Business Intelligence
SS 2015

Text Mining

W. Grossmann

Content

- Introduction and Terminology
- Data Preparation and Modelling
- Descriptive Analysis of the DTM
- Analysis of a Text Corpus
- Further Aspects of Text Mining

Introduction and Terminology, Data

- Text documents may be of different origin
  - Reports, abstracts, journal articles, blogs, tweets, email,…
- There are many different formats
  - .txt, .pdf, .doc, html, xml,…

Introduction and Terminology, Approaches

- Two different views:
  - Metadata view: a description of the document
    - There exist a number of standards for describing resources
    - One popular standard is the Dublin Core Metadata Initiative (DCMI): http://dublincore.org/

Introduction and Terminology, Approaches

- Example for Metadata using R:

  author : Rinderle-Ma, Grossmann

  datetimestamp: 2014-09-28 08:09:19

  description : Brief Task description

  heading : 1.3.5 Evaluation and Reporting Task

  id : 11

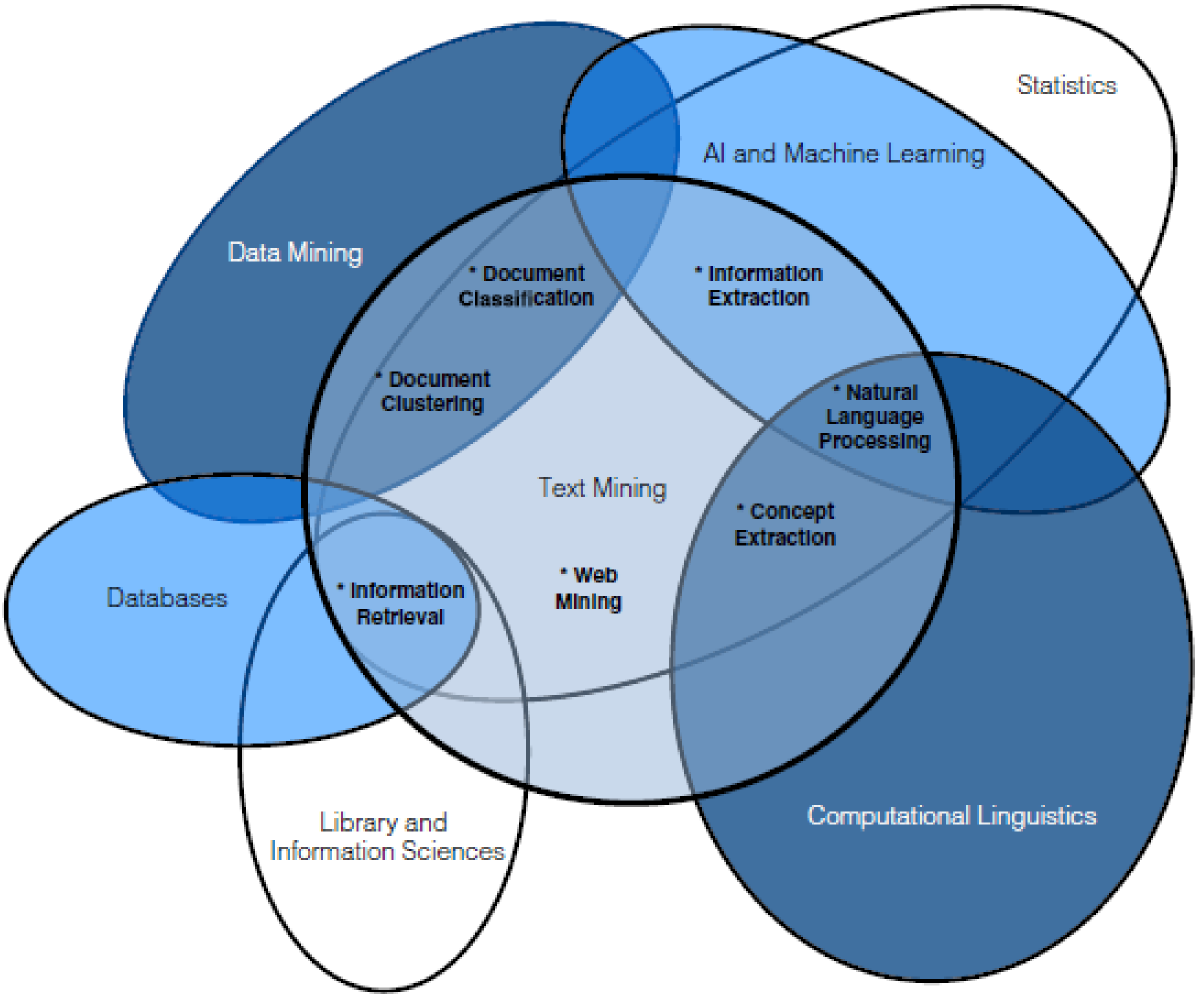
  language : en

  origin : Fundamentals of Business Intelligence V1.0Text Mining (Text Analytics view)

Introduction and Terminology, Approaches

- Knowledge and techniques from different areas are combined (graphic from Miner et al. Practical Text Mining)

Introduction and Terminology, Approaches



Introduction and Terminology, Levels of Text Mining

- Text Mining can be done at different levels
  - Word level
  - Sentence level
  - Document level
  - Corpus level = Collection of documents
- A Document can be defined in different ways:
  - Sections of a document, paragraphs in text, …
  - A tweet, an email,…

Introduction and Terminology, Analytical Goals

We will focus on analysis of a corpus

Analytical Goals in Text Mining

- Descriptive Goals: Description of the contents of the documents in a corpus based on properties of word frequencies in the documents.
- Understanding goals: Find clusters of documents which are similar with respect to content identify the topics in these groups.

Introduction and Terminology, Methodology

Template: Text Mining for a Corpus

- **Relevant Business and Data:** A text corpus defined by a collection of text documents
- **Analytical Goals:**
  - Description of the documents in the corpus
  - Clustering the documents in the corpus
  - Finding topics of the corpus
- **Modeling Task:** Definition of the document term matrix by appropriate data preparation steps
- **Analysis task:**
  - *Description of Corpus:* Determination of type-token relation and association measures; visualization of the content in the corpus using word clouds and correlation plots.
  - *Clustering documents:* Use cluster analysis methods of chapter 5 for cluster the documents based on the document term matrix.
  - *Topic Models:* Define a number of topics and find the probability of assignment of the documents to the topics.
- **Evaluation and Reporting Task:** Represent the results of the analysis by word clouds, by correlation plots and by characterization of the topics with terms.

Data Preparation and Modeling, Transformations

- Usually not the original text is used for text mining but a transformed (purged) text
- Basic standard transformations
  - Removal operations (punctuation, numbers, special characters (@, /,…), email address,
  - White space operations
  - Lower case letters
  - Stop words (articles, prepositions,…)
  - Stemming (words without endings)

Data Preparation and Modeling, Transformations

- Example sentence
  - Its main goals are the interpretation of the results in reference to domain knowledge and coming to a decision of how to proceed further.
- Transformed sentence
  - main goals interpretation results reference domain knowledg coming decision proceed

Data Preparation and Modeling, Document Term Matrix

- After the transformations the corpus consists of a number of documents with preprocessed terms
- These terms are organized in a list of tokens and the frequency of the tokens by tokenization
  - A token is defined by a n-gram = n contiguous words in the document, usually 1-grams(one term) or bigrams (2 words)

Data Preparation and Modeling, Document Term Matrix

- The basic unit for analysis is the

  document term matrix (DTM)

$$DTM = (t_{ij}), \quad i = 1,\ldots,d,\ j = 1,\ldots.,n$$

$$t_{ij} = \textit{frequency of term } j \textit{ in document } i$$

- Sometimes also the transposed matrix is used and called TDM(term document matrix)
- Other name for the DTM: Bag of words

Data Preparation and Modeling, Document Term Matrix

- An alternative to the DTM is often to replace the frequency simply by an indicator

$$DTMI = (d_{ij}), \quad i = 1,\ldots,d,\ j = 1,\ldots.,n$$

$$d_{ij} = \begin{cases} 1 & if\ \ term\ \ j\ occurs\ in\ document\ i \\ 0 & otherwise \end{cases}$$

Data Preparation and Modeling, Document Term Matrix

- Usually the DTM has many columns and contains many terms with low frequency
- General assumption:
  - Frequency of a term informs about the importance of the term for the contents
  - There are terms occurring frequently due to linguistic reasons, for example verbs like have, is, …

Data Preparation and Modeling, Document Term Matrix

- Solution of the problem:
  - Define upper and lower thresholds for the terms
  - Use instead of the DTM the TF-IDF = Term frequency – inverse document frequency matrix
- Inverse document frequency (IDF) = Number of documents divided by the frequency of the documents which contain the term
  - Reduces the importance of terms which occur in many documents

Data Preparation and Modeling, Document Term Matrix

- Formulas:

$$D = \{d_1, d_2, \ldots\} \; Documents$$

$$W = \{w_1, w_2, \ldots\} \; Words$$

$$IDF_{ij} = \frac{|D|}{1 + DF_{ij}}, \quad DF_{ij} = card\{d_i : w_j \in d_i\}$$

$$TF - IDF_{ij} = t_{ij} * \log(IDF_{ij})$$

- TF-IDF is of special interest for key-words differentiating between documents

Data Preparation and Modeling, Document Term Matrix

- Formulas:

$$D = \{d_1, d_2, \ldots\}\ Documents$$

$$W = \{w_1, w_2, \ldots\}\ Words$$

$$IDF_{ij} = \frac{|D|}{1 + DF_{ij}}, \quad DF_{ij} = card\{d_i : w_j \in d_i\}$$

$$TF - IDF_{ij} = t_{ij} * \log(IDF_{ij})$$

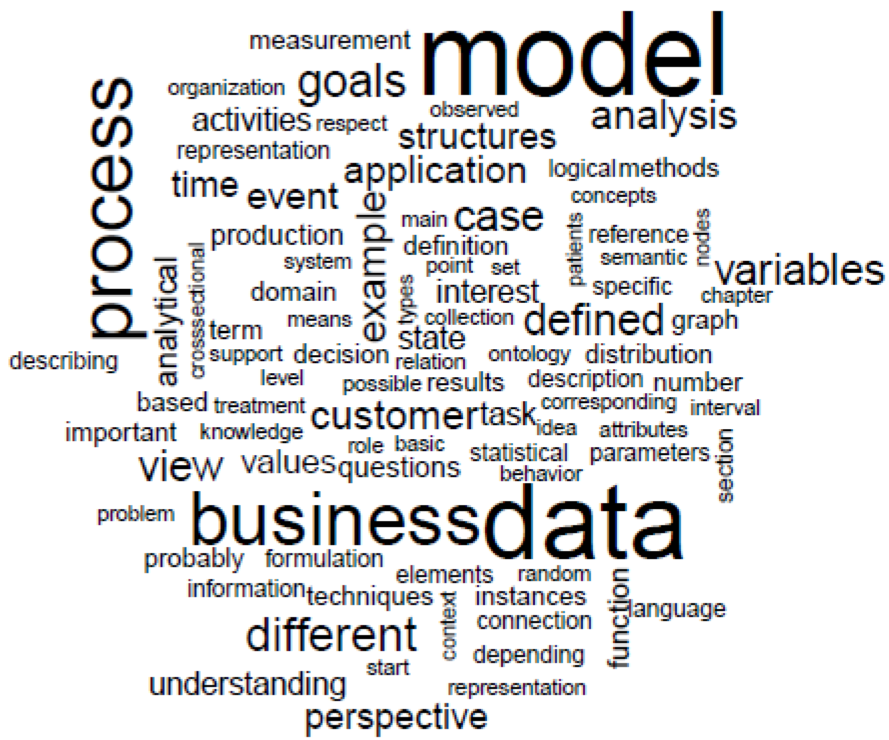
- TF-IDF is of special interest for key-words differentiating between documents

Descriptive Analysis of the DTM, Word Clouds

- A useful representation of a DTM is using a word cloud
  - Representation of the terms in the DTM with size according to the frequency of the terms
  - Usually the most frequent terms are in the center
  - Terms can be also rotated and colored

Descriptive Analysis of the DTM, Word Clouds

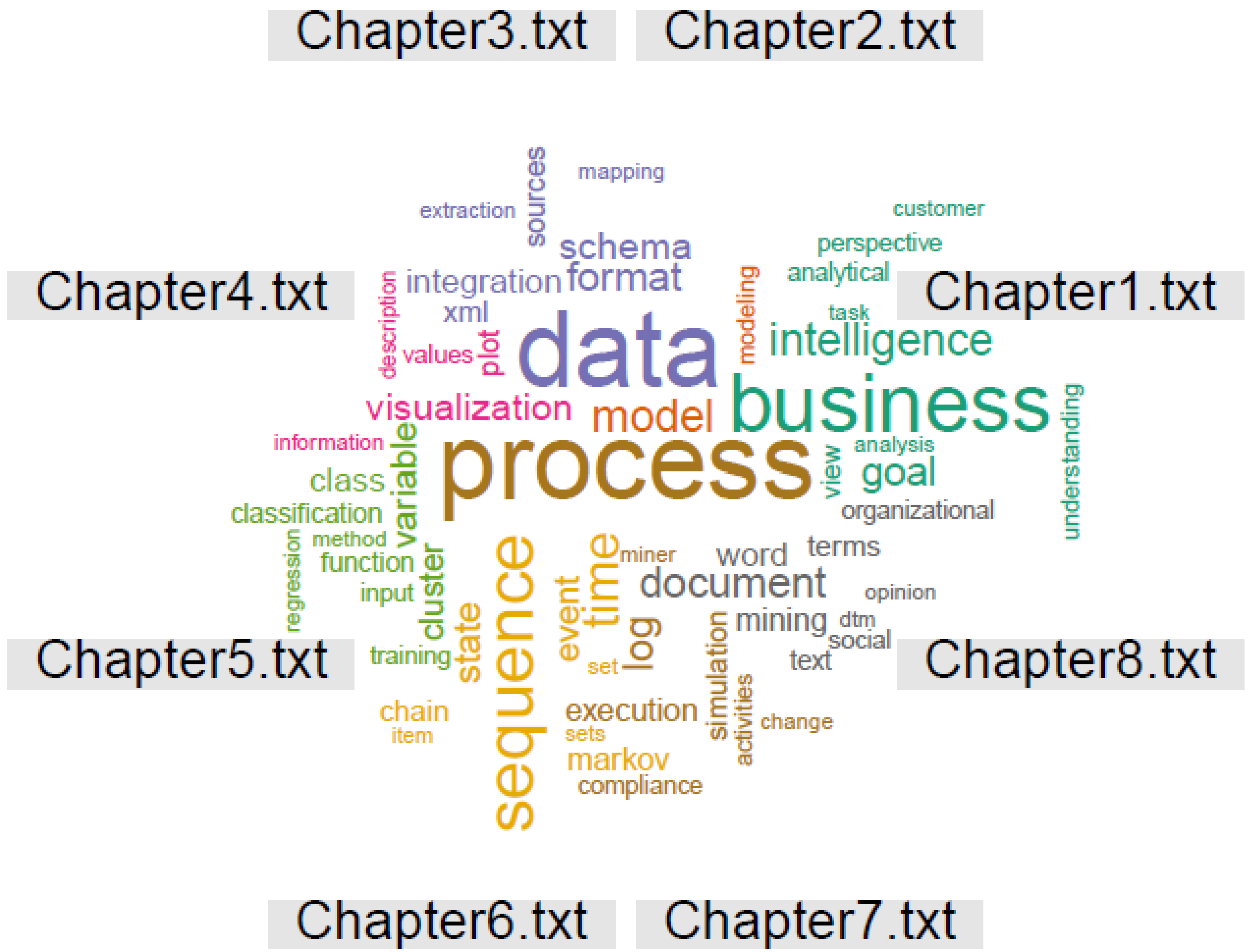
- Example of a word cloud



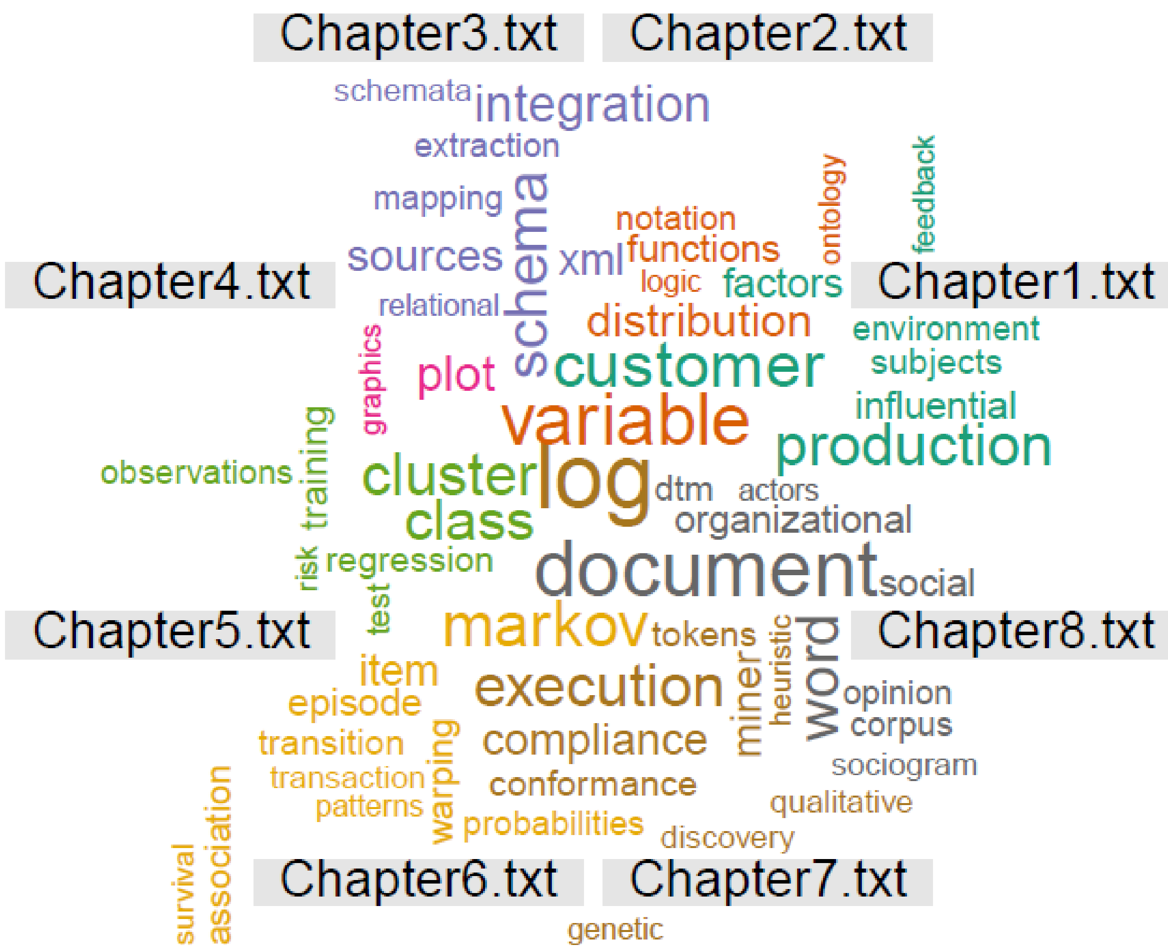
Descriptive Analysis of the DTM, Word Clouds

- For comparison of documents a comparison cloud is a useful tool
  - The documents are organized in an outer circle in the graphic
  - Terms are shown with size according to their frequency and are positioned according to their occurrence in the documents

Descriptive Analysis of the DTM, Comparison Cloud for terms



Descriptive Analysis of the DTM, Comparison Cloud for TF-IDF



Descriptive Analysis of the DTM, Associations between terms

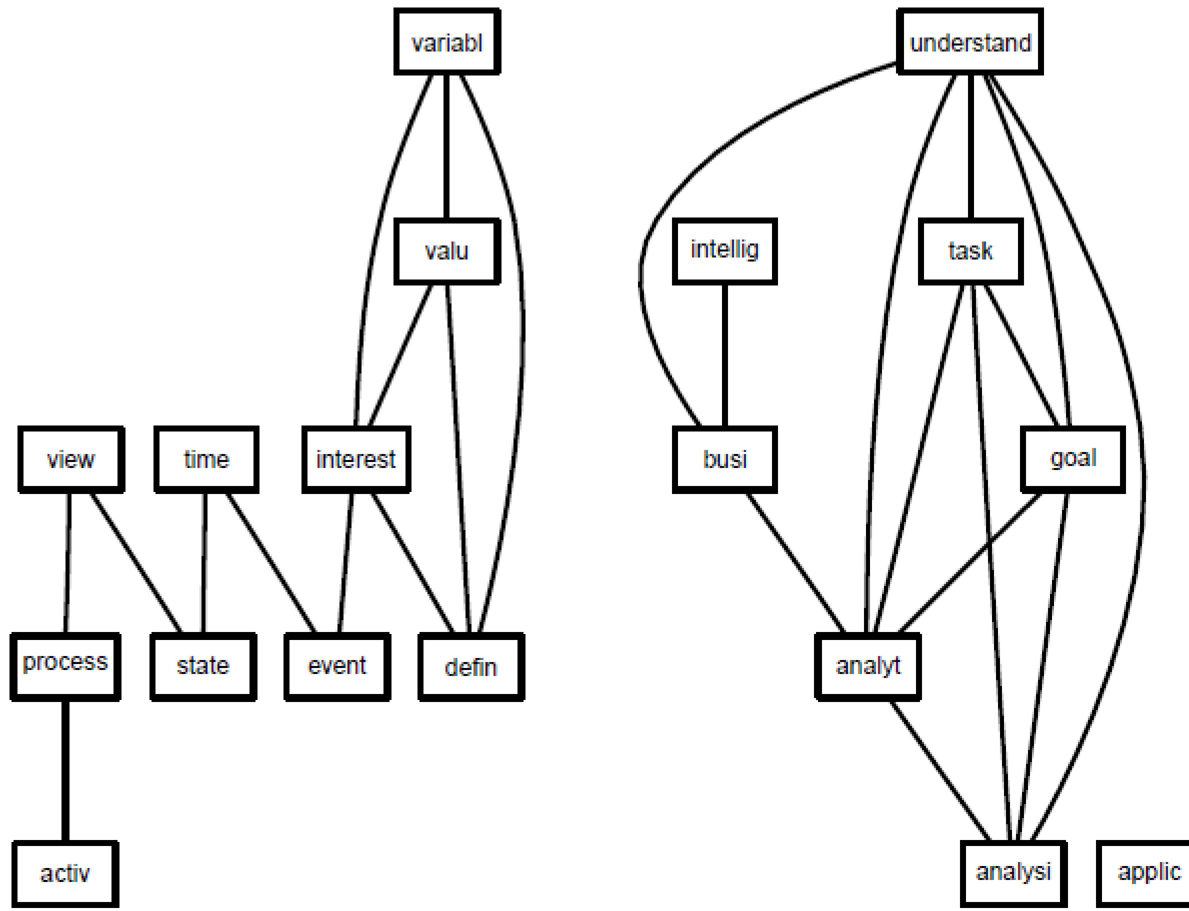
- Another way to describe the contents of a the document is to use correlation between the term frequencies in the different documents
- One can use also the indicator matrix for such associations

Descriptive Analysis of the DTM, Associations between terms

- Example

of association

above 0.7 for



Analysis of a Text Corpus, Clustering

- Cluster analysis of text data is based on the definition of similarities between documents
- For definition of the similarity the most popular measure is the cosine measure of the term frequencies in the documents

$$sim(d_i, d_j) = \frac{t_{i\bullet} \cdot t_{j\bullet}}{\left\| t_{i\bullet} \right\| \cdot \left\| t_{j\bullet} \right\|}$$

$$t_{i\bullet} = \text{ } frequency\ vector\ of\ terms\ in\ document\ i$$

Analysis of a Text Corpus, Clustering

- Based on this distance one can apply any cluster analysis algorithm (hierarchical or k-means)
- Many other methods have been proposed
  - Co-Clustering:
    - Interpret the DTM as a bipartite graph: Terms and documents
    - Partition the two sets in such a way that the edges between different clusters are minimized

Analysis of a Text Corpus, Classification

- Classification of documents can be done by interpretation of the terms as variables (features) describing the documents
  - Hence the DTM is a classical feature matrix for the documents and we can apply any classification algorithm
  - Frequently the indicator version DTMI is used instead of the DTM
  - Classical application: Spam detection in emails

Analysis of a Text Corpus, Topic Models

- Topic models is an advanced for method grouping documents and terms into topics
- Model:
  - Define a number of topics
  - For each document a distribution of the topics is assumed
  - For each topic the terms have a topic specific characteristic distribution

Analysis of a Text Corpus, Topic Models

- A topic model estimates the parameters of the distributions of the topics within the different documents and identifies the most frequent terms in each topic
- Usually the algorithm is applied for a different number of topics and the results are compared

Further Aspects of Text Mining, Analysis at the Word Level

- Words allow the representation of concepts with different words (synonyms)
- Concepts have many times and ordering
  – Hypernyms: terms representing a narrower concept
  – Hyponyms: Terms representing a broader concept
  – Part of relation between concepts

Further Aspects of Text Mining, Analysis at the Word Level

- Representation of such relations in a database for words
- For English terms WordNet (http://wordnet.princeton.edu) is an important resource which is free available
  - Examples: business, model, busy

Further Aspects of Text Mining, Analysis at the Sentence Level

- Analysis at the sentence level allows the syntactic analysis of a sentence

POS = Part of Speech Tagging

- Taggers identify the role of the words in a sentence
- Apache Open NLP is a frequently used tool (available in R) (https://opennlp.apache.org/)

Further Aspects of Text Mining, Analysis at the Sentence Level

- For tagging a standard are the Penn Treebank Tags (http://web.mit.edu/6.863/www/PennTreebankTags.html)
- Example:

Further Aspects of Text Mining, Analysis at the Sentence Level

- Example

The evaluation and reporting task looks at the analysis results from a global business perspective.

```
{(TOP
    (S
        (NP (DT  The) (NN evaluation) (CC and)
            (NN reporting) (NN task))
        (VP (VBZ looks)
        (PP (IN at) (NP (DT the) (NN analysis)
                        (NNS results)))
        (PP (IN from) (NP (DT a) (JJ global)
                    (NN business) (NN perspective)))))}
```

Further Aspects of Text Mining, Keyword Extraction

- Keyword extraction is usually done in a number of steps for creating features
  - TF-IDF for keyword search in a corpus
  - POS
  - Identifying words at the beginning of a text
  - Relation of the words to words in a thesaurus
  - Using such features each word gets a score and high scores define keywords
  - Learning the scores is based on supervised learning

Further Aspects of Text Mining, Opinion Mining

- Opinion Mining and Sentiment Analysis
- General setup
  - Basic unit is a document (cf. questionnaire)
  - Opinion holder : author of the document (cf. surveyed person)
  - Objects and features about which an opinion is stated (cf. questions in the questionnaire)
  - Polarity of the opinion (cf. answers in the questionnaire)

Further Aspects of Text Mining, Opinion Mining

- Main tasks in opinion mining
  - Finding in a document all opinionated sentences
  - Identify the objects and features about which an opinion is stated
  - Classify the opinion (polarity)

Further Aspects of Text Mining, Opinion Mining

- Finding opinionated sentences
- Different cases
  - Direct opinion (frequently based on adjectives like good, nice, bad, …)
  - Negation (non, not, negative prefixes)
  - Comparative opinion (better worse, comparative form of adjectives, …)
- POS is of utmost importance

Further Aspects of Text Mining, Opinion Mining

- Identification of objects and features
  - Many times only one object, in simple cases in the header of the document (metadata)
  - One of the first nouns in a document represent the objects
  - Features can be identified using data bases which characterize objects, for example products or movies
  - Knowledge about synonyms, hypernyms and hyponyms is necessary (WordNet)

Further Aspects of Text Mining, Opinion Mining

- Opinion classification
  - Polarity is usually based on wordlists or dictionaries for adjectives stating the polarity (twitter) http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon
  - A more elaborated dictionary is SentiWordNet (http://sentiwordnet.isti.cnr.it/)
  - Other methods are based on statistical classification

Further Aspects of Text Mining, Opinion Mining

- Problems with opinion mining
  - Opinion is not so well defined as objective features of products
  - Opinion is based on common sense (cf. SenticNet for such an approach http://sentic.net/sentics/)
  - Terms like precision and recall are difficult to apply
  - What is the group of opinion holders?
  - Can we identify spam opinion?