Text Mining (1/2)

Lecture 16

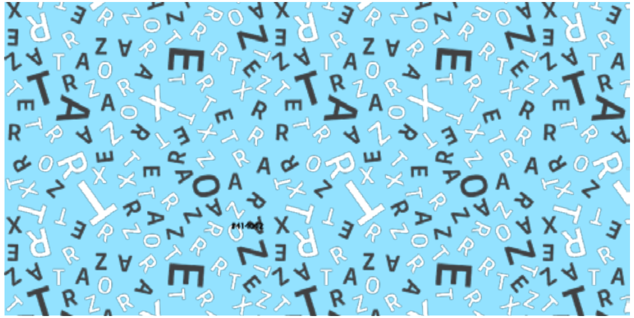
IDS-L16

Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Outline of Today’s Lecture

- **Basic definitions and preliminaries**
- **Preprocessing text**
- **Modeling text**



Moving from structured tabular data to unstructured text (supervised and unsupervised problems).

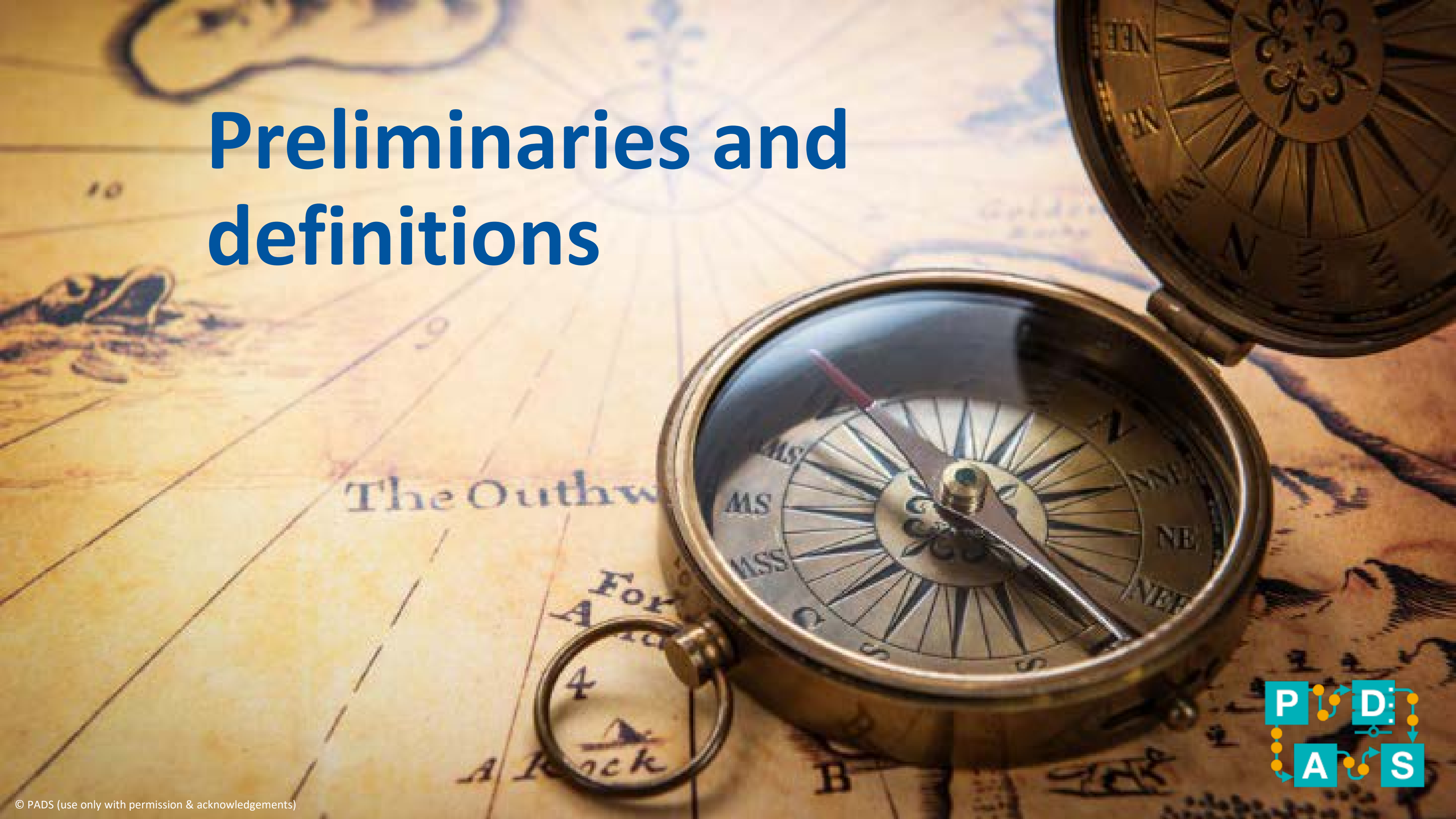
Chair of Process and Data Science

Material (only as background information or if you want to dive deeper)

- **Jurafsky, Martin, *“Speech and Language Processing”*, chapters 4 through 8: https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf**
- **Manning, Ragavan, Schütze, “Introduction to Information Retrieval”, chapters 6 and 13 through 18: https://nlp.stanford.edu/IR-book/pdf/irbookprint.pdf**

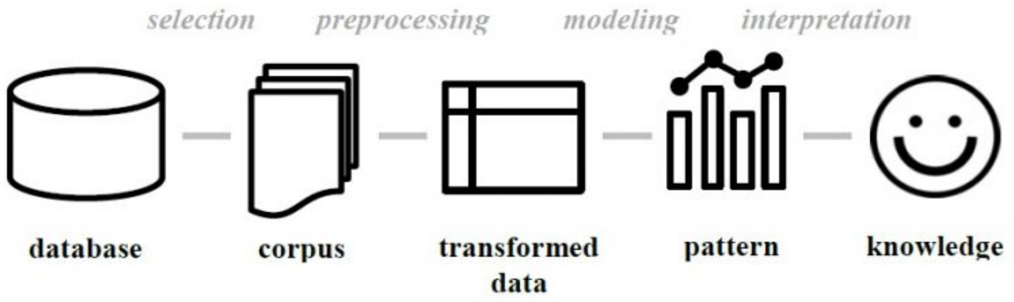
Chair of Process and Data Science

Preliminaries and definitions



Text Mining

Text Mining: the extraction of (structured) knowledge from (unstructured) text.



Text Mining Pipeline

From https://hecc.ubc.ca/

Text Mining: nomenclature

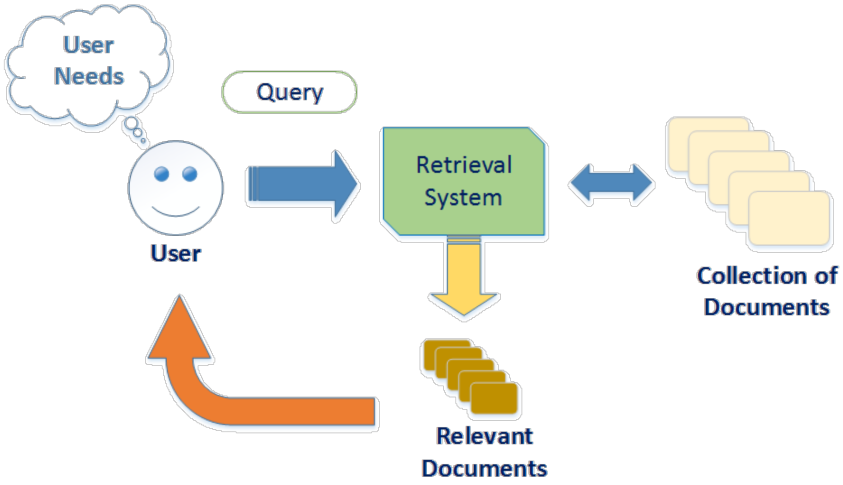
“Text Mining” is a very broad concept with “fuzzy edges”. It also relates to many other domains (e.g. linguistics & information retrieval).

From https://psb.stanford.edu/

Chair of Process and Data Science

Information Retrieval (IR)

- Typically considered as part of Text Mining, although older.
- Consists of obtaining relevant information from a large collection of text through querying the database of an information system.



From http://ir.cs.ui.ac.id/

Natural Language Processing (NLP)

- Sometimes used as a synonym for Text Mining.
- More properly, it indicates the subfield of Text Mining that applies Machine Learning techniques to extract meaning from text (as intended by a human).
- Computational Linguistics is often used as a synonym for NLP.

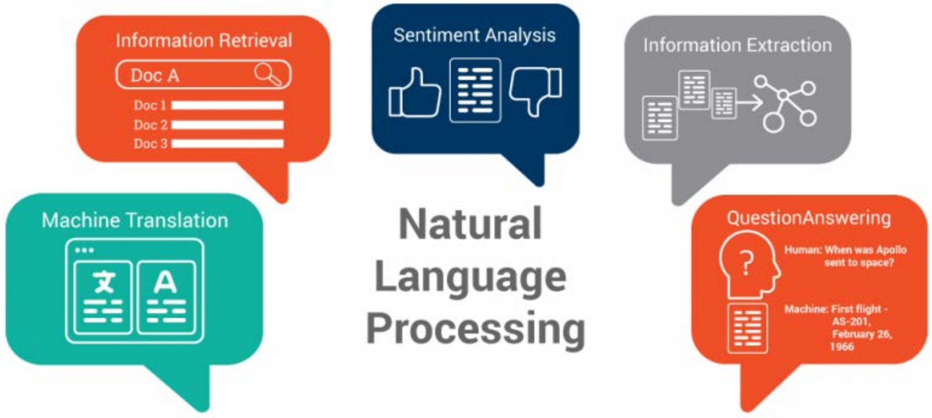
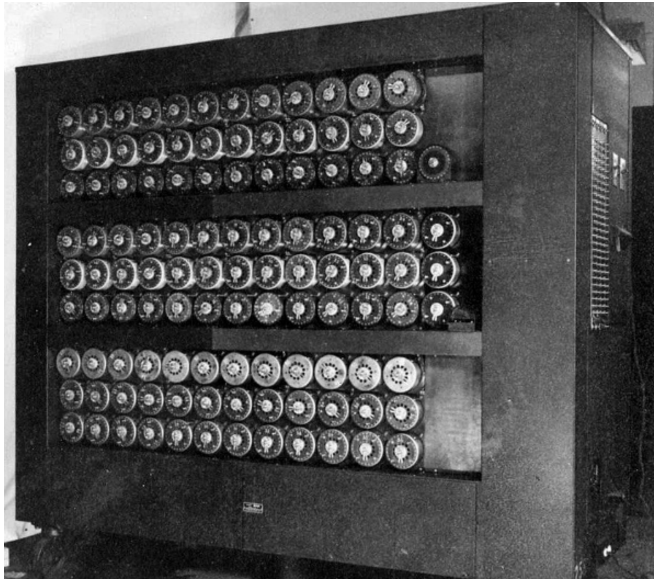


Image by Srishti Sawla

Text Mining

- “Text Mining” sounds cutting edge, but it has a long history.
- Modern Text Mining emerged in the ‘90s, but it is based on way older concepts – especially Information Retrieval.

The “Bombe” machine, used to decode Enigma ciphers (1940). Part of the decryption exploited the statistical analysis of word frequency. Exploiting for example the occurrence of the letter “e” in English (most frequent).



Chair of Process and Data Science

Text Mining applications

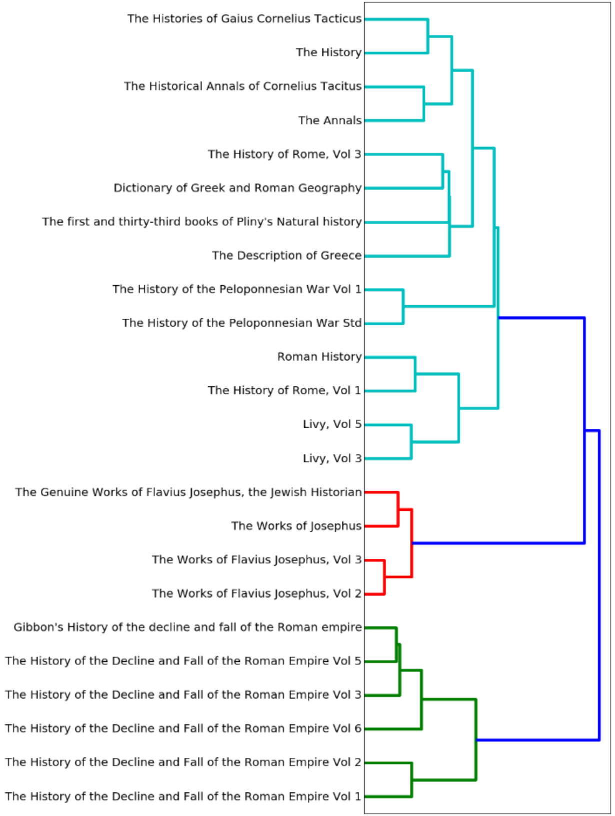
Text Mining, intended as a broad concept including IR and NLP, has a lot of different applications!

- **Some are unsupervised (e.g. Document Clustering)**
- **Some supervised (e.g. Sentiment Analysis)**
- **Some are language-dependent (e.g. Machine Translation)**
- **Some are language-independent (e.g. Keyword Extraction)**

A small set of examples is given next …

Document Clustering

Document Clustering: grouping together documents based on text, topic and content similarity.



From https://github.com/utkuozbulak/

Document Classification

Document classification: predict a specific label for a document based on the word content. As regular classification it can be binary or multilabel.

From https://hackernoon.com/

Named Entity Recognition

Named Entity Recognition (NER): recognizing named entities in the text and labeling them with a type (“person”, “location”, etc.) based on contextual information.

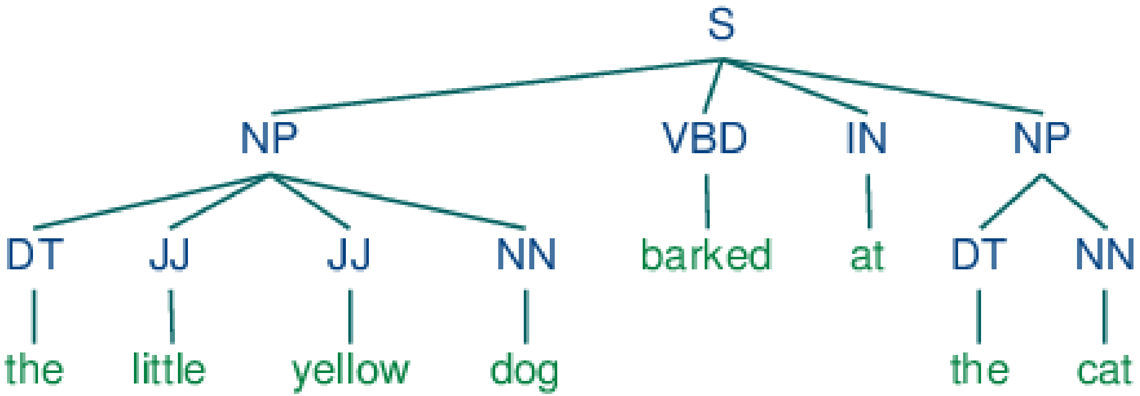
"There was nothing about this storm that was as expected," said Jeff Masters, a meteorologist and founder of Weather Underground. "Irma could have been so much worse. If it had traveled 20 miles north of the coast of Cuba, you'd have been looking at a (Category) 5 instead of a (Category) 3."

Person | Organization | Location

From https://blog.paralleldots.com/

Part Of Speech Tagging

Part-Of-Speech (POS) Tagging: labeling a word with the corresponding part of speech (noun, verb, adjective, etc.)



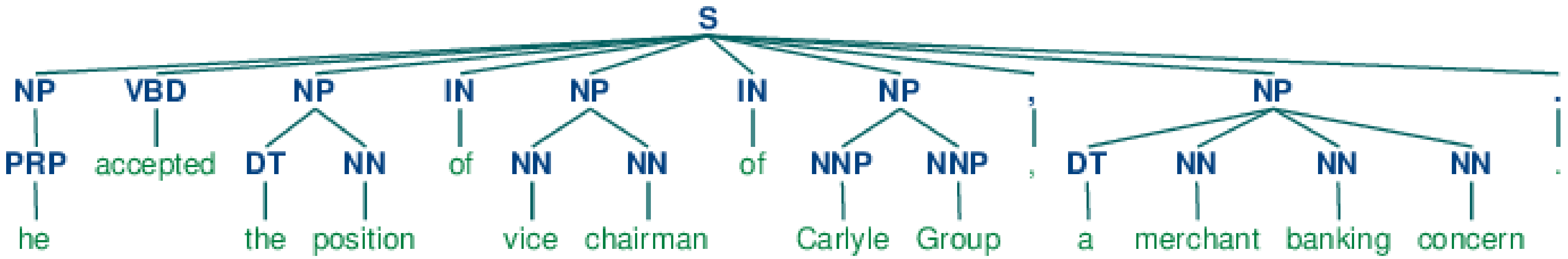
Tag	Description	Tag	Description
CC	Coordinating conjunction	PRP$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determinant	RBR	Adverb, comparative
EX	Existential *there*	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition or subordinating conjunction	SYM	Symbol
JJ	Adjective	TO	to
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBN	Verb, past participle
NN	Noun, singular or mass	VBG	Verb, gerund or present participle
NNS	Noun, plural	VBP	Verb, non-3rd person singular present
NNP	Proper noun, singular	VBZ	Verb, 3rd person singular present
NNPS	Proper noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoun
POS	Possessive ending	WP$	Possessive wh-pronoun
PRP	Personal pronoun	WRB	Wh-adverb

Parsing tree with part of speech labels. The words are labeled as verbs (VBD), nouns (NN) adjectives (JJ) and so on. The whole sentences is split in parts (here noun phrases, NP) before the tagging.

From http://www.nltk.org/

Chair of Process and Data Science

Part Of Speech Tagging



Tag	Description	Tag	Description
CC	Coordinating conjunction	PRP$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determinant	RBR	Adverb, comparative
EX	Existential *there*	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition or subordinating conjunction	SYM	Symbol
JJ	Adjective	TO	to
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form

LS	List item marker	VBD	Verb, past tense
MD	Modal	VBN	Verb, past participle
NN	Noun, singular or mass	VBG	Verb, gerund or present participle
NNS	Noun, plural	VBP	Verb, non-3rd person singular present
NNP	Proper noun, singular	VBZ	Verb, 3rd person singular present
NNPS	Proper noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoun
POS	Possessive ending	WP$	Possessive wh-pronoun
PRP	Personal pronoun	WRB	Wh-adverb

From "Parts-of-Speech Tagger Errors Do Not Necessarily Degrade Accuracy in Extracting Information from Biomedical Text" by Ling, Lefevre, and Nicholas

Chair of Process and Data Science

Coreference Resolution

Coreference Resolution: identification of words and expressions that refer to another object in a sentence or in a piece of text.

"I voted for Nader because he was most aligned with my values," she said.

From https://nlp.stanford.edu/

Sentiment Analysis

Sentiment Analysis: identification of positive/negative attitude of the writer from text, via the recognition of emotions, opinions, or mood.

Emotional Criteria	Example topic sentences
Trust	"Forbes Article Predicts Bitcoin Value will "Explode"" / "Good news for the Bitcoin" / "Don't panic, China is NOT banning bitcoin"
Fear	"Mining cartel attack" / "OMG! What has Satoshi created? He has opened Pandora's box" / "We are victims of our own success"
Surprise	"Whatever happened to the Bitcoin Police?" / "I think the rapture happened.....?" / "Blockchain.info "firstbits" changing/disappearing?!"

doi:10.1371/journal.pone.0132944.t011

From https://nlp.stanford.edu/

Keyword Extraction

Keyword Extraction: automatic identification of important terms inside a piece of text.

"For search managers, developers & data scientists finding ways to innovate"

~~For~~ **search managers**~~,~~ **developers** ~~&~~ **data scientists finding** ~~ways to~~ **innovate**

By Wolters Kluwer

Chair of Process and Data Science

Machine Translation

Machine Translation: automatic translation of text from one language to another.

From https://www.capitatranslationinterpreting.com

Machine Translation

In recent years, data science emerged as a new and important discipline. It can be viewed as an amalgamation of classical disciplines like statistics, data mining, databases, and distributed systems. Existing approaches need to be combined to turn abundantly available data into value for individuals, organizations, and society. Moreover, new challenges have emerged, not just in terms of size (“Big Data”) but also in terms of the questions to be answered. The course aims to provide a comprehensive overview of data science and expose students to real-life data sets and tools.

In den letzten Jahren hat sich die Datenwissenschaft zu einer neuen und wichtigen Disziplin entwickelt. Es kann als ein Zusammenschluss klassischer Disziplinen wie Statistik, Data Mining, Datenbanken und verteilte Systeme betrachtet werden. Bestehende Ansätze müssen kombiniert werden, um reichlich vorhandene Daten in Wert für Einzelpersonen, Organisationen und die Gesellschaft zu verwandeln. Darüber hinaus haben sich neue Herausforderungen ergeben, nicht nur in Bezug auf die Größe ("Big Data"), sondern auch in Bezug auf die zu beantwortenden Fragen. Ziel des Kurses ist es, einen umfassenden Überblick über die Datenwissenschaft zu geben und die Studierenden mit realen Datensätzen und Tools vertraut zu machen.

In de afgelopen jaren is de datawetenschap als een nieuw en belangrijk vakgebied naar voren gekomen. Het kan worden gezien als een samensmelting van klassieke disciplines zoals statistiek, datamining, databases en gedistribueerde systemen. Bestaande benaderingen moeten worden gecombineerd om overvloedig beschikbare data om te zetten in waarde voor individuen, organisaties en de maatschappij. Bovendien zijn er nieuwe uitdagingen ontstaan, niet alleen in termen van grootte ("Big Data") maar ook in termen van de te beantwoorden vragen. Het doel van de cursus is om een uitgebreid overzicht te geven van de datawetenschap en de studenten bloot te stellen aan real-life datasets en tools.

Translated using DeepL

Text mining

Each of these topics could have a course on its own!

We will focus in data representation and some application examples.

Text mining

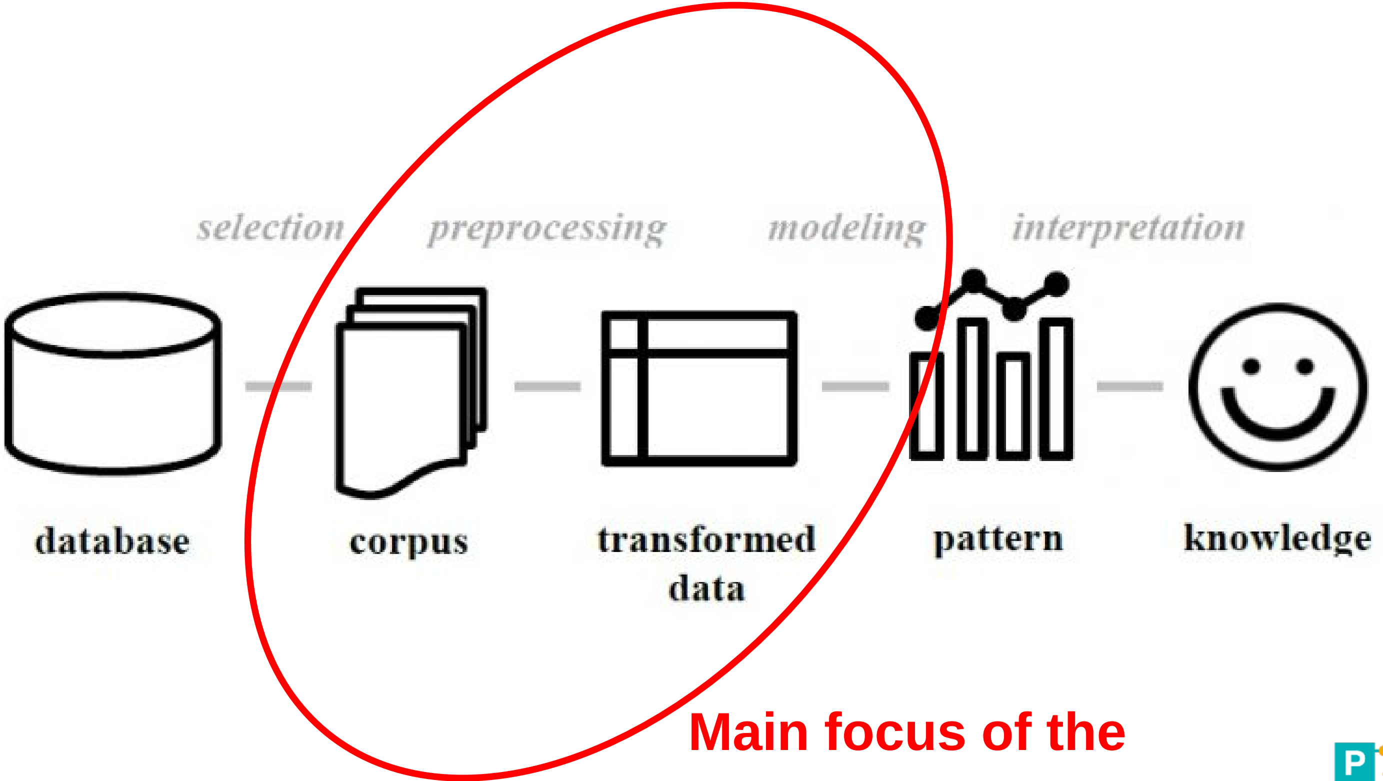
It is crucial to preprocess and model text as data (numbers) in a proper way.

Once you do that, you can apply other techniques to it

- **Supervised/unsupervised learning**
- **Statistical methods**

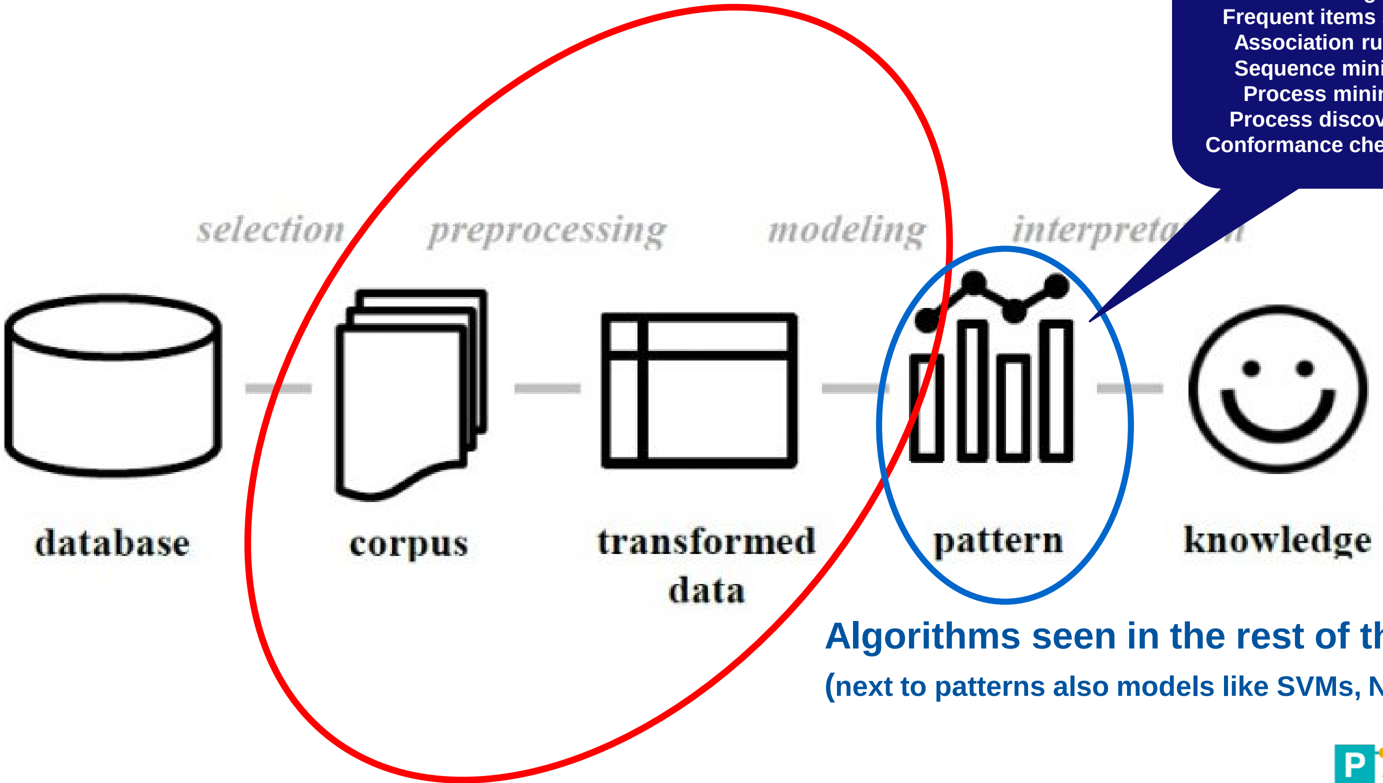
Chair of Process and Data Science

Text mining



Main focus of the Text Mining lectures

Text mining



Decision trees
Regression
Support vector machines
Neural networks
Clustering
Frequent items sets
Association rules
Sequence mining
Process mining
Process discovery
Conformance checking

Algorithms seen in the rest of the course

(next to patterns also models like SVMs, NNs, DTs, etc.)

Chair of Process and Data Science

Reminder: Data science is like a making cocktail. Mix the proper ingredients in the right way!

- Python
- Visualization
- Decision trees
- Regression
- Support vector machines
- Neural networks
- Evaluation
- Clustering
- Frequent items sets
- Association rules
- Sequence mining
- Process mining
- Process discovery
- Conformance checking
- Text mining
- Preprocessing
- Visual analytics
- Encryption
- Anonymization
- Big data infra
- Distribution

Text Preprocessing

Structuring text

The first challenge in Text Mining is to go from unstructured data (text) to structured data (ideally numbers).

Text is extremely unstructured!

- **Do I consider sentences or words?**
- **The "length" of the single unit of information is variable!**
- **What are the "features" in text?**

Chair of Process and Data Science

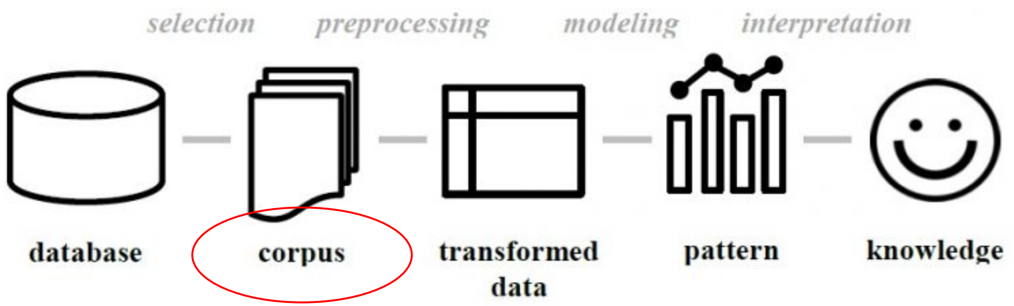
Structuring text

In order to make text "processable" it has to go through multiple steps of transformation.

Many of these steps are common to all or almost all applications. Some, instead, are specific.

Structuring text

Let's take a look at the pipeline again! The first step is extracting a corpus from some database.



What is a corpus?

Corpus

A corpus (pl. corpora) is a collection of pieces of text.

It is the first level of "structuredness" of text:

- The text is divided in pieces that are consistent
  e.g.: words, sentences, paragraphs, tweets, forum posts

Corpus

- **Often available with meta-information about the text**
  e.g. Forum posts can be labeled with the author or the thread title

- **Usually in a single language, sometimes multi-lingual.**

- **The pieces of text in the corpus are often called documents (regardless of nature and size).**

Chair of Process and Data Science

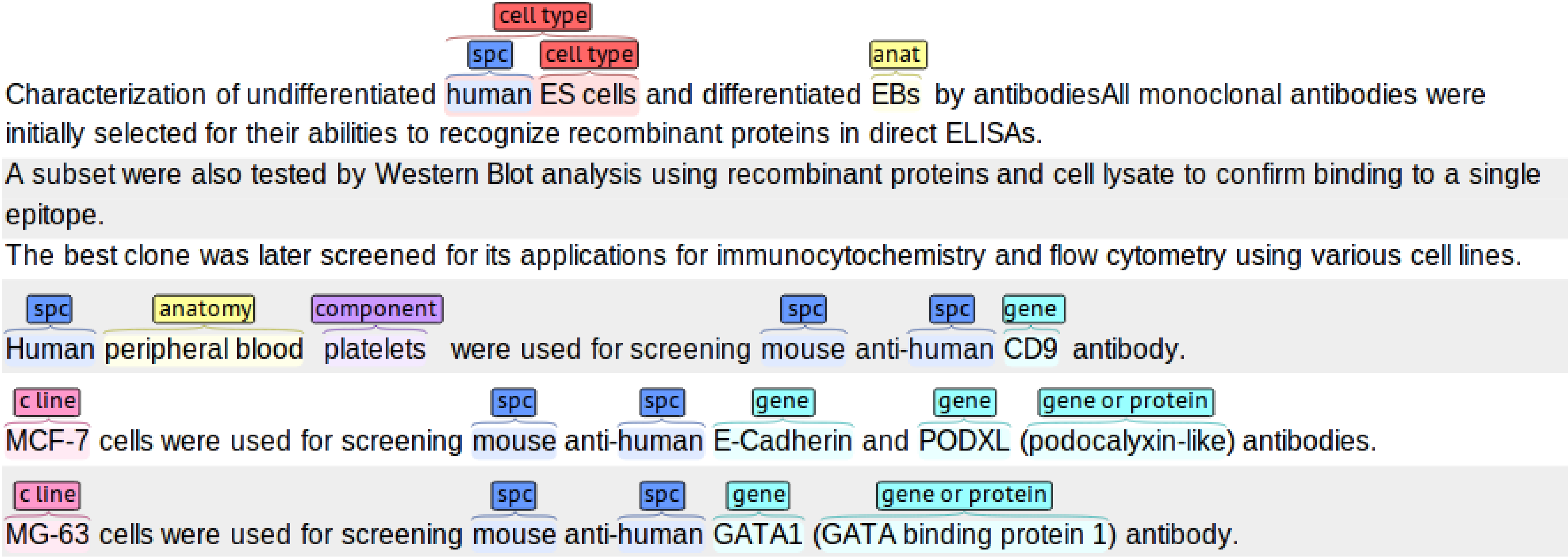
Annotated corpus

An annotated corpus is a corpus in which the “units” or fractions of “units” of text have been annotated with additional information in order to work as a training set for a specific application.

- Corpora are usually annotated by hand
  - You need to compare the accuracy of your algorithms with human accuracy!

- Innovations in Text Mining are possible thanks to people that annotated by hand tens of millions of words back in the ’80s!

Annotated corpus



Corpus annotated for domain-specific (medical) Named Entity Recognition.

From http://brat.nlplab.org/

Chair of Process and Data Science

Annotated corpus

Word	Tag 1 (Main)	Tag2	Tag3
مبايعة	N	ON	F
تعلم	V	PV	F
نور	N	ON	M

Main POS Noun	Definition
Original Noun	The word that describes an event that is not connected with any time period.
Agent Noun	The noun that is used to describe the person how has done an action or the subject of the verb.
Patient Noun	The noun that is used to describe the object of the verb.
Adjective Noun	The adjective noun in Arabic language is actually similar to it in English language; it is the noun that is used to describe something.
Superlative Noun	The superlative is the noun that used to compare two nouns in a specific adjective.

Main Part of Speech Tags	Tag (Main)
Noun	N
Verb	V
Particle	P

Main POS Verb	Definition
Perfect/ Past Verb	Describes an action that happened in the past.
Imperfect/ Progress Verb	Describes an action that is happening in the present time.
Imperative Verb	Describes an order to do some action.

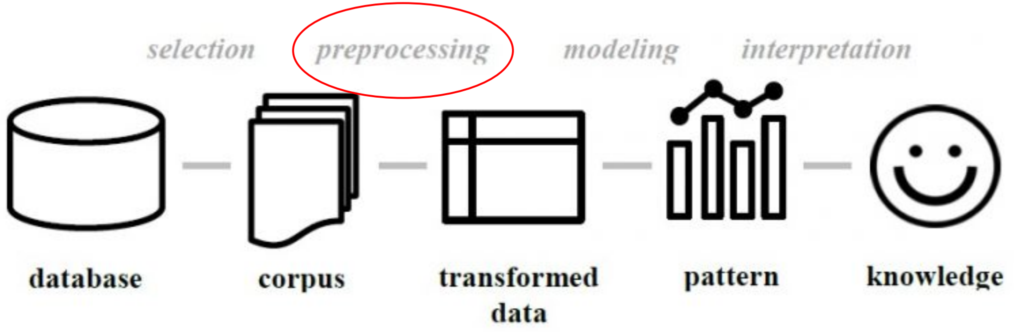
Corpus annotated for Part-Of-Speech Tagging in Arabic.

Tags refer to noun/verb, type of noun/verb, male/female.

From Abumalloh et al., “Building Arabic Corpus Applied to Part-of-Speech Tagging”

Structuring text

Once obtained/generated a corpus (annotated or not), the text has to go through a preprocessing phase.



Chair of Process and Data Science

Text Preprocessing

Text in a corpus has to go through different preprocessing steps in order to be then modeled and used in analysis:

1. **Tokenization**
2. **Stopword removal**
3. **Token Normalization: Stemming/Lemmatization**

Tokenization

Tokenization of text means splitting it in smaller units called "tokens".

- **Usually into "words" (this is what we are going to do)**

  But could also be characters, ideograms, phonemes, syllables, sentences, phrases, clauses, and more.

Word Tokenization

Wow, that’s really easy! Just split on spaces, right?

Chair of Process and Data Science

Word Tokenization

Wow, that’s really easy! Just split on spaces, right?

s = “He’s been talking to Bill de Blasio, the 109th New York City mayor.”

Chair of Process and Data Science

Word Tokenization

Wow, that's really easy! Just split on spaces, right?

s = "He's been talking to Bill de Blasio, the 109th New York City mayor."

```
s.split(' ')
```

['He's', 'been', 'talking' 'to' 'Bill' 'de' 'Blasio,', 'the', '109th', 'New', 'York', 'City', 'mayor.']

Chair of Process and Data Science

Word Tokenization

Wow, that’s really easy! Just split on spaces, right?

s = “He’s been talking to Bill de Blasio, the 109th New York City mayor.”

Hmm, does this actually mean something?
Should I have a token for each number?

Wait, this one token should actually be 2. Or not?

Should these be 3 tokens or one?

[‘He’s’, ‘been’, ‘talking’ ‘to’ ‘Bill’ ‘de’ ‘Blasio,’, ‘the’, ‘109th’, ‘New’, ‘York’, ‘City’, ‘mayor.’]

Should these ones here be 3 tokens or one?

These should not end up in there, right…?

Chair of Process and Data Science

Word Tokenization

Some other languages are harder than English.

“Lebensversicherungsgesellschaftsangestellter”
(44 letters)

Is that one token…? It depends!

Word Tokenization

Other examples

- German:
  “Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz” (63)
  (law delegating beef label monitoring)
- Dutch:
  “bestuurdersaansprakelijkheidsverzekering” (40)
  (drivers' liability insurance)

Are these tokens …? It depends!

Chair of Process and Data Science

Word Tokenization

Western languages are relatively easy to tokenize.

(some languages have no spaces between words)

我喜欢新西兰花	Unsegmented Chinese sentence
我　喜欢　新西兰　花	*I like New Zealand flowers*
我　喜欢　新　西兰花	*I like fresh broccoli*

From http://what-when-how.com/

Word Tokenization

Tokenization is not trivial, not even into words!

- **There is specific software to tokenize.**
- **Quite often tokenization need to be designed ad-hoc for the task at hand.**
- **It is application-dependent.**

Chair of Process and Data Science

Stopword Removal

Stopword removal refers to removing from the text words that do not add meaning.

Usually, common parts of speech are to be removed, such as ‘the’, ‘as’, ‘in’, ‘me’, ‘you’, ‘which’, ‘on’, etc.

Chair of Process and Data Science

Stopword Removal

A stoplist will commonly contain:

- **'Be' and 'have' verbs: 'is', 'are', 'am', 'was', 'have', 'has', etc.**
- **Articles: 'the', 'a', 'an', etc.**
- **Auxiliary verbs: 'will', 'should', 'would', 'shall', 'must', etc.**
- **Prepositions: 'in', 'to', 'from', 'through', 'of', 'by', 'on', etc.**
- **'who', 'what', 'which', 'where', 'when', 'how', 'why', etc.**

Why do we remove stopwords?

Stopword Removal

“The cats, which were seven, started to climb the tree.”

“The Saxons, which were outnumbered, started to prepare the siege.”

What happens if I try to compare the similarities between sentences counting the number of words in common?

Stopword Removal

“The cats, which were seven, started to climb the tree.”

“The Saxons, which were outnumbered, started to prepare the siege.”

What happens if I try to compare the similarities between sentences counting the number of words in common?

6/10 words are the same! These two sentences should be very similar!

Chair of Process and Data Science

Stopword Removal

"The cats, which were seven, started to climb the tree."

"The Saxons, which were outnumbered, started to prepare the siege."

Five of the six common words are stopwords …

Stopword Removal

"~~The~~ cats, ~~which were~~ seven, started ~~to~~ climb ~~the~~ tree."

"~~The~~ Saxons, ~~which were~~ outnumbered, started ~~to~~ prepare ~~the~~ siege."

Removing the stopwords shows that the two sentences are very different despite the initial overlap.

Chair of Process and Data Science

Stopword Removal

Removing stopword is very useful in every Text Mining algorithm that takes into account the frequency of words in the documents.

Without stopword removal statistical methods on text would not work!

Stopword Removal

A list of stopwords is called a stoplist. Stoplists are language-dependent.

Stopword Removal

Again, designing a good stoplist is not trivial!

- **Are words like “want” or “between” stopwords?**
  - **It depends on the application!**
  - **Stoplists can be conservative (small) or aggressive (long).**
  - **There is no definitive answer on what to include.**

Stopword Removal

Again, designing a good stoplist is not trivial!

- **Stoplists can be domain dependent!**
  - **If I am working with a corpus extracted from curricula, words like "experience" or "school" can be stopwords.**
  - **In a healthcare-related corpus, words like "patient" or "hospital" can be stopwords.**
  - **Later in the lecture we will see a method to find candidate stopwords (based on frequency).**

Chair of Process and Data Science

Stopword Removal

Be careful about what you remove!

"The Who's seventh studio album is titled "The Who by Numbers"."

It is very hard to remove stopwords from sentences like this one!

Stopword Removal

Be careful about what you remove!

Assume:
- **'Be' and 'have' verbs: 'is', 'are', 'am', 'was', 'have', 'has', etc.**
- **Articles: 'the', 'a', 'an', etc.**
- **Auxiliary verbs: 'will', 'should', 'would', 'shall', 'must', etc.**
- **Prepositions: 'in', 'to', 'from', 'through', 'of', 'by', 'on', etc.**
- **'who', 'what', 'which', 'where', 'when', 'how', 'why', etc.**

~~The Who's~~ seventh studio album ~~is~~ titled "~~The Who by~~ Numbers"

⇒ seventh studio album titled numbers

Chair of Process and Data Science

Stemming

Let’s look at some other examples of similarities between sentences:

Chair of Process and Data Science

Stemming

Let's look at some other examples of similarities between sentences:

"The IBM 5150 was the first IBM personal computer."
"The IBM 5150 machine revolutionized home computing."

Should "computer" and "computing" count as a match?

Stemming

Stemming: chopping off suffixes of words in order to match tokens with a common root

“compute”, “computer”, “computers”, “computing”, “computational” → “comput”

- **Works *most* of the times.**
- **In many cases the root word carries the meaning.**

Stemming

In many cases you can choose where to chop.

Stemmers algorithms can be more aggressive or more conservative.

- **Aggressive stemmers will accentuate similarities between documents.**
- **Conservative stemmers will accentuate differences.**

Stemming

Popular stemmers for English:

- **Porter stemmer (conservative)**
  - https://tartarus.org/martin/PorterStemmer/index.html
- **Snowball stemmer (more aggressive)**
  - https://github.com/snowballstem

Both are widely available in a number of languages.

Lemmatization

Instead of chopping off words, I can swap them for their lemma (the basic form).

“compute”, “computer”, “computers”, “computing”, “computational” → “compute”

Way harder to implement than stemming!

Often rule-based, but every language has exceptions.

Chair of Process and Data Science

Stemming vs Lemmatization

Word	Stem	Lemma
bakery	baker	bake
bakeries	baker	bake
police	polic	police
policy	polic	policy
numerical	numer	numerical

Very different words can have the same stem

Lemmatizer are often unable to reconstruct the lemma, the default choice is to conserve the word

Both stemming and lemmatization are prone to errors!

Token Normalization

Stemming and lemmatization are the main forms of token normalization: transforming tokens to make them comparable.

Other forms of normalization:

- **Case-folding: convert everything into lowercase.**
- **Alternative spelling: “color” and “colour”.**
- **Transliterations: “Brno” and “Brünn”.**
- ä → ae, ö → oe, ü → ue, Ä → Ae, Ö → Oe, Ü → Ue, ß → ss

Text Preprocessing

“Process Discovery and Conformance Checking are part of Process Mining.”

now looks like this

[‘process’, ‘discover’, ‘conform’, ‘check’, ‘part’, ‘process’, ‘min’]

=

[‘process’^{2}, ‘discover’^{1}, ‘conform’^{1}, ‘check’^{1}, ‘part’^{1}, ‘min’^{1}]

Chair of Process and Data Science

Text Preprocessing

No amount of preprocessing can fully manage the complexity of natural language.

"Buffalo killed a buffalo in Buffalo."

A human understands such sentences, preprocessing this to use for mining purposes is extremely hard.

Homonyms

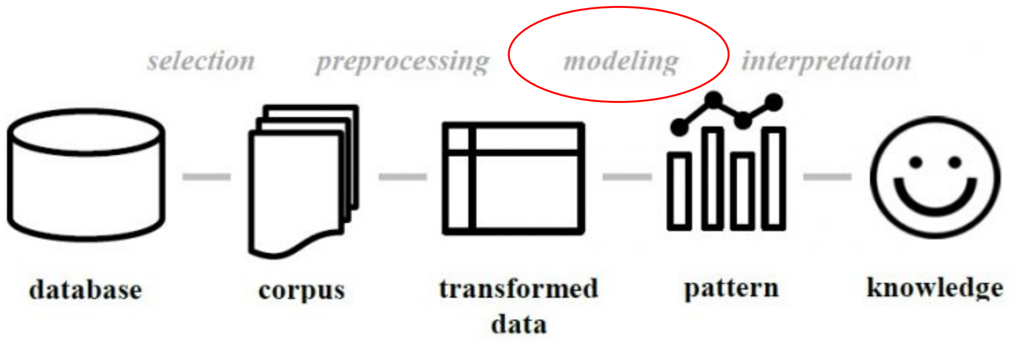
- Address - to speak to / location
- Air - oxygen / a lilting tune
- Arm - body part / division of a company
- Band - a musical group / a ring
- Bark - a tree's out layer / the sound a dog makes
- Bat - an implement used to hit a ball / a nocturnal flying mammal
- Bright - very smart or intelligent / filled with light
- Circular - taking the form of a circle / a store advertisement
- Current - up to date / flow of water
- Die - to cease living / a cube marked with numbers one through six
- Express - something done fast / to show your thoughts by using words
- Fair - equitable / beautiful
- Jag - a sharp, jutted object / a crying spree
- Kind - type / caring
- Lie - to recline / to tell a falsehood
- Match - to pair like items / a stick for making a flame
- Mean - average / not nice
- Pole - a person from Poland / a piece of metal that holds a flag
- Pound - unit of weight / to beat
- Quarry - a site for mining stone / to extract or obtain slowly
- Ream - a pile of paper / to juice a citrus fruit
- Ring - a band on a finger / something circular in shape
- Right - correct / direction opposite of left
- Rock - a genre of music / a stone
- Rose - to have gotten up / a flower
- Spring - a season / coiled metal
- Stalk - a part of a plant / to follow or harass someone
- Tender - gentle / offer of money
- Tire - to grow fatigued / a part of a wheel
- Well - in good health / a source for water in the ground

https://examples.yourdictionary.com/examples-of-homonyms.html

Text Modeling

Modelling

Now that we have obtained our preprocessed data, we can focus on building a usable model from the text.



Bag-of-Words Model

The simplest model: Bag-of-Words (BoW) model.

It represents documents in the corpora as bags (multisets).

Bag-of-Words Model

If W is the set of possible words, a document $\boldsymbol{d}$ is a multiset of words: $\boldsymbol{d} \in \mathbb{B}(\mathrm{W})$.

If $\boldsymbol{D} = \mathbb{B}(\mathrm{W})$ is the set of all possible documents, a corpus $\mathbf{c}$ is a multiset of documents : $\mathbf{c} \in \mathbb{B}(\mathrm{D})$.

Bag-of-Words Model

Rumsfeld 2012

"… there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns - the ones we don't know we don't know."

['known'^{11}, 'unknown'^{3}, 'thing'^{2}, …]

Chair of Process and Data Science

Bag-of-Words Model

Very naïve model: multiset representation loses the order of the items:

“James loves watching movies, Kate hates it.”
“Kate loves watching movies, James hates it.”

“Chickens hatch from eggs.”
“Eggs hatch from chickens.”

It can be a bad model for these sentences (and for applications that strongly depend on order of words).

Bag-of-Words Model

Why is it useful then?

- Very simple to implement.
- In some applications, it just works well enough.
  - Historically tied to the first information retrieval systems.
  - Used successfully for a long time in spam detection software.

- Today it is often used not as a final model, but as an intermediate step (feature extraction) in combination with more advanced techniques.

Chair of Process and Data Science

Document-Term Matrix (Bag-of-Words in the form of a table)

When in tabular form, the same idea is called document-term matrix.

Article ID	biolog	biopsi	biolab	biotin	almost	cancer-surviv	cancer-stage	Article Class
00001	12	1	2	10	0	1	4	breast-cancer
00002	10	1	0	3	0	6	1	breast-cancer
00014	4	1	1	1	0	28	0	breast-cancer
00063	4	0	0	0	0	18	7	breast-cancer
00319	0	1	0	9	0	20	1	breast-cancer
00847	7	2	0	14	0	11	5	breast-cancer
03042	3	1	3	1	0	19	8	lung-cancer
05267	4	4	2	6	0	14	11	lung-cancer
05970	8	0	4	9	0	9	17	lung-cancer
30261	1	0	0	11	0	21	1	prostate-cancer
41191	9	0	5	14	0	11	1	prostate-cancer
52038	6	1	1	17	0	19	0	prostate-cancer
73851	1	1	8	17	0	17	3	prostate-cancer

doi:10.1371/journal.pone.0162721.t001

Chair of Process and Data Science

Term Frequency

The document-term matrix contains the so-called **term frequencies (tf)**, the number of occurrences of a word in a document.

$$tf(w,d) = \#of\ occurrences\ of\ word\ w\ in\ document\ d$$

Inverse Document Frequency

Another important metric is the inverse document frequency (idf). It is a measure of the specificity of a word in the corpus that contains it.

$$idf(w) = log_2(\frac{N}{\#of\ documents\ that\ contain\ w\ at\ least\ once})$$

with N equal to the number of the documents in the corpus.

Intuitive explanation: $idf(w) = log_2(\frac{1}{P(w)})$, i.e., the more unlikely, the higher the value.

Chair of Process and Data Science

Inverse Document Frequency

- Words with a very low idf score appear in many documents.
- This means that those words are not very representative in discriminating between documents.
- Some of them can be candidate stopwords, and could be added to the stoplist.

Tf-idf Scoring

Combining tf and idf gives the tf-idf scoring

$$tfidf(w,d) = tf(w,d) * idf(w)$$

- **An essential ingredient of information retrieval.**
- **Mediates between**
  - **the relevance of a word in a corpus ($idf(w)$), and**
  - **the strength of the association between a word and a document ($tf(w,d)$).**

Tf-idf Scoring

‘Cats are the only pet of the <u>felines</u> family, while dogs are canids.’

‘Cats are the third-most popular pet in the US.’

‘Dogs have been selected for millennia as pet animals.’

‘Normally, dogs are not aggressive towards other dogs outside their territory.’

Stem: feline

tf = 1

idf = log(4/1) = 2

tfidf = 1*2 = 2

$$tf(w,d) = \#of\ occurrences\ of\ word\ w\ in\ document\ d$$

$$idf(w) = log_2\left(\frac{N}{\#of\ documents\ that\ contain\ w\ at\ least\ once}\right)$$

Tf-idf Scoring

‘<u>Cats</u> are the only pet of the felines family, while dogs are canids.’

‘<u>Cats</u> are the third-most popular pet in the US.’

‘Dogs have been selected for millennia as pet animals.’

‘Normally, dogs are not aggressive towards other dogs outside their territory.’

Stem: cat

tf = 1

idf = log(4/2) = 1

tfidf = 1*1 = 1

$$tf(w,d) = \#of\ occurrences\ of\ word\ w\ in\ document\ d$$

$$idf(w) = log_2\left(\frac{N}{\#of\ documents\ that\ contain\ w\ at\ least\ once}\right)$$

Tf-idf Scoring

‘Cats are the only pet of the felines family, while dogs are canids.’

‘Cats are the third-most popular pet in the US.’

‘Dogs have been selected for millennia as pet animals.’

‘Normally, dogs are not aggressive towards other dogs outside their territory.’

Stem: pet

tf = 1

idf = log(4/3) = 0.41

tfidf = 1*0.41 = 0.41

$$tf(w,d) = \#of\ occurrences\ of\ word\ w\ in\ document\ d$$

$$idf(w) = log_2(\frac{N}{\#of\ documents\ that\ contain\ w\ at\ least\ once})$$

Tf-idf Scoring

‘Cats are the only pet of the felines family, while dogs are canids.’

‘Cats are the third-most popular pet in the US.’

‘Dogs have been selected for millennia as pet animals.’

‘Normally, dogs are not aggressive towards other dogs outside their territory.’

Stem: dog

tf = 2

idf = log(4/3) = 0.41

tfidf = 2*0.41 = 0.82

$$tf(w,d) = \#of\ occurrences\ of\ word\ w\ in\ document\ d$$

$$idf(w) = log_2(\frac{N}{\#of\ documents\ that\ contain\ w\ at\ least\ once})$$

Tf-idf Scoring

Even if extremely simple, many querying systems rely on (variations of) tf-idf!

- Given a query and a corpus
- For each document in the corpus
  - Compute $\boldsymbol{score(query, d) = \sum_{w \in query} tfidf(w, d)}$
- Rank documents by score
- Return first n documents

Chair of Process and Data Science

Document-Term Matrix

The document-term matrix can also be built with the tf-idf scores.

w1	w2	w3	w4	...

each column is a word/term

each row is a document

each cell contains the tf-idf score

The matrix allows us to apply a wide range of data science techniques.

Chair of Process and Data Science

Document Classification

Article ID	biolog	biopsi	biolab	biotin	almost	cancer-surviv	cancer-stage	Article Class
00001	12	1	2	10	0	1	4	breast-cancer
00002	10	1	0	3	0	6	1	breast-cancer
00014	4	1	1	1	0	28	0	breast-cancer
00063	4	0	0	0	0	18	7	breast-cancer
00319	0	1	0	9	0	20	1	breast-cancer
00847	7	2	0	14	0	11	5	breast-cancer
03042	3	1	3	1	0	19	8	lung-cancer
05267	4	4	2	6	0	14	11	lung-cancer
05970	8	0	4	9	0	9	17	lung-cancer
30261	1	0	0	11	0	21	1	prostate-cancer
41191	9	0	5	14	0	11	1	prostate-cancer
52038	6	1	1	17	0	19	0	prostate-cancer
73851	1	1	8	17	0	17	3	prostate-cancer

doi:10.1371/journal.pone.0162721.t001

Every document is represented by a vector of constant length

I can use this class attribute as target to train a neural network for classification

Chair of Process and Data Science

Supervised and unsupervised

w1	w2	w3	w4	…

Next to the tf-idf values there may be metadata or annotations/computations providing additional features that may be used as target features.

w1	w2	w3	w4	…	f1	f2	f3	…

Business as usual!

Document Clustering

The document-term matrix allows for another immediate application: document clustering.

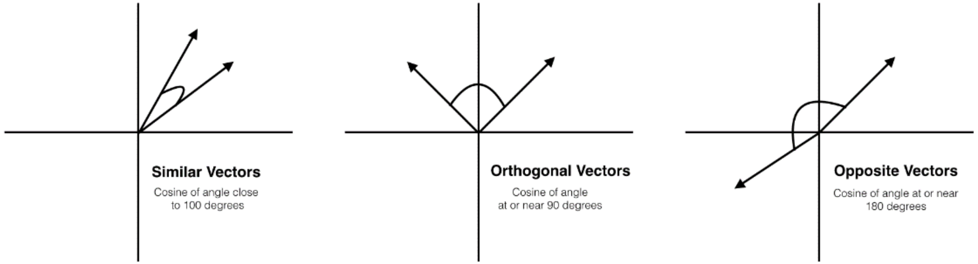
Having fixed length vector, in order to perform clustering we need a distance/similarity measure

Document Clustering

Recall the lecture on Clustering: we can use cosine similarity

$$sim(x, y) = \frac{x \cdot y}{||x|| \, ||y||}$$

approaches 1.0 | approaches 0.0 | approaches -1.0



Document Vector or Term-Frequency Vector

Document	*team*	*coach*	*hockey*	*baseball*	*soccer*	*penalty*	*score*	*win*	*loss*	*season*
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

Document Clustering

Once we can have a distance metric on text data, we can perform clustering with any standard algorithm (k-means, k-mediods, dbscan, etc.).

Cosine similarity is very well suited to compute for very sparse vectors or vectors of different lengths.

Conclusion

Short summary of the lecture

- Basics and definitions of Text Mining.
- Text preprocessing: tokenization, stopword removal, stemming and lemmatization.
- The BoW model and some applications.
- Tfidf weighting.

Next lecture

In the next lecture we will address the two main limitations of BoW models:

- **How to retain the order of the words?**
- **How to deal with the sparseness of the resulting document vectors?**

#	Lecture	date	day
Lecture 1	Introduction	10/10/2018	Wednesday
Lecture 2	Crash Course in Python	11/10/2018	Thursday
Instruction 1	*Python*		
Lecture 3	Basic data visualisat		
Lecture 4	Decision trees		
Instruction 2	*Decision trees and*		
Lecture 5	Regression		
Lecture 6	Support vector mac		
Instruction 3	*Regression and sup*		
Lecture 7	Neural networks (1/		
Instruction 4	*Neural networks an*		
Lecture 8	Neural networks (2/		
Lecture 9	Evaluation of super		
Instruction 5	*Neural networks an*		
Lecture 10	Clustering		
Lecture 11	Frequent items sets		
Lecture 12	Association rules		
Lecture 13	Sequence mining		
Instruction 6	*Clustering, frequent*		
Lecture 14	Process mining (uns		
Lecture 15	Process mining (supervised)	29/11/2018	Thursday
Instruction 7	*Process mining and sequence mining*	*30/11/2018*	*Friday*
Lecture 16	Text mining (1/2)	05/12/2018	Wednesday
Instruction 8	*Text mining and process mining*	*06/12/2018*	*Thursday !!*
Lecture 17	Text mining (2/2)	12/12/2018	Wednesday
Lecture 18	Data preprocessing, data quality, binning, etc.	13/12/2018	Thursday
Lecture 19	Visual analytics & information visualization	19/12/2018	Wednesday
backup		20/12/2018	Thursday
Instruction 9	*Text mining, preprocessing and visualization*	*21/12/2018*	*Friday*
Lecture 20	Responsible data science (1/2)	09/01/2019	Wednesday
Lecture 21	Responsible data science (2/2)	10/01/2019	Thursday
Instruction 10	*Responsible data science*	*11/01/2019*	*Friday*
Lecture 22	Big data (1/2)	16/01/2019	Wednesday
Lecture 23	Big data (2/2)	17/01/2019	Thursday
Instruction 11	*Big data*	*18/01/2019*	*Friday*
Lecture 24	Closing	23/01/2019	Wednesday
backup		24/01/2019	Thursday
Instruction 12	*Example exam questions*	*25/01/2018*	*Friday*
backup		30/01/2019	Wednesday
backup		31/01/2019	Thursday
extra	*Question hour*	*01/02/2019*	*Friday*

Lecture 14	Process mining (unsupervised)	28/11/2018	Wednesday
Lecture 15	Process mining (supervised)	29/11/2018	Thursday
Instruction 7	*Process mining and sequence mining*	*30/11/2018*	*Friday*
Lecture 16	Text mining (1/2)	05/12/2018	Wednesday
Instruction 8	*Text mining and process mining*	*06/12/2018*	*Thursday !!*
Lecture 17	Text mining (2/2)	12/12/2018	Wednesday
Lecture 18	Data preprocessing, data quality, binning, etc.	13/12/2018	Thursday
Lecture 19	Visual analytics & information visualization	19/12/2018	Wednesday
backup		20/12/2018	Thursday
Instruction 9	*Text mining, preprocessing and visualization*	*21/12/2018*	*Friday*

Tag der Informatik 2018

Freitag, 07.12.2018, 12:00 Uhr

Wir freuen uns sehr, Sie auch in diesem Jahr wieder zum Tag der Informatik begrüßen zu dürfen. Im Namen der Fachgruppe Informatik der RWTH Aachen lädt der Lehrstuhl Informatik 9 (i9) zur festlichen Absolventenfeier ein.

Anfahrt

Der Tag der Informatik 2018 findet im Informatik-Zentrum statt. Finden Sie hier weitere Informationen zur Anfahrt.

Programm

Der Tag der Informatik beginnt um **12:00 Uhr** mit der Firmenkontaktmesse im Foyer.

Uhrzeit	Programmpunkt
12:00 - 17:00	Firmenkontaktmesse
13:45 - 14:00	Eröffnung
14:00 - 15:00	3MM ("three-minute madness") der Firmen
KAFFEEPAUSE	
15:15 - 15:30	Begrüßung
15:30 - 16:30	Gastvortrag: Process Mining – Discovering Process Maps from Data
KAFFEEPAUSE	
16:45 - 17:15	Ansprache an die Absolventen
17:15 - 17:45	Preisverleihung
17:45 - 19:00	Absolventenehrung
19:00 - 19:15	Sektempfang
19:15 - 21:00	Buffet
21:00 - 23:00	Party

Für die Teilnahme an der Abendveranstaltung wird ein Teilnahmebändchen benötigt. Alle Absolventen erhalten diese für sich und ihre Begleitpersonen bei der Zeugnisverleihung.

Wil van der Aalst
@wvdaalst

Looking forward to the keynote of Anne Rozinat @arozinat on Friday as part of the "Tag der Informatik" @RWTH. She is the co-founder of @fluxicon and a role model for all bright, entrepreneurial, female, computer scientists! informatik.rwth-aachen.de/cms/Informatik ... #RWTH #Informatik

Tweet übersetzen

12:18 - 3. Dez. 2018

10 Retweets **36** „Gefällt mir"-Angaben

Fluxicon, PADS - Process And Data Science und Anne Rozinat

10 36