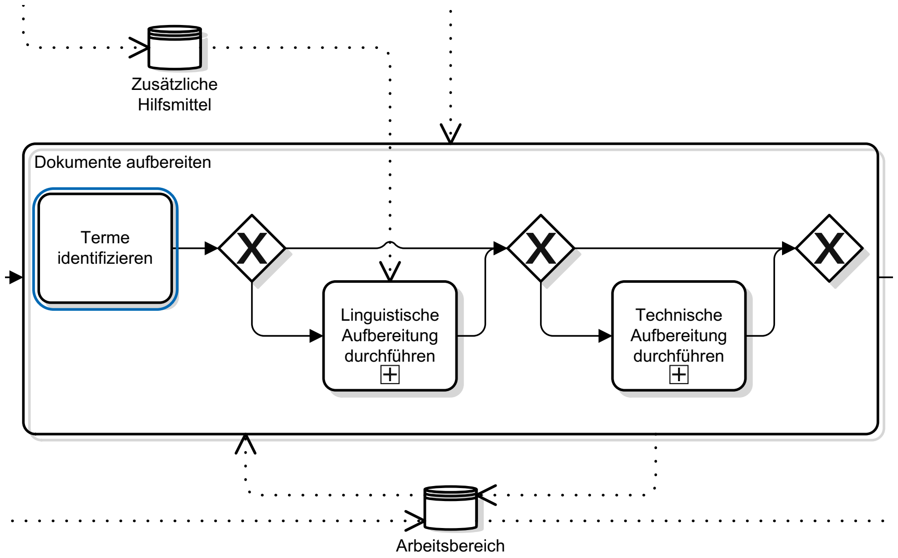
The decomposition of texts into terms is called tokenization and is used to structure the unstructured text data.

Due to the gained structuring the text data are prepared for classic data mining methods

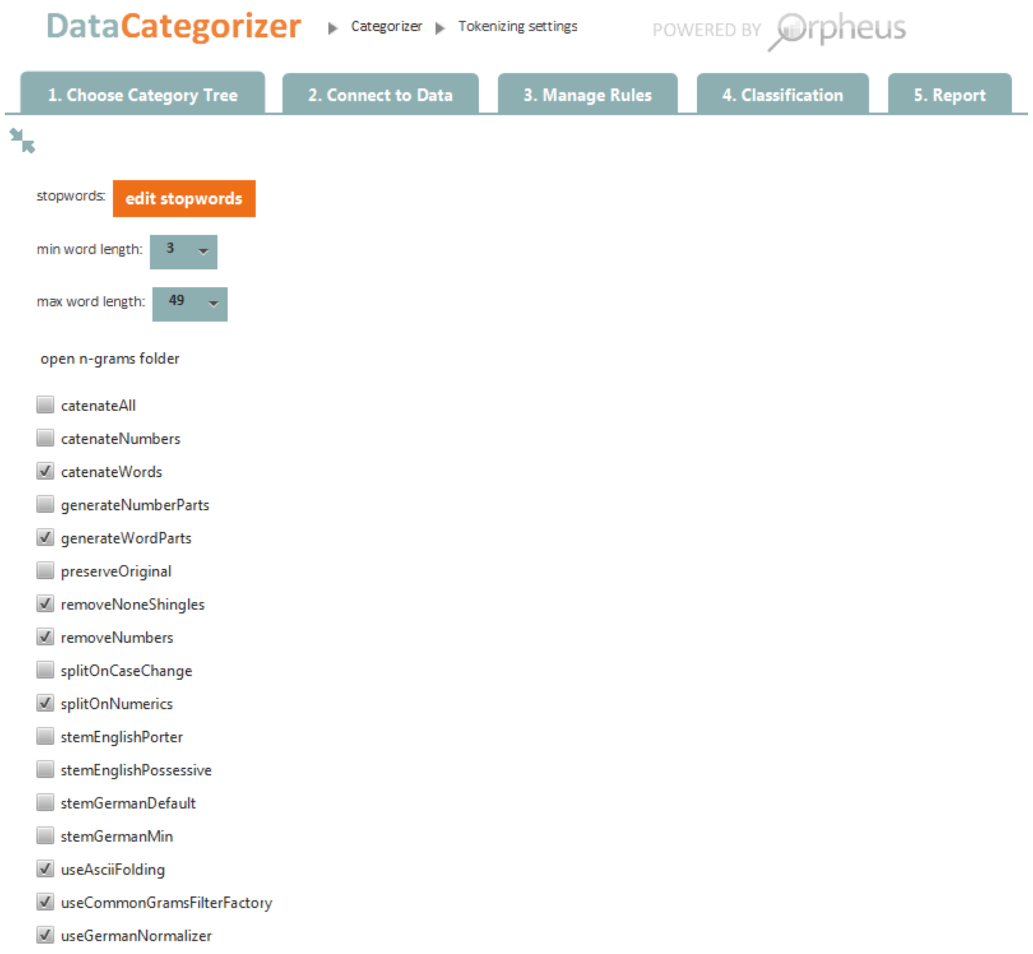
Tokenization is usually performed by separation based on

Space or

Punctuation marks performed.



Document preparation
Use case

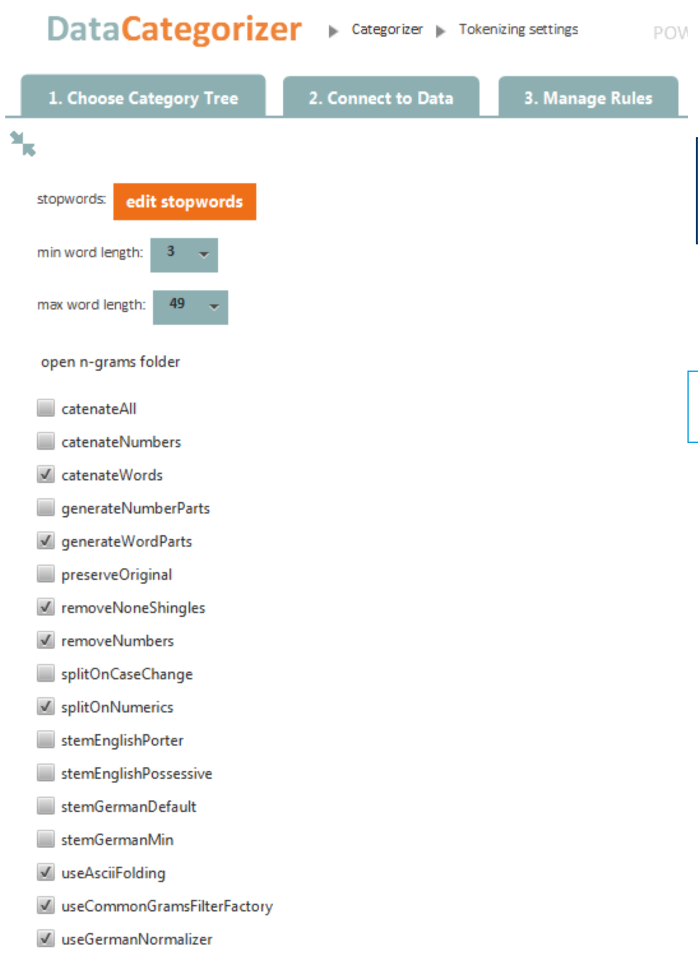


Orpheus **DataCategorizer**

Destination
Automatic categorization of order and invoice documents

Procedure
- Learning on a **pre-classified sample**
- **Automatic identification of** selective terms in the sample
- Applying the learned categorization to the sample to **determine the classification quality**
- Possibly manual corrections or extension of the automated term search
- **Applying** the final categorization to a new data set

Document preparation
Use case



Supplier	Product group	Invoice item

RS Components GmbH Hardware Mini-ITX Intel Core i5 660 Q57 4GB DDR3

Stop word

Compound term

Contains only digits

Less than three characters

Result of tokenization

components hardware miniitx mini itx intel core ddr

Document preparation
Linguistic preparation

Activities
Filter terms

Perform lexical analysis

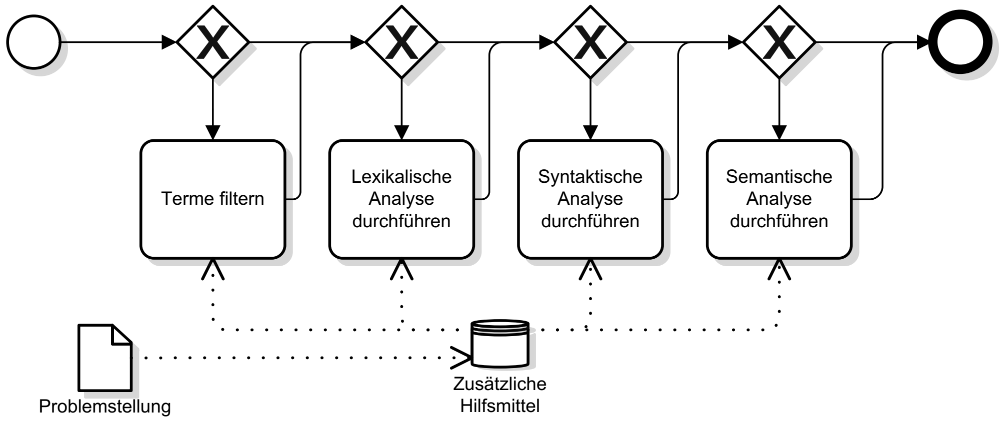
Perform syntactic analysis

Perform semantic analysis

Focal points of the technical preparation
Examination of **linguistic aspects** of the texts

Processes that define terms on the basis of linguistic properties

Recognize syntactic and semantic elements



Document preparation
Lexical analysis

Target:

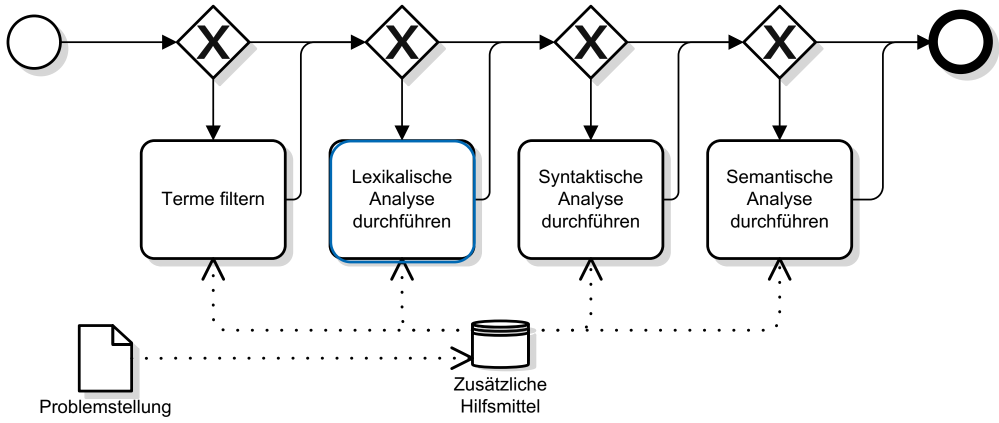
Examination of **linguistic aspects of** the texts

Reduction of the amount of data

Procedure: depending on the problem, optional activities

Term filtering: filtering of irrelevant terms, e.g. by stop word lists

Lexical analysis:
- Conversion of terms into their **root form** (stemming); Ex: "[I] went" and "[he went]" are stemmed. went" are transformed into the base form "to go
- **Disadvantage**: depending on the reduction, the interpretability of the terms can be reduced drastically



Document preparation
Use case: master form reduction

Effects of stem form reduction using the example of the English language

Depending on the stemming algorithm, the terms are converted to a uniform root form; Ex:

The terms "am", "are", "is" are transformed into "be".

The terms "car", "cars", "car's", "cars'" are transformed into "car".

The result of stemming could therefore look like this:

Original: "the boy's cars are different colors"

Stemmed: "the boy car be differ color"

The example illustrates both

the potential with respect to the **reducibility of the terms**

as well as the significantly **reduced interpretability** compared to the original text or terms.

Document preparation
Syntactic analysis

Target:

Examination of **linguistic aspects of** the texts

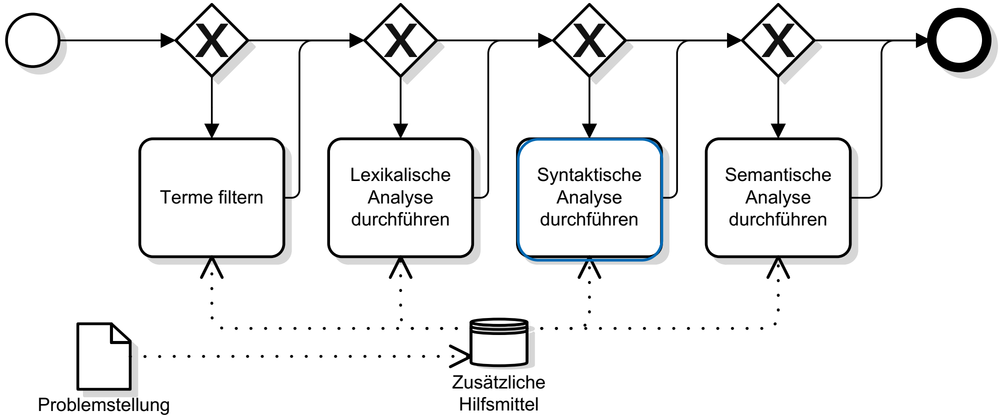
Study of **word types and sentence structure**

Thereby: **targeted extraction of information from specific, syntactic units**

Procedure:

Recognition of adjectives, nouns etc. by a probabilistic model; also as **part-of-speech (POS) tagging.** tagging

Examination of the sentence structure by so-called **parsing**



Document preparation
Semantic analysis

Target:

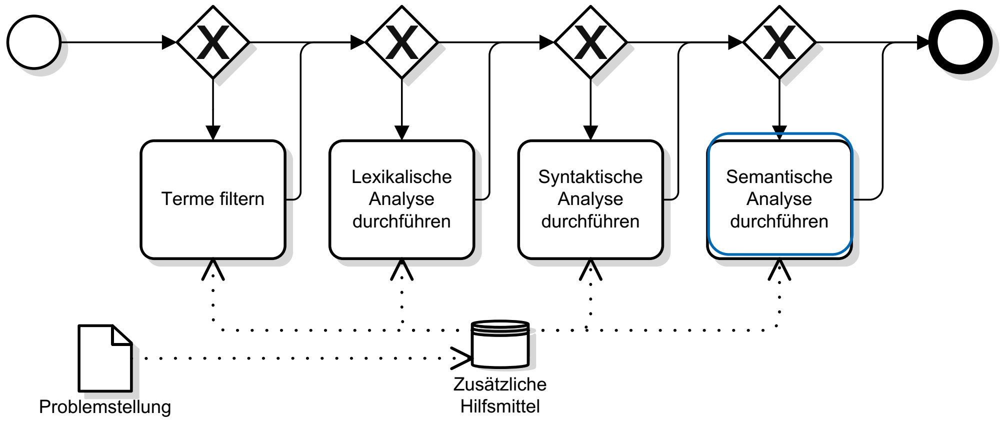
Examination of **linguistic aspects of** the texts

Study of the **meaning of words** in the sentence

Procedure:

Collocations: Collocations frequently occurring word combinations that are found in a certain discourse area.

Thesauruses and **ontologies**:
Ontologies form concepts and relationship between these terms and serve the representation of knowledge structures; thus they enable computers both logical reasoning and the recognition of context.



Document preparation
Use case: Irony detection

Target:

Identification of ironic statements in customer reviews

Separate treatment of these statements in the context of **opinion mining**

Procedure:

Determining the **context of** statements

Recognizing clues to ironic statements

Linguistic Foundations:

Definition according to Lapp (1992): Irony is used to express attitudes or feelings that one does not have and at the same time to make one understand that one does not have them.

Main motive of use: **evaluation** of persons, actions, objects and properties.

Use almost exclusively for negative reviews

Document preparation
Use case: Irony detection

Detection in direct communication:

Facial expression,

Gestures,

Emphases

Recognition in texts:

Problem: Facial expressions, gestures and intonations cannot be clearly depicted in texts.

Irony signals in texts:

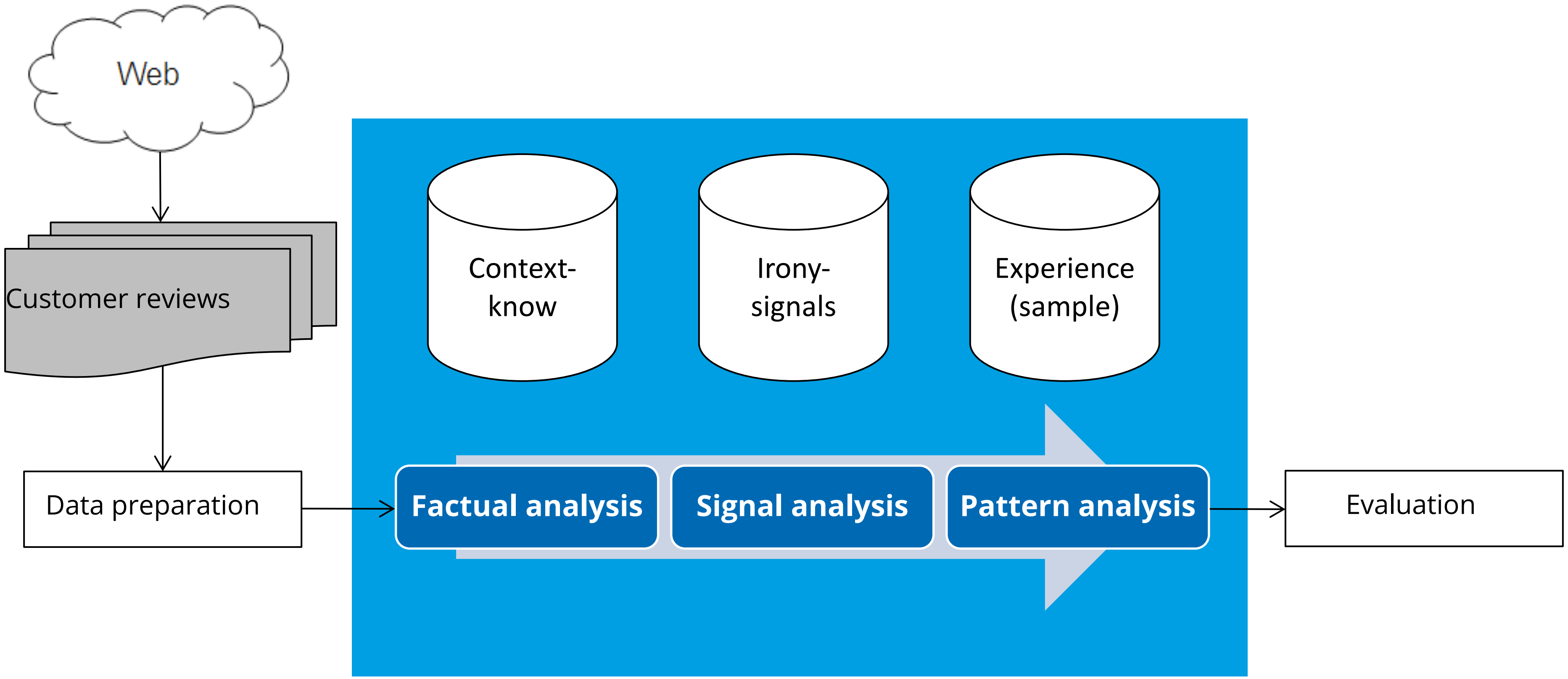
— Phonological-graphemic level (quotation marks, exclamation marks, capital letters, smileys,...)

— Morphological-syntactical level (superlatives)

Facilitating identification through **knowledge of the context** and situation of a statement

Document preparation
Use case: Irony detection

Irony detection process



Document preparation
Technical preparation

Activities
Indexing and weighting terms

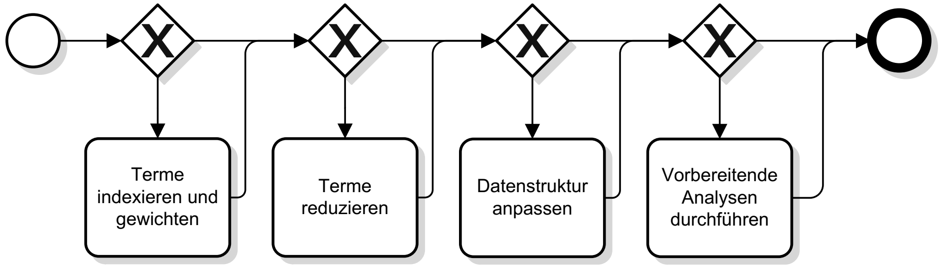
Reduce terms

Customize data structure

Perform preparatory analyses

Focal points of the technical preparation
Methods that reduce terms on the basis of statistical ratios

Transformation steps to convert the database into a different structure



Document preparation

Weighting and reducing terms

Objectives:

Weighting of terms based on statistical ratios; this allows representative terms to be identified in the data

Reduction of terms by defining threshold values; this allows data reduction without linguistic processing

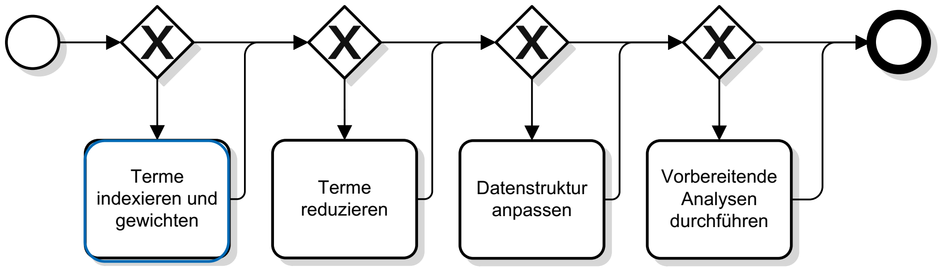
Widespread metrics:

Binary key figures express only the occurrence of a term in a document

Term frequencies indicate how often a term occurs in a document

The **document frequency** specifies in how many documents a term occurs

Weighted Frequency: The **TF-IDF measure** combines the two frequency measures



Document preparation

Customize data structure

Many text mining techniques have their origins in data mining, and often classical data mining techniques can also be applied to text data, provided that the data is available in a suitable structure.

In the literature, the vector space model according to Salton et al. (1975) is mostly used for this purpose, whereby the text data are available in a structured tabular form:

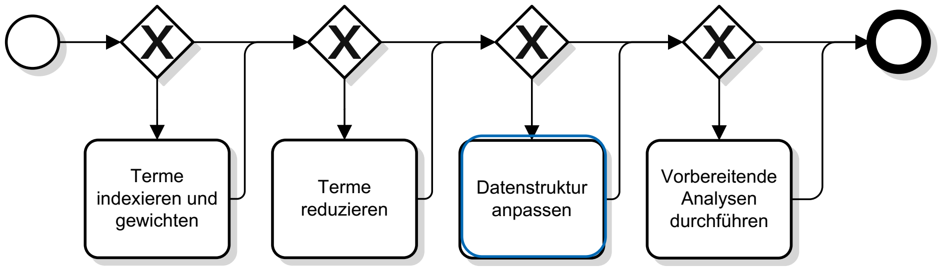
Terms and documents are transformed into a matrix

The lines list the terms

In the columns the documents are listed

The intersections show whether a term occurs in a document or not

	D1	D2	D3
T1	1	0	1
T2	0	0	1
T3	1	1	0



Document preparation
Customize data structure

Instead of the occurrence of terms in documents, other key figures can also be displayed in the matrix, e.g.

The **frequency** how often a term is mentioned in the document

TF-IDF values of the terms

etc.

The vector space model makes it easy to identify similarities between documents and terms

	D1	D2	D3
Text	1	0	1
Data	0	1	0
Mining	1	1	0
Informatio n	1	1	1
Retrieval	0	0	1

Identification of documents with similar content

	D1	D2	D3
Text	1	0	1
Data	0	1	0
Mining	1	1	0
Information	1	1	1
Retrieval	0	0	1

Identification of terms with similar contexts

TECHNISCHE UNIVERSITÄT DRESDEN

CHAIR OF BUSINESS INFORMATION SYSTEMS
ESP. INTELLIGENT SYSTEMS AND SERVICES

DRESDEN concept

Prof. Dr. Alfred Benedikt Brendel
Chair of Business Information Systems, esp. Intelligent Systems and Services

Text Mining

Procedure model: Text mining procedure

Text mining methods

Input

Problem

Workspace

Activities

Apply classification methods

Segmentation procedure apply

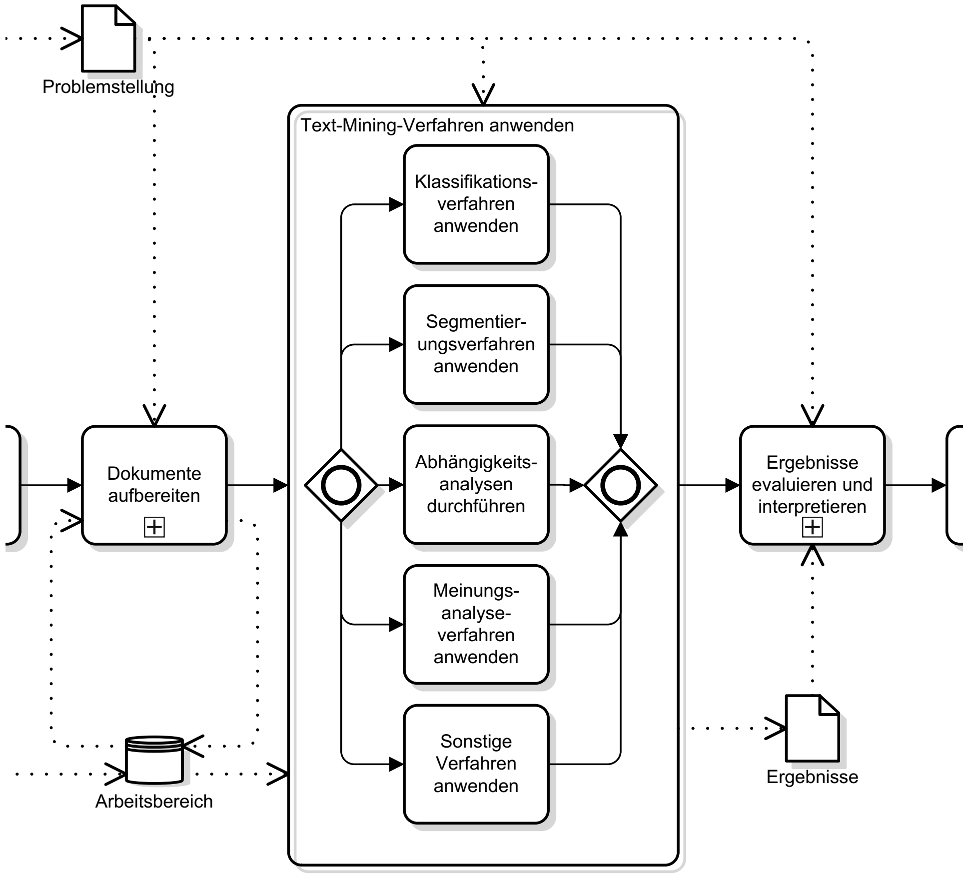
Dependency analyses perform

Opinion analysis procedures apply

Use other methods

Output

Results



Text mining method
General

Most of the methods have their origin in classical data mining methods

They can be applied in the same way to text data, provided that they have been prepared for the procedures

The preparation of the data includes in particular

tokenization, which identifies and separates individual terms in the texts

The transformation of the data into the **vector space model**, which gives the data a structure with which data mining methods can be applied.

However, there are also methods that have their origin in computational linguistics

Such methods are more specialized to the needs of text data, but many of these methods still require a vector space model as a basis

Text mining methods
Classification methods

Classification methods group data sets into a given category system

The procedures learn their model using pre-classified training data sets

The learned model is then applied to new data sets

Methods used in the literature:

Decision trees

Artificial Neural Networks

Support Vector Machine

Text mining methods
Segmentation method

Segmentation methods group data sets without depending on a predefined category system

The methods examine the data sets and calculate their similarity in terms of their attributes

In the context of text analysis, the main attributes available are the terms contained in the text

Documents are therefore similar if they have many terms in common

Methods used in the literature:

K-Means algorithm

Topic Modeling

Text mining methods
Dependency analyses

Dependency analyses map dependencies between terms

In contrast to classification and segmentation methods, these dependencies are calculated over the entire data set

This means that these results apply to all documents in the dataset

By determining term dependencies and linking these terms to term networks, ontologies can be created automatically

Methods used in the literature:

Association analysis

Coocurrence analysis

Text mining method
Opinion analysis

Opinion analysis techniques attempt to determine the mood of a text

Depending on the desired granularity (and simultaneously increasing complexity), the opinion or sentiment can be based on

of the entire document,

of a set or

be determined in relation to properties of an evaluated object

Tonality is best determined on the basis of adjectives and adverbs

A particular challenge here is to recognize the relationships between evaluating terms and evaluated objects

Methods used in the literature:

Opinion Observer

Red opal

Text mining methods
Other analysis methods

Text summarization
The procedures select essential sentences of a document on the basis of statistical key figures
The user can grasp the core content of a document more quickly

Information Extraction
The procedures search defined patterns in documents and can thereby convert terms into information
However, to do so, the procedures require manually created tools such as dictionaries

Method for the visualization of interrelationships (information visualization)
The methods represent documents or their keywords in hierarchies, graphs or networks
This enables the user to navigate through the documents and easily grasp content or contexts

Prof. Dr. Alfred Benedikt Brendel
Chair of Business Information Systems, esp. Intelligent Systems and Services

Text Mining

Procedure model: Evaluation of results

Outcome evaluation

Input

Problem

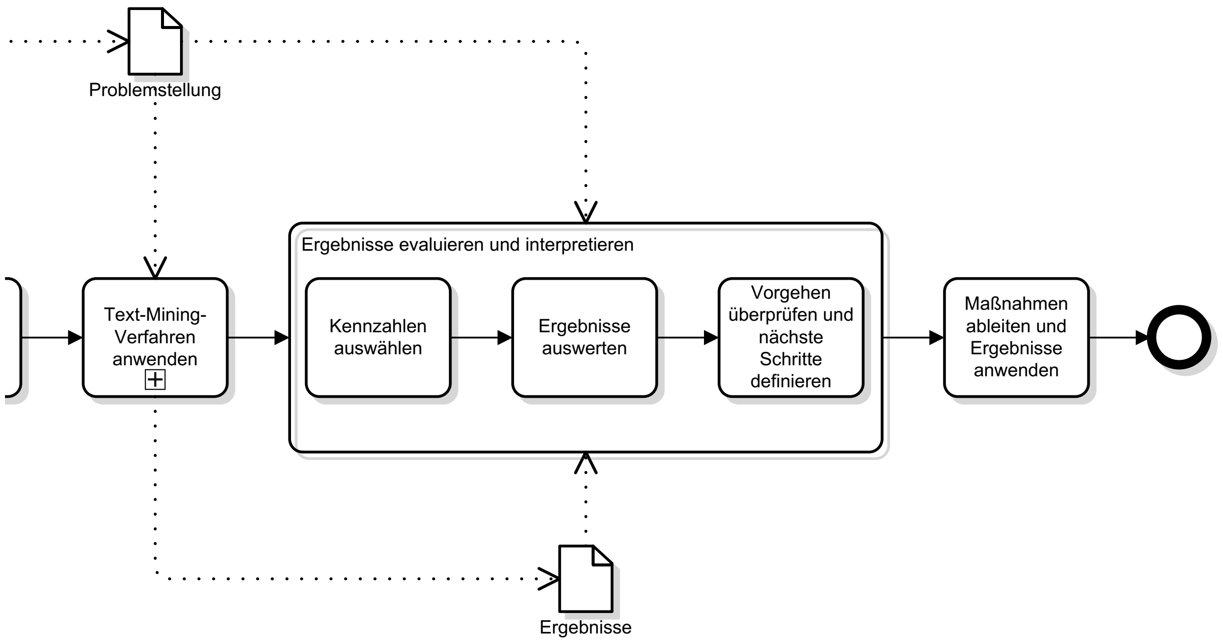
Results

Activities

Select key figures

Evaluate results

Review procedure and define next steps



Outcome evaluation

In this step, it is checked whether the initial question could be answered by the text mining process. Here are

assess the results from the upstream step of applying the analytical methods using **quality criteria appropriate** to the method, and

Match the results from the upstream step of applying the analysis procedures **with the problem definition**

If the check is positive, the next process step is executed.

If the test is negative,

it must first be analyzed **where improvements are** needed (preparation, data analysis,...)

and then the process must **jump back to this point** to be executed again with changed parameters or other procedures

Prof. Dr. Alfred Benedikt Brendel
Chair of Business Information Systems, esp. Intelligent Systems and Services

Text Mining

Procedure model: Application

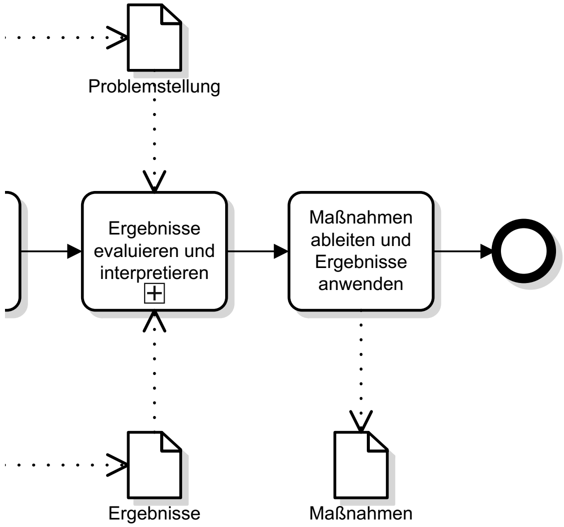
Application

Activities

Derive measures and apply results

Output

Measures



Prof. Dr. Alfred Benedikt Brendel
Chair of Business Information Systems, esp. Intelligent Systems and Services

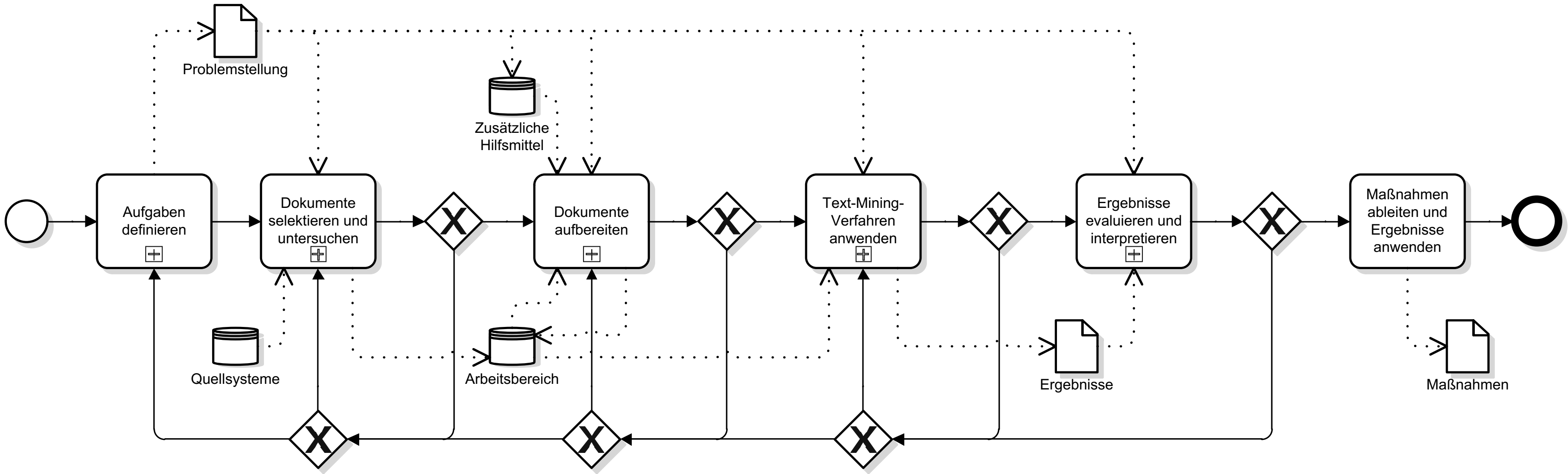
Text Mining

Procedure model: feedback loops

Feedback loops

Feedback loops are an important part of the process model

They allow iterative procedures and improvements within a text mining project, e.g. to adjust parameters of the preprocessing or - if necessary - to correct the task.



Feedback loops

BU -> DS -> DP -> M -> E -> Depl

2 -> 1
3 -> 2
4 -> 1
4 -> 2

4 -> 3
5 -> 1
5 -> 4

The literature mentions seven situations in which feedback loops are absolutely necessary:

1. **From phase *Document Selection* to phase *Business Understanding*:**
   To improve document selection and examination, more information is needed regarding the application domain.
2. **From phase *Document Processing* to phase *Document Selection*:**
   To improve document preparation, the properties of the texts must be further investigated.
3. **From phase *Text Mining Procedure* to phase *Business Understanding* :**
   The loop allows the adjustment of the analysis objectives if the results are not satisfactory.
4. **From phase *Text Mining Procedure* to phase *Document Selection* :**
   The loop is necessary in case the wrong text mining methods were selected due to lack of document investigation, resulting in erroneous results.

Feedback loops

5. From phase *Text Mining Procedure* to phase *Document Processing*:
   To improve the analysis results, the documents must be reprocessed, e.g. because the requirements of the procedure were not yet known at the time of reprocessing.
6. From phase *Result Interpretation* to phase *Business Understanding*:
   In case of invalid results, the entire process must be repeated; the reason for this is misinterpretation of the task or poor execution of the process.
7. From phase *Result Interpretation* to phase *Text Mining Procedure* :
   To improve the results, the text analysis must be performed again, e.g. with new procedures or new parameters.

The model takes these situations into account and also provides further feedback options after all core phases, so that phases can be repeated recursively if required.

Prof. Dr. Alfred Benedikt Brendel
Chair of Business Information Systems, esp. Intelligent Systems and Services

Text Mining

Procedure model: Application examples

Prof. Dr. Alfred Benedikt Brendel
Chair of Business Information Systems, esp. Intelligent Systems and Services

Text Mining

Python example

TECHNISCHE UNIVERSITÄT DRESDEN

CHAIR OF BUSINESS INFORMATION SYSTEMS
ESP. INTELLIGENT SYSTEMS AND SERVICES

DRESDEN concept

Prof. Dr. Alfred Benedikt Brendel
Chair of Business Information Systems, esp. Intelligent Systems and Services

Thank you for your attention