Topic Modeling: Automated Extraction of Topics from a Corpus

Session 8

Jonas Widmer, University of Bern

WORCK Training School 2, February 2024

Outline

What is Topic Modeling?

Use Cases

Application

Preprocessing with Python

Topic Modeling with **Mallet**

Analysis and Visualization

What Is Topic Modeling?

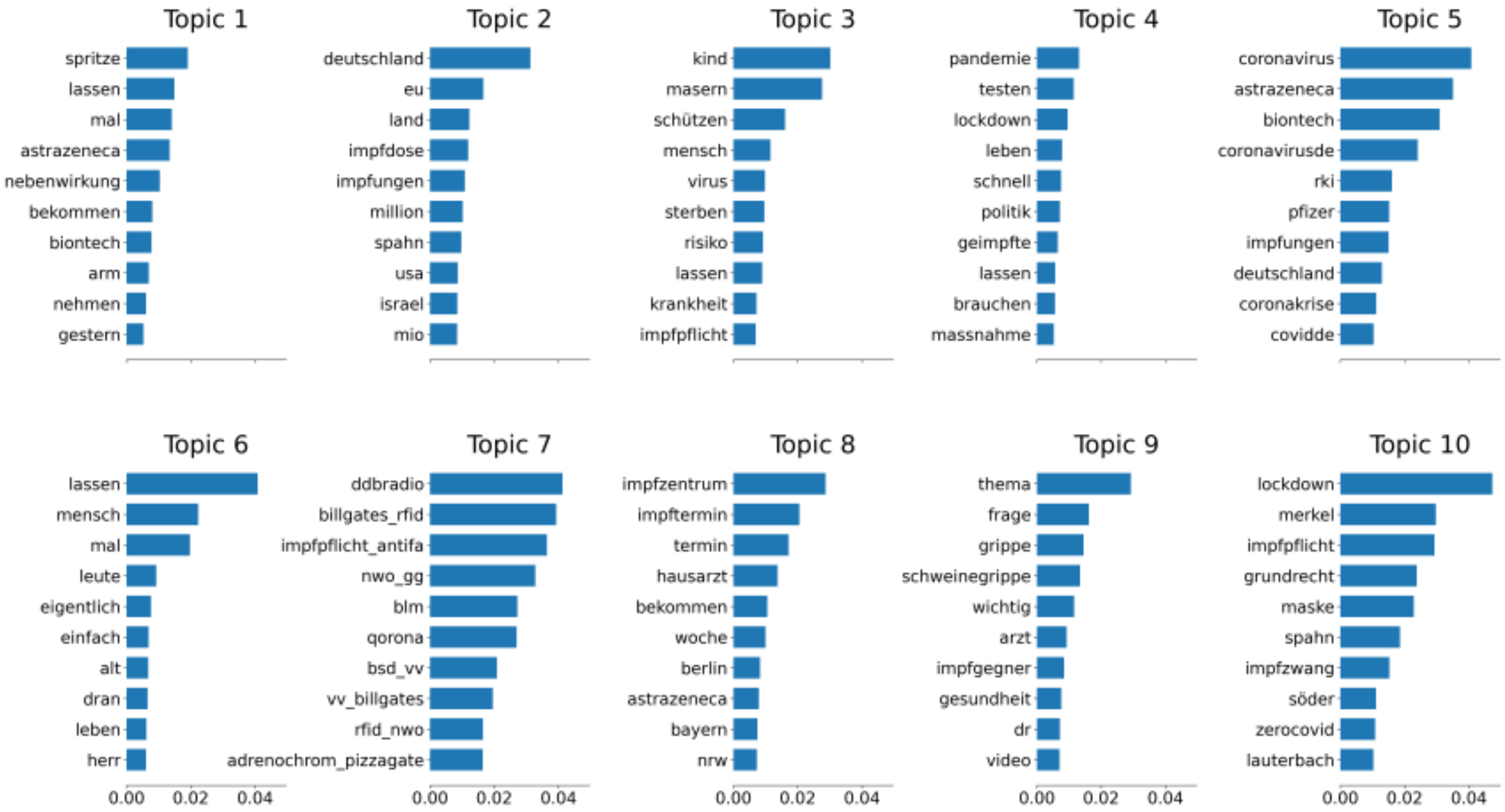
Automated, quantitative analysis of a large text collection with the help of machine learning

Automated detection of topic groups: **probabilistic, unsupervised clustering**

Result: Groups of words that can give indications to topics of the corpus

Objective/added value: Summary and visualization of content & help with the development of theoretical concepts/hypotheses

Example: Tweets 2021 with #impfen/ #impfung

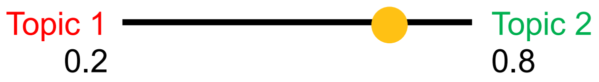


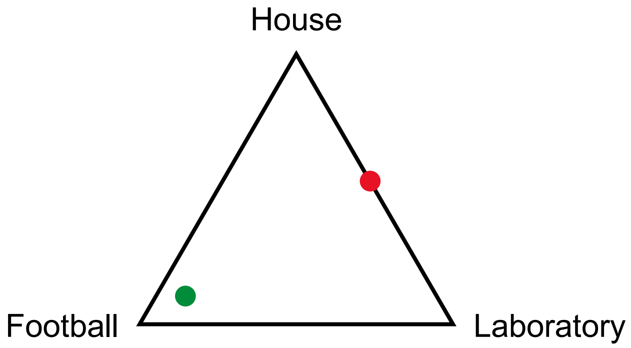
Topic Modeling Theory

- Per document in the corpus (e.g. tweet or article) representation with **bag-of-words**
→ Sequence, grammar and semantics are not taken into account
- Because bag-of-words: multilingual, advantage e.g. for historical research
- **Fixed number of topics** must be determined manuall
- Words in document = context of a word
- Orientation to the **vocabulary** of the corpus

Latent Dirichlet Allocation (Blei et al., 2003)

- Latent = hidden
- Dirichlet Allocation:
  - Probability of topics in document
  - Probability of words in topic
- Algorithm:
  - Genrating random Dirichlet Allocation
  - Random selection of topic from document Dirichlet Allocation
  - Random selection of word from topic Dirichlet Allocation
  - Iterative approximation (= Machine Learning)





	House	Football	Laboratory
T1	0.4	0.1	0.5
T2	0.1	0.8	0.1

Train Topic Structure

Algorithms for approximation

Gibbs Sampling
Oriented towards random samples
Used by Mallet
Variational Inference
Proposed by Blei et al. & continuously optimized
Used by Python Libraries (e.g. Gensim)

Objective: To know the probability distribution across the topics for each word in the vocabulary

Use Cases

- **Quantitative Analysis of Corpora**
  - Extraction of existing topics
    E.g. changes in existing topics in newspaper articles over time
  - Assignment of topics to individual documents
    E.g. analysis of a large letter corpus – which letter contains which topic
- Topics must be tagged manually
- **Quality measures** exist; often good results through manual optimisation of the number of topics (interesting: test through word exchange)

Preprocessing – Example: Tweet

🔥Hey Guys, #ZenithSwap has launched at just $ 55,000 USD Marketcap. The ChatGPT of DEX - Reimagining DeFi with AI-Powered Yield Farming. Now at 4X. Lot of up potential at such low marketcap.🔥😇 $ARB $ZSP #Arbitrum https://t.co/V4pqKF43XN

Preprocessing – Clean Documents

- **Decisions to be taken:**
  - Upper/lower case
  - Stop words: Clean yourself or with Mallet? Own stop word list?
  - Lemmatisation? Part-of-Speech-Tagging (only nouns, verbs, adjectives)?
  - Clean special characters (emojis, #, @, …)? Numbers?
  - Filter short/long words?
  - URLs, e-mail addresses, mentions, hashtags?
  - Germana: ss to ß? (sometimes better lemmatisation)
  - Bigram/trigram?
- **Caution:** Pay attention to the order of preprocessing
  e.g. ‚GPT-3.5‘ and ‚gpt-4‘ should not necessarily both become ‚gpt‘

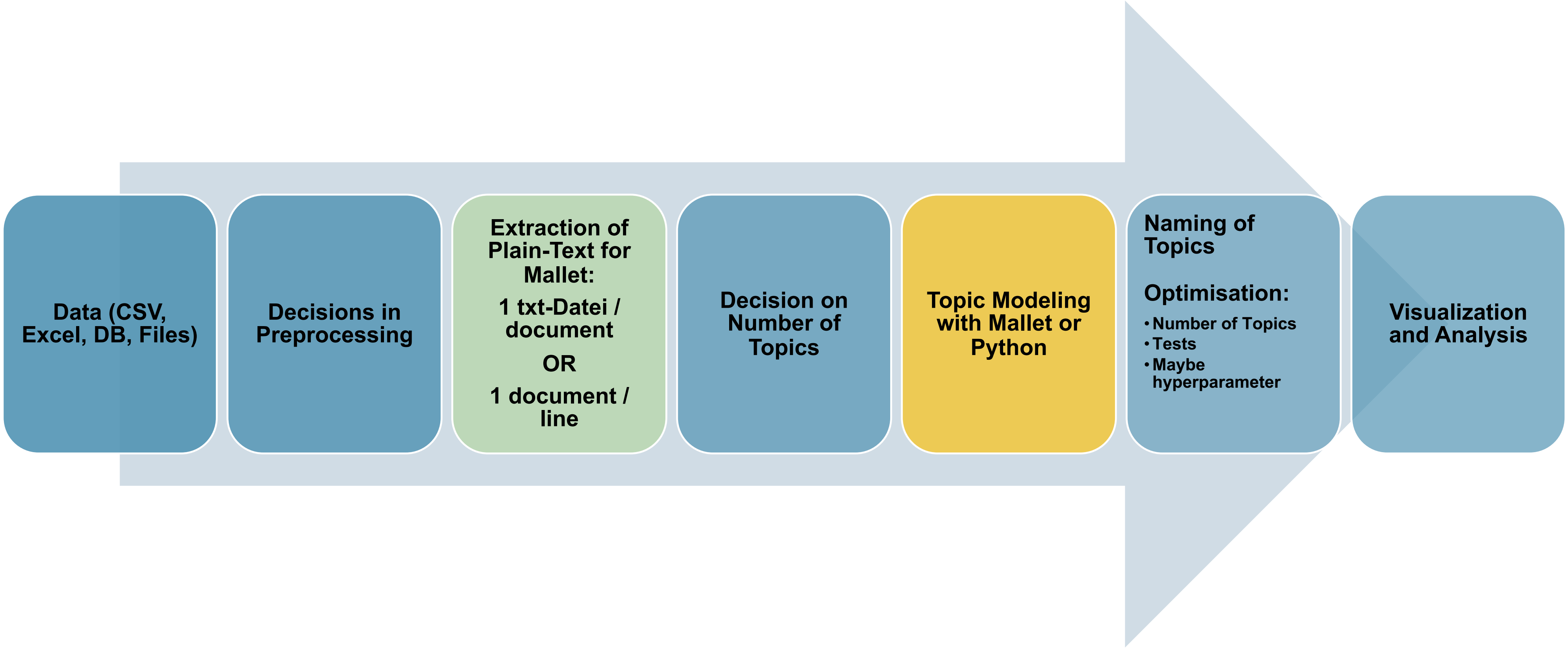
Preprocessing – Example: Tweet

```
🔥Hey Guys, #ZenithSwap has launched at just $ 55,000 USD
Marketcap. The ChatGPT of DEX – Reimagining DeFi with AI-
Powered Yield Farming. Now at 4X. Lot of up potential at
such low marketcap.🔥😇 $ARB $ZSP #Arbitrum
https://t.co/V4pqKF43XN
```

⇩

```
fire launch marketcap reimagine defi powered yield farming
potential marketcap fire smiling_face_with_halo
```

Application



Data (CSV, Excel, DB, Files)

Decisions in Preprocessing

Extraction of Plain-Text for Mallet:

1 txt-Datei / document

OR

1 document / line

Decision on Number of Topics

Topic Modeling with Mallet or Python

Naming of Topics

Optimisation:

- Number of Topics
- Tests
- Maybe hyperparameter

Visualization and Analysis