

# Analysis of text data

Extracting information from natural text

# Learning goals

1. Understand the role and concept of text mining and web mining.
2. Learn the basic workflow in mining natural text.
3. Understand the potential uses for text and web mining.

# Text mining

ve all expansions and have consid  
ra dividers to go along with a coup  
ning insert but would need to purc  
) addition dividers! This would cost  
I'm wondering if the rigid, fixed di  
t great for going in between every  
d type since you can't scoot them  
cal card stock-like dividers and, ev  
se dividers are clear acrylic, it migh  
kingdom cards with so many of th  
d in place so close together since t

vs.

ID	Likes_travel	Likes_nature	Likes_bars_clubs	Income	Gender
1	9	5	4	3176	F
2	9	5	6	2633	F
3	10	0	2	4180	M
4	3	2	5	2229	F
5	6	5	5	3089	F
6	7	2	5	2197	F
7	1	1	5	4383	M
8	9	5	10	2213	F
9	0	5	4	3847	F
10	2	7	3	4183	M
11	10	7	7	4809	M
12	5	4	5	2087	M

- Finding previously unknown information from text masses in an automated way.
- The challenge arises from the format of text data.
  - Words, sentences and paragraphs vs. variables and values.
- The content is often in the web.
  - Text mining from web content is a subdomain of web mining.

# Popular examples

- Word cloud
  - <https://worditout.com/word-cloud/create>
- Trending vocabulary in social media
  - <https://www.trendsmap.com/>

# Subdomains of Web mining

1. Web content mining
  - Parsing HTML documents and analyzing their content
2. Web structure mining
  - Analysis of the structure of the hyperlink network
  - Analogy to the analysis of social media networks.
3. Web usage mining
  - Analysis of log files.

# Web content mining in Python

```
from bs4 import BeautifulSoup
from urllib.request import urlopen

url = "https://www.nngroup.com"
page = urlopen(url)
soup = BeautifulSoup(page, 'html.parser')
print(soup.get_text())
```

- In Python, BeautifulSoup library contains web scraping functionality.
- However, API access is preferable to 'scraping'.

# Focus

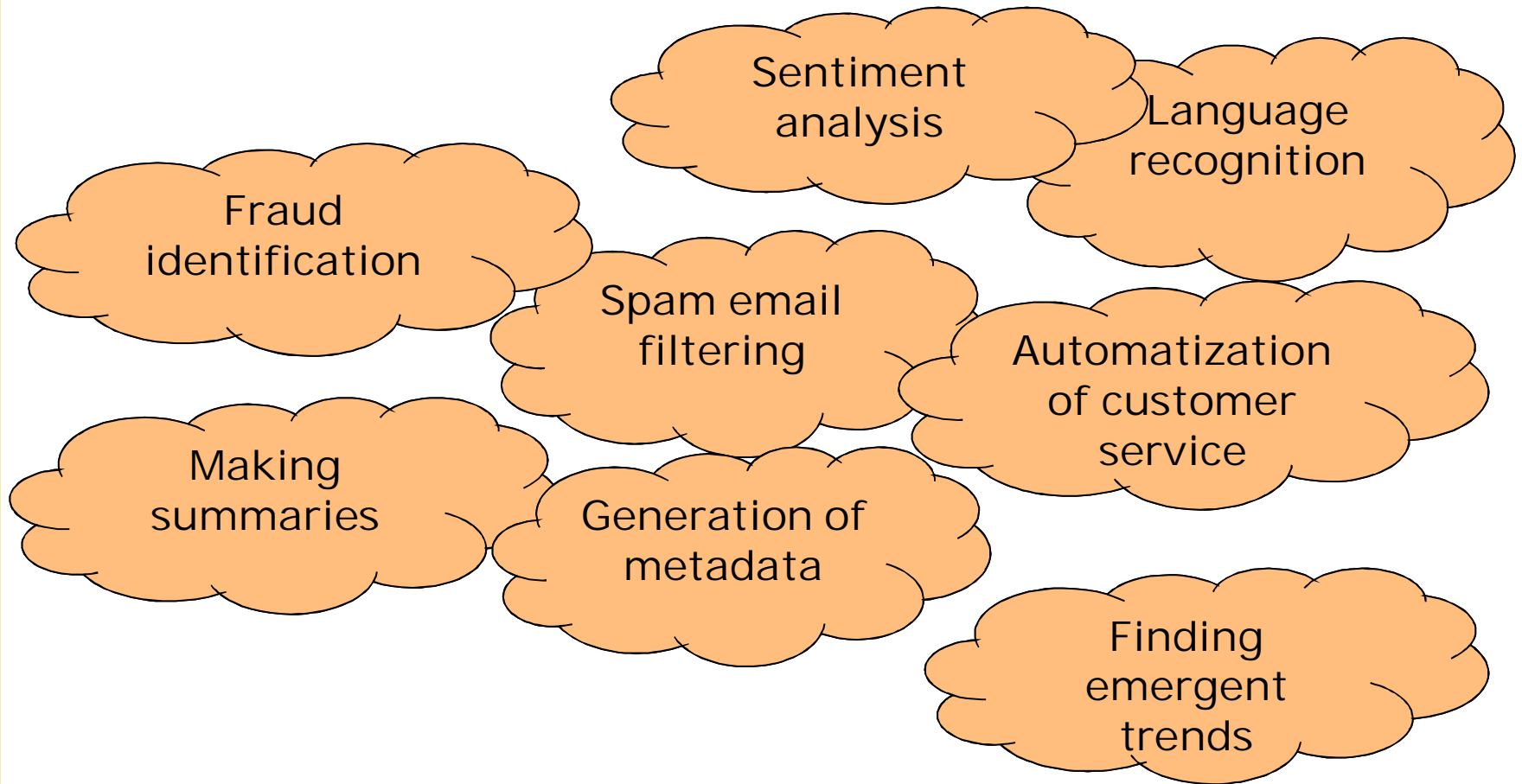
- In this session we focus in the techniques of text mining.
- The techniques are also applicable to the analysis of web documents (web content mining)
  - Special challenges:
    - Larger data size
    - Need suitable access methods (web scraping vs. API access)
    - Need for preprocessing (usually in machine-generated 'messy' HTML format)
    - Dynamic nature of data
- In the realm of deep neural networks, word2vec models can be used for recommending/prediction.
  - In this presentation, we focus on the 'shallow' word-count based approach for analysis of text data.

# Idea of text mining

- The text data is first transformed into tabular format.
  - The variables can represent the number of occurrences of each word in each text.
  - This transformation can be easier done with tools such as:
    - Pattern matching languages
    - Regular expressions
- Once the data is in tabular, numeric format, it can be analyzed using 'traditional' data mining methods.
  - Classification
  - Clustering
  - Association rule mining



# Fields of use



# Text mining process

- Let's get acquainted with the text mining process by following an example.
  - The core text-to-tabular processing phase is similar irrespective of the mining objective or the chosen analysis method.
- Example:
  - Perform sentiment analysis: evaluate the inauguration speeches of US presidents from Washington to Trump.
  - Goal: rank the speeches on their emotional content ('affection status').

# Text processing



Row No.	aamir	aaron	aaronah	aback	abandon	abbott	abdulrashidh	aber	aberlour	abhishek
1	0	0	2	0	5	0	0	1	1	0
2	1	106	0	1	1	1	1	0	0	1
3	0	0	0	0	2	0	0	0	0	0

- The purpose of text processing phase is to transform the texts into tabular format.
- Each word in text turns into a new variable.
- The value of the variable is the frequency (i.e. number of occurrences) of the word in the data set.
  - In the example, there are three documents. Each one produces an observation (a row) into the data set.
- The variable derived from the word, together with its values, is called a word vector.
- It is possible to derive the word vectors for group of subsequent words instead of individual words.
  - These are called n-grams.
  - In practice, only 2-grams are considered.

# Text processing

- The processing step is comprised of four phases:
  1. Transform case.
    - Make capital and lower-case words to be treated uniformly.
  2. Tokenize.
    - Find individual words (or n-grams).
  3. Filter stopwords.
    - Ignore common words with little value to the analysis.
  4. Stem/Lemmatize
    - Merge variants of the same word.
- Let's pay further attention into these steps.

# Phase 1: Transform case

This is a nice sentence. This sentence is even nicer.

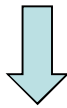


this is a nice sentence. this sentence is even nicer.

- The distinction between capital and lower-case letters is removed.
- As a consequence, it will make no difference whether the word appears in the beginning of a sentence or not.

## Phase 2: Tokenize

this is a nice sentence. this sentence is even nicer.



this	is	a	nice	sentence	even	nicer
2	2	1	1	2	1	1

- In this phase, each word is turned into a variable.
- The punctuation marks are removed at this point.
- Instead of individual words, 2-grams could be considered.

## Phase 3: Filtering stop words

this is a nice sentence even nicer  
2 2 1 1 2 1 1



nice sentence nicer  
1 2 1

- Stopwords are common words that occur frequently in all texts.
- They are not considered interesting for text mining. Thus, they are often removed.
- A stopwords list is required.

# Phase 4: Stemming and/or lemmatization

nice    sentence    nicer  
1        2        1

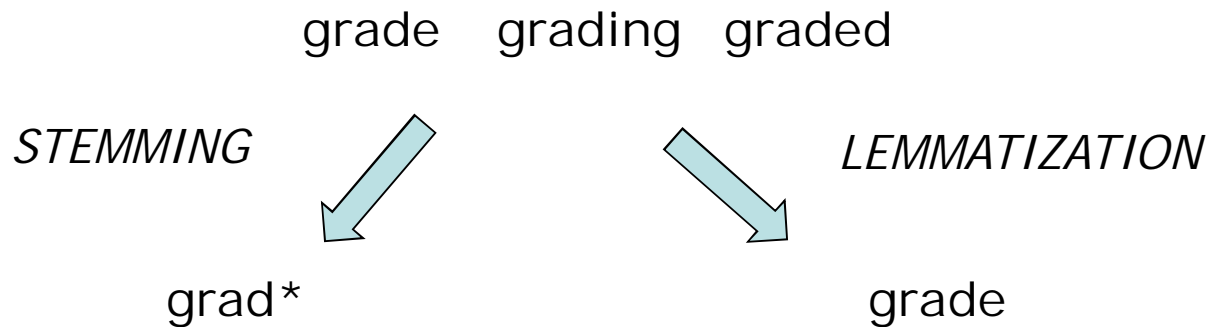


nice\*    sentence\*  
2        2

- Depending on the language, words can have multiple inflected forms.
- It is not feasible to consider forms of the same word as different words.
  - The problem is most evident in agglutinative languages such as Finnish.
- The merger of inflected words can be done in multiple ways:
  - The words can be replaced by a shorter substring that is free from inflection (see image above).
  - Alternatively, the stems can be recognized by morphological analysis (harder to implement).



# Stemming vs. lemmatization



- Stemming finds the common root of the word.
  - Not necessarily a proper word.
- Lemmatization generates the dictionary form of the word.

# What about synonyms?

- Words with multiple meanings may 'dilute' the data.
- This problem is usually ignored.
- Rationale: in Big Data context, the slight degradation of data cause by synonyms is compensated by a large data size.

## sentence

noun [C]   /ˈsen.təns/



### sentence noun [C] (WORD GROUP)

**A1** a group of words, usually containing a verb, that expresses a thought in the form of a statement, question, instruction, or exclamation and starts with a capital letter when written:

*He's very impatient and always interrupts me mid-sentence.*

*Your conclusion is good, but the final sentence is too long and complicated.*

### More examples

### sentence noun [C] (PUNISHMENT)

**B2** a punishment given by a judge in court to a person or organization after they have been found guilty of doing something wrong:

*He got a **heavy/light** sentence (= he was severely/not severely punished).*

*The offence **carries** a **jail/prison/life/five-year** sentence.*

*He was given a **non-custodial/suspended** sentence.*

Image: Cambridge dictionaries online.

# Analysis possibilities

Row No.	a	abid	abound	abund	access	acclaim
1	0	0	1	0	0	0
2	0	2	1	1	1	7
3	0	0	0	1	0	0

- Once the word vectors have been produced, 'traditional' data mining methods can be applied:
  - Clustering: which documents are similar in vocabulary?
    - Finding an unknown author.
  - Association analysis: which words are used together?
    - Gathering domain-specific vocabularies.
  - Decision tree / neural network
    - Automatic classification of customer feedback (angry complaint, error report, suggestion etc.)
    - An learning set is required.
- The absolute term frequencies are usually replaced by measures taking into account:
  1. The length of each document
  2. The rareness of the word in the entire corpus.
- TF-IDF is a commonly used measure.
  - Several weighting schemes, see e.g. <https://en.wikipedia.org/wiki/Tf%E2%80%93idf>

# Text mining in Python

- In the course's Oma workspace, see Documents/Methods/Data/Inaugural speeches (demo).
  - Sentiment analysis
  - Simple word-based clustering
- The text mining functionality is available in the NLTK package.