# Text Mining & Tools

## Knowledge Discovery and Data Mining 2 (VU) (706.715)

Roman Kern

ISDS, TU Graz

2018-03-15

# Text Mining

## Computer Linguistics, Natural Language Processing

# Overall Goal

Starting with (usually written) natural text provide means to automatically make use of the information encoded in the text.

# Application

- Information Extraction
  - e.g. identify mentions of persons
- Machine Translation
- Support other tasks
  - e.g. Information Retrieval
- Document based operations
  - Classification (e.g. spam), clustering (e.g. for navigation), summarisation

# Challenges

**Natural language ...**

- is ambiguous (word level, syntactic level)
- is noisy
- often contain errors (spelling mistakes, grammatical errors)
- can only understood with “world knowledge” (context information)

# General Approach

**From shallow to deep parsing**

1. Split sentences
2. Split tokens (words)
3. Apply Part-of-Speech tagging (word groups)
4. Chunking (phrases)
5. Build sentence tree (constituency parsing)
6. Extract grammatical relationship between words (dependency parsing)

Everything up to POS is considered to be shallow parsing; building a sentence tree is considered to be deep parsing.

# Tasks of NLP

**Tasks (or sub-fields) of NLP**

- Sentence Splitter, Tokeniser, Chunker
- Parser (different types)
- Named Entity Recognition
- Anaphora Resolution, Co-reference Resolution
- Word Sense Disambiguation
- Semantic Role Labelling
- ...

# Basics

## Bag of Words, Vectorisation, Vector Space Model

# Bag of Words

Replace a document (or sentence, paragraph)

- ... by a simple representation
- ... consisting of the words that appear in the document
- ... without keeping the sequence information

## Example

"The green house is next to the blue building" $\rightarrow$
$\{\textit{blue} : 1, \textit{building} : 1, \textit{green} : 1, \textit{house} : 1, \textit{is} : 1, \textit{next} : 1, \textit{the} : 2, \textit{to} : 1\}$

# Bag of Words

**Vector Space Model** (basic)

- Each document is an instance
- Each word represents an attribute
- The value of the attribute is the number of times the work appears in the document
- $\rightarrow$ **Document-Term Matrix**
  - Documents are rows, and terms are columns
  - The resulting matrix is very sparse (typically approx. $2\%$ non-zero entries)
- Suitable representation for many machine learning algorithms
- The process of transforming the text to an vector is called **vectorisation**
  - i.e. each word is a assigned to a fixed dimension
  - ... which needs to be the same for all documents

# Basic Techniques

## Vector Space Model

- Comparison of documents is then a comparison of vectors
  - Often via the cosine similarity
  - i.e. the angle between documents defines their relatedness
- Stop word list
  - Manually assembled list of non-content words
  - e.g. the, a, with, to, ...
  - Remove words without semantics
- Stemming
  - Remove inflexions (using rules)
  - Usually modify the suffix
- n-grams (e.g. bi-gram, tri-grams, skip-grams)
  - e.g. concatenate two adjacent words into a single term (bi-gram)

# Linguistics Basics

## And what do linguists tell us

# Linguistics Basics

- Phonetics
  - What are the acoustic building blocks of speech
- Phonology
  - How sound work in sequence to form language
  - ... allows to identify the some “words” spoken by different people
- Morphology
  - How words are formed
- Syntax
  - What are the rules of words combinations
- Semantics
  - Meaning of words/sentences
- Pragmatics
  - How the context influences the meaning

# Linguistics Basics

- Phoneme
  - Unit of sound
- Grapheme
  - Unit of writing system
  - Realisations are called glyphs (surface form)

Note: There is no strict 1:1 mapping between phonemes and graphemes.

**Semantics of words**

## Distributional Hypothesis

First described by Harris in 1954, which states that words which tend to occur together are semantically related. Firth describes this intuition as “a word is characterised by the company it keeps”.

## Strong Contextual Hypothesis

Proposed by Miller and Charles in 1991, says that the more similar the contexts of words the more semantically related the words are.

Note: Linguists also use the term context to refer to situational or social context (pragmatics).

# Linguistics Basics

**Syntactic Parsing**

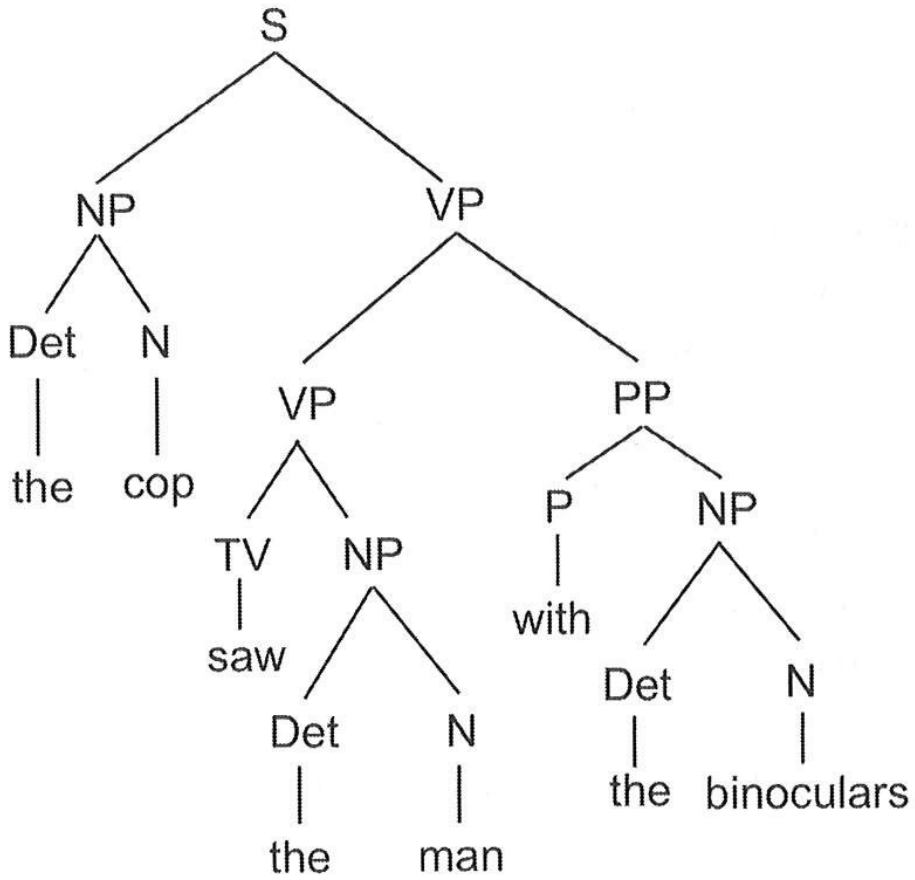
- Transform a sentence into a tree representation
- ... which reflects the grammatical structure of the sentence

## Example sentence

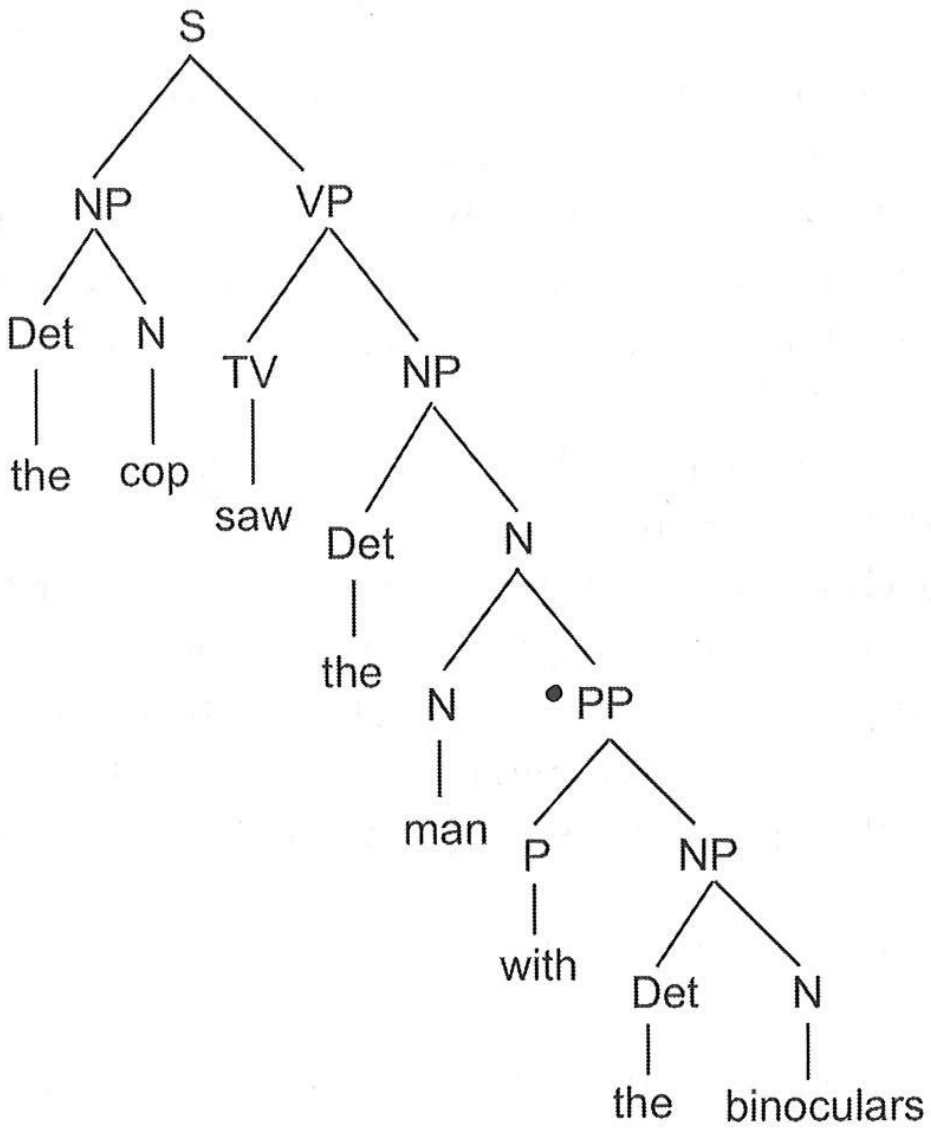
`The cop saw the man with the binoculars.`

Taken from: Bergmann, A., Hall, K. C., & Ross, S. M. (2007). Language files: Materials for an introduction to language and linguistics. Ohio State University Press.

# Linguistics Basics



# Linguistics Basics



# Linguistics Basics

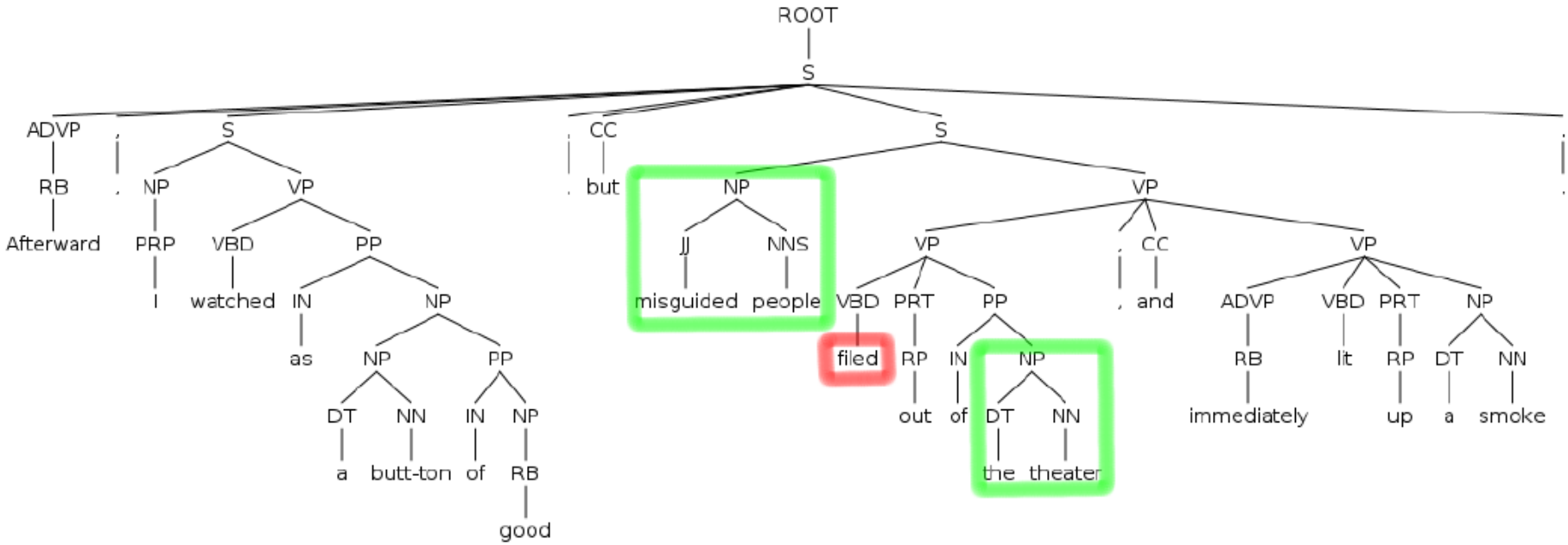
**Dependency Parsing**

- Transform a sentence into a graph representation
- ... where each vertex is a word
- ... and each edge represents a grammatical relationship

## Example sentence

```
Afterward , I watched as a butt-ton of good , but misguided
people filed out of the theater , and immediately lit up a
smoke .
```

# Linguistics Basics



# Linguistics Basics

| relation | gov ▲ | dep |
|---|---|---|
| **pobj** | as-5 | butt-ton-7 |
| **det** | butt-ton-7 | a-6 |
| **prep** | butt-ton-7 | of-8 |
| **nsubj** | filed-14 | people-13 |
| **prt** | filed-14 | out-15 |
| **prep** | filed-14 | of-16 |
| **cc** | filed-14 | and-20 |
| **conj** | filed-14 | lit-22 |
| **advmod** | lit-22 | immediat... |
| **prt** | lit-22 | up-23 |
| **dobj** | lit-22 | smoke-25 |
| **pobj** | of-16 | theater-18 |
| **pobj** | of-8 | good-9 |
| **amod** | people-13 | misguide... |
| **det** | smoke-25 | a-24 |
| **det** | theater-18 | the-17 |
| **advmod** | watched-4 | Afterwar... |
| **nsubj** | watched-4 | I-3 |
| **prep** | watched-4 | as-5 |
| **cc** | watched-4 | but-11 |
| **conj** | watched-4 | filed-14 |

# Information Extraction

## Knowledge Base Population

# Information Extraction

**Traditional Information Extraction**

- Given a block of text (usually sentence)
- ... identify all named entities from a predefined list of entity types
  - 4 categories: person, organisation, location and miscellaneous
- Often tackled using sequence classification algorithms (plus external resources)
  - e.g. Hidden Markov Models, Conditional Random Fields

Sang, E. F. T. K., & De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. CONLL '03 Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, 4, 142–147.

# Information Extraction

## Wikification

- The list of extracted terms to expanded to all Wikipedia articles
- i.e. each Wikipedia is treated as an entity
- Also called entity linking

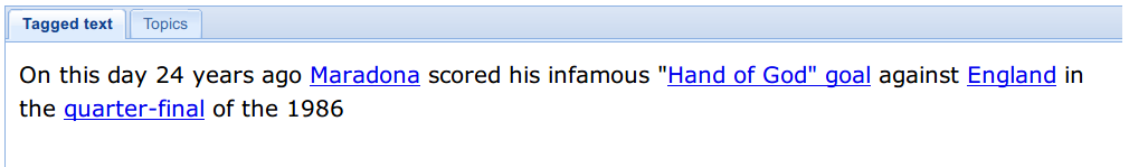



Figure: Screenshot of the TAGME system: `https://tagme.d4science.org/tagme/`

Mendes, P. N., Jakob, M., García-silva, A., & Bizer, C. (2011). DBpedia Spotlight : Shedding Light on the Web of Documents. Proceedings of the 7th International Conference on Semantic Systems (I-Semantics)., 95, 1–8.

# Information Extraction

## Open Information Extraction

- No restriction of the type of information being extracted
- The syntactic structure conveys the hints what to extract
- ... sometimes also called fact extraction

### Example

"In May 2010, the principal opposition parties boycotted the polls after accusations of vote-rigging."
$\rightarrow$
("the principal opposition parties", "boycotted", "the polls")
("the principal opposition parties", "boycotted the polls in", "May 2010")
("the principal opposition parties", "boycotted the polls after", "accusations of vote-rigging")

Gamallo, P. (2014). An overview of open information extraction. OpenAccess Series in Informatics, 38, 13–16.

# Information Extraction

**Opinion Mining / Sentiment Analysis**

- Given a text (often a review)
- ... provide a classification into
  - `Positive` or `negative`
  - ... other classification schemes also common
- Often computed using trigger words/phrases and machine learning (classification)
  - Specialised corpora available, e.g. SentiWordNet

Liu, B. (2007). Web data mining: exploring hyperlinks, contents, and usage data. Springer Science & Business Media.

# Machine Translation

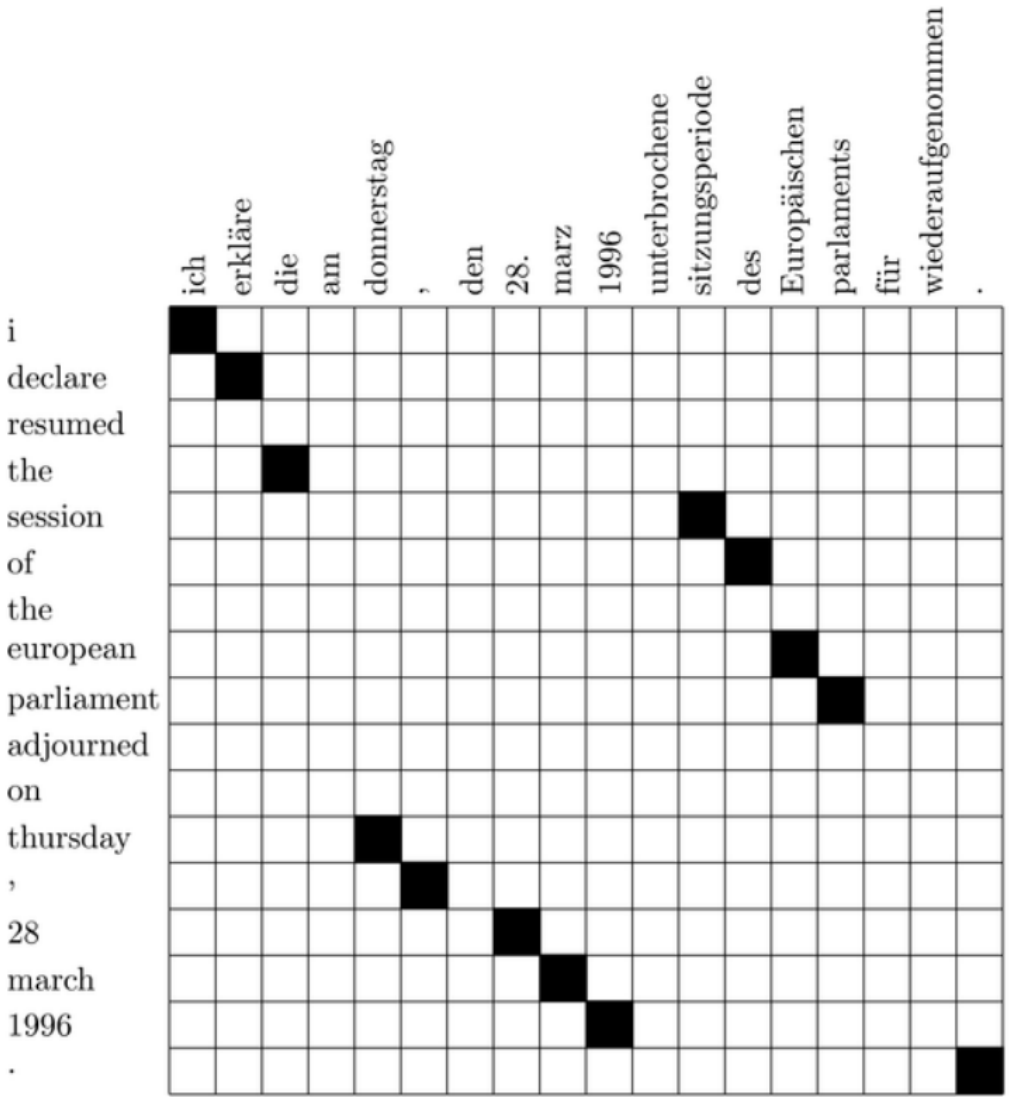
## Cross-Language Techniques

# Machine Translation

- Task: Given a piece of text in the source language, translate it into the target language
- Basic approach
  - Provide word for word translation candidates
    - ★ Based on (sentence, word) aligned corpora
  - Based on sequence information, generate a sentence in the target language (reordering)
- Production system typically use an interlingua for translation

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., ... & Dyer, C. (2007, June). Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions (pp. 177-180). Association for Computational Linguistics.

# Machine Translation



# Text Reuse

## Authorship Attribution, Plagiarism Detection

# Text Reuse

## Sentence Similarity

- There are multiple ways to express the same meaning using natural text
- $\rightarrow$ express the semantic similarity with a scalar
- Typically semantic similarity is computed using topical similarity
  - i.e. overlap of words

## Textual Entailment

- Related concept
  - For any two sentences X and Y, X entails Y, if whenever X is true, Y is true as well.
  - “Joe is a oran utan.” $\rightarrow$ “Joe is a mammal.”

## Text Reuse

- Two documents sharing “similar” fragments of text
- Often tackled by building a reference corpus
- ... by indexing the reference document via sliding windows
- The document in question is also processed via sliding windows
- ... which are compared with the reference corpus
  - The features use to compare the windows are often topical features, i.e. the words.
- Any longer sequences of matching windows indicate text reuse

Muhr, M., Kern, R., Zechner, M., & Granitzer, M. (2010). External and Intrinsic Plagiarism Detection using a Cross-Lingual Retrieval and Segmentation. In 2nd International Competition on Plagiarism Detection.

# Authorship Attribution

## Stylometric Features

| Feature | Description |
|---|---|
| alpha-chars-ratio | the fraction of total characters in the paragraph which are letters |
| digit-chars-ratio | the fraction of total characters in the paragraph which are digits |
| upper-chars-ratio | the fraction of total characters in the paragraph which are upper-case |
| white-chars-ratio | the fraction of total characters in the paragraph which are whitespace characters |
| type-token-ratio | ratio between the size of the vocabulary (i.e., the number of *different* words) and the total number of words |
| hapax-legomena | the number of words occurring once |
| hapax-dislegomena | the number of words occurring twice |
| yules-k | a vocabulary richness measure defined by Yule |
| simpsons-d | a vocabulary richness measure defined by Simpson |
| brunets-w | a vocabulary richness measure defined by Brunet |
| sichels-s | a vocabulary richness measure defined by Sichel |
| honores-h | a vocabulary richness measure defined by Honore |
| average-word-length | average length of words in characters |
| average-sentence-char-length | average length of sentences in characters |
| average-sentence-word-length | average length of sentences in words |

Rexha, A., Klampfl, S., Kröll, M., & Kern, R. (2015). Towards Authorship Attribution for Bibliometrics using Stylometric Features. In Mining Scientific Papers: Computational Linguistics and Bibliometrics.

# Advanced Topics

## Word Embeddings, LDA, ...

# Advanced Topics

**Latent Semantic Analysis** [1]

- Idea: Apply thin SVD on the document-term matrix
  - Where the SVD is limited to the $k$ most important singular values
- Requires as input:
  - Document/term matrix
  - Fixed number of topics
- Provides:
  - Mapping of document to a (dense) lower-dimensional representation
- Probabilistic version: pLSA [2]

[1] Landauer, T. K., Dutnais, S. T., Anderson, R., Carroll, D., Fbltz, P., Pumas, G., ... Streeter, L. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge.
[2] Hofmann, T. (1999). Probabilistic latent semantic indexing.

## Latent Dirichlet Allocation (LDA)

- Requires as input:
  - Document/term matrix
  - Fixed number of topics
- Provides:
  - Mapping of document to topics (as vector of probabilities)
  - Mapping of terms to topics (as vector of probabilities)
- Can be seen as fuzzy co-clustering

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation.

# Advanced Topics

| Topic #247 | Topic #5 | Topic #43 | Topic #56 |
|---|---|---|---|
| drugs (.096) | red (.202) | mind (.081) | doctor (.074) |
| drug (.060) | blue (.099) | thought (.066) | dr. (.063) |
| medicine (.027) | green (.096) | remember (.063) | patient (.061) |
| effects (.023) | yellow (.073) | memory (.037) | hospital (.049) |
| body (.019) | white (.048) | thinking (.030) | care (.046) |

Figure: Example of LDA build using the TASA corpus

Steyvers, M., & Griffiths, T. (2007). Probabilistic Topic Models. Handbook of Latent Semantic Analysis.

## Word Embeddings

- Main idea: replace a single word by a representation
  - where similar words are close to each other
  - ... see “distributional hypothesis”
- Each word is represented by a (dense) vector (or 50-300 dimensions)
  - Ideal as input to a neural network
- Best known realisations
  - Word2Vec [1]
  - GloVe [2]

[1] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv Preprint arXiv:1301.3781.
[2] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In Empirical Methods in Natural Language Processing (EMNLP) (pp. 1532–1543).

# Advanced Topics

## Word Sense Disambiguation

- Given a polysemous word (or homonyms)
  - The same word, but different senses (or meaning)
  - e.g. bank (the financial institute) vs. bank (side of the river) vs. bank (building in with a financial instance is located)
- Basic approach
  - Exploit the distributional hypothesis
  - Different context (surrounding words) imply a different meaning

Kern, R., Muhr, M., & Granitzer, M. (2010). KCDC: Word Sense Induction by Using Grammatical Dependencies and Sentence Phrase Structure. In Proceedings of SemEval-2. Uppsala, Sweden, ACL.

# Tools

## Machine Learning, NLP, ...

# Tools

**Python**

- Read/manage data: pandas
- Machine learning: scikit-learn
- NLP
  - NLTK
  - Spacy
    - `http://www.spacy.io/`
  - fastText
    - `https://research.fb.com/fasttext/`
- Coding: Jupyter notebooks
- Point'n'click: Orange

# Tools

**Deep Learning**

- Python: tensorflow + keras
  - NLP based on tensorflow: SyntaxNet
- Java: deeplearning4j
- R: H2O
- Word embeddings: Word2Vec, GloVe

# Tools

**Weka**

- Machine learning library for Java
- Extensive array of algorithms available, plus many 3rd party packages, e.g. time series prediction
- Can be used as application or as library
- Extensions for multi-label problems
  - Meka, Mulan
- Extension for streaming data
  - Moa

# Tools

**Stanford CoreNLP**

- NLP library for Java
- Heavily used in research

**Mallet**

- NLP library for Java
- Implements HMMs and CRFs (plus many more)
- ... for sequence classification

**Gate**

- Application & library for text mining
- Good starting point for rule based extractions

# Tools

**Sensium**

- SaaS for basic NLP tasks
- `https://www.sensium.io/`

**CODE Annotator**

- Web-based research prototype for manual annotations and sequence learning
- `http://code-demo.know-center.tugraz.at`

# The End

Next: Time Series