UNIVERSITY OF AMSTERDAM
Economics & Business

BSc Business Analytics Y3 – 2023/2024

Text Retrieval - Text Mining
Week 3 – Topic Modeling

Julien ROSSI

Week 3

After this lecture, you will:

- Understand the challenges of **Topic Modeling**
- Know about **LDA**, **BERTopic**

Previously in Text Representations

Lexical Representations

- Vocabulary
- 1 dimension = 1 item of the vocabulary

- Preprocessing (lemmatizing, stemming, stopping, …)
- Raw Counts or TF-IDF

- Text = assembly of words

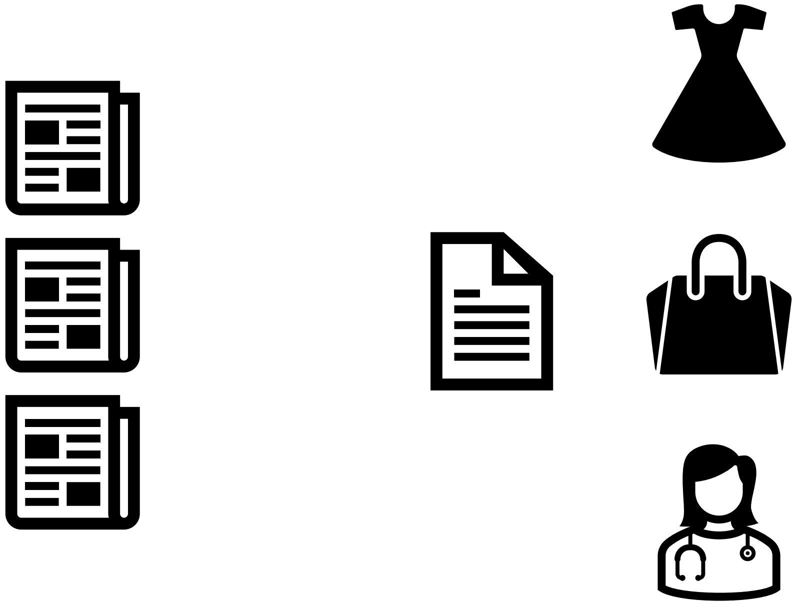
Semantic Representations

Semantic = deals with meaning

- **Text** = mix of **topics** (“science”, “business”, “sport”, …)
- These **topics** are responsible for the **terms** that appear

University of Amsterdam
Economics & Business

Problem



University of Amsterdam
Economics & Business

Problem

"How did **Vogue Magazine** talk about **Health** ?"

- **Without reading 100,000 articles**
- Which words ?

- How many articles ?

Idea 1

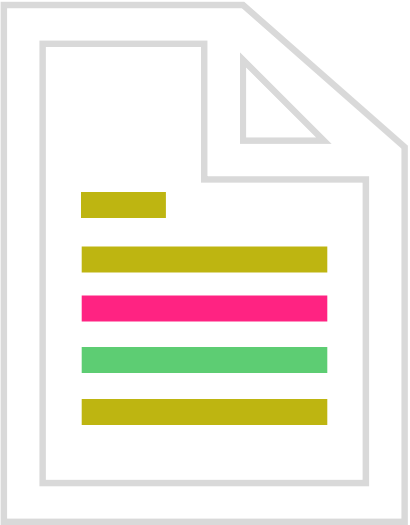
- Split each article into topics:
    - Is it about health?
    - Is it about fashion?
    - Is it about cooking?
    - …

No nuance… and I need to read articles

University of Amsterdam
Economics & Business

Idea 2

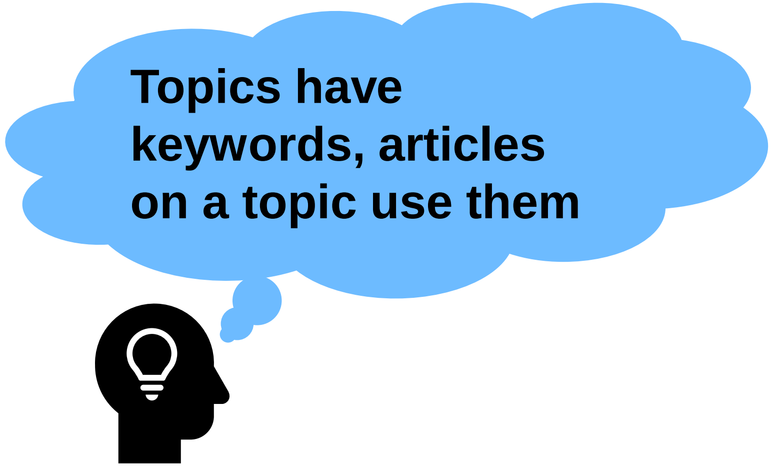
- Split each article into topics:
  - How much % about health?
  - How much % about fashion?
  - How much % about cooking?
  - Etc…



Better… but I need to read articles

University of Amsterdam
Economics & Business

Idea 3



TOPICS

KEYWORDS

ARTICLE

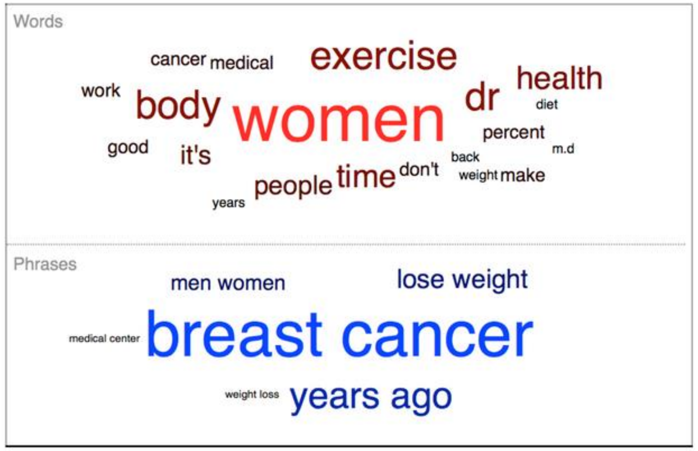
Idea 4


OR… Learn topics +
keywords from articles
TOPICS
??
??
??
KEYWORDS
?? ?? ??
?? ?? ??
?? ?? ??
ARTICLES

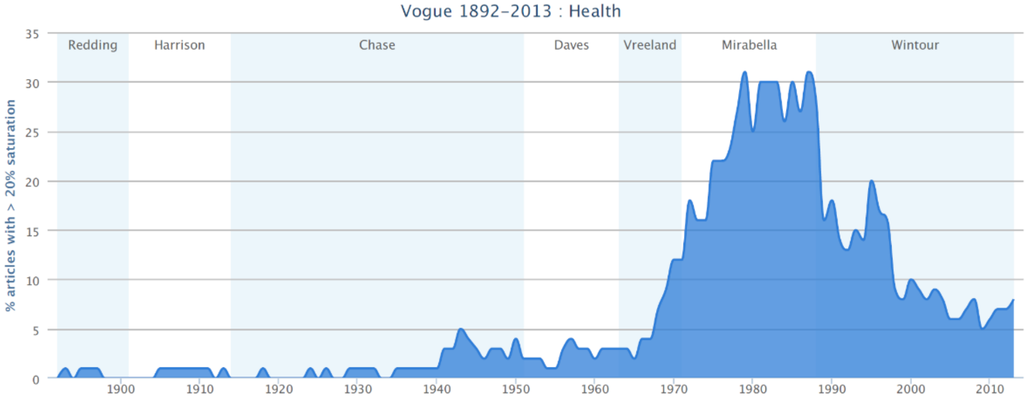
Result

"How did **Vogue Magazine** talk about **Health** ?"

- Which words ?



- How many articles ?



http://library.ifla.org/930/1/119-leonard-en.pdf

University of Amsterdam
Economics & Business

Result

"How did **Vogue Magazine** talk about **Health** ?"

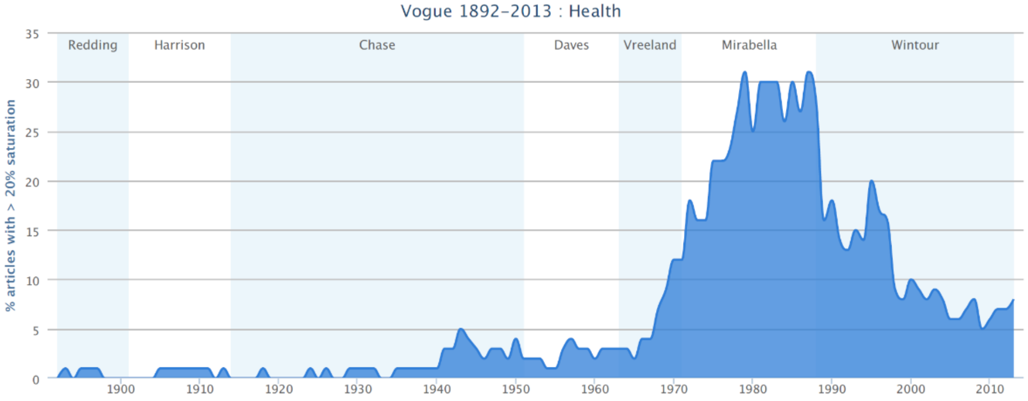
- **Top articles on "Health" (titles only)**

  "Q&A: The pill" (Dec 1987) – 99% about health

  "Facts on Fat: Obesity" (Aug 1979)

  "Inner info: Contraception" (Aug 1978)

  "Crash Diets Don't Work" (Aug 1979)

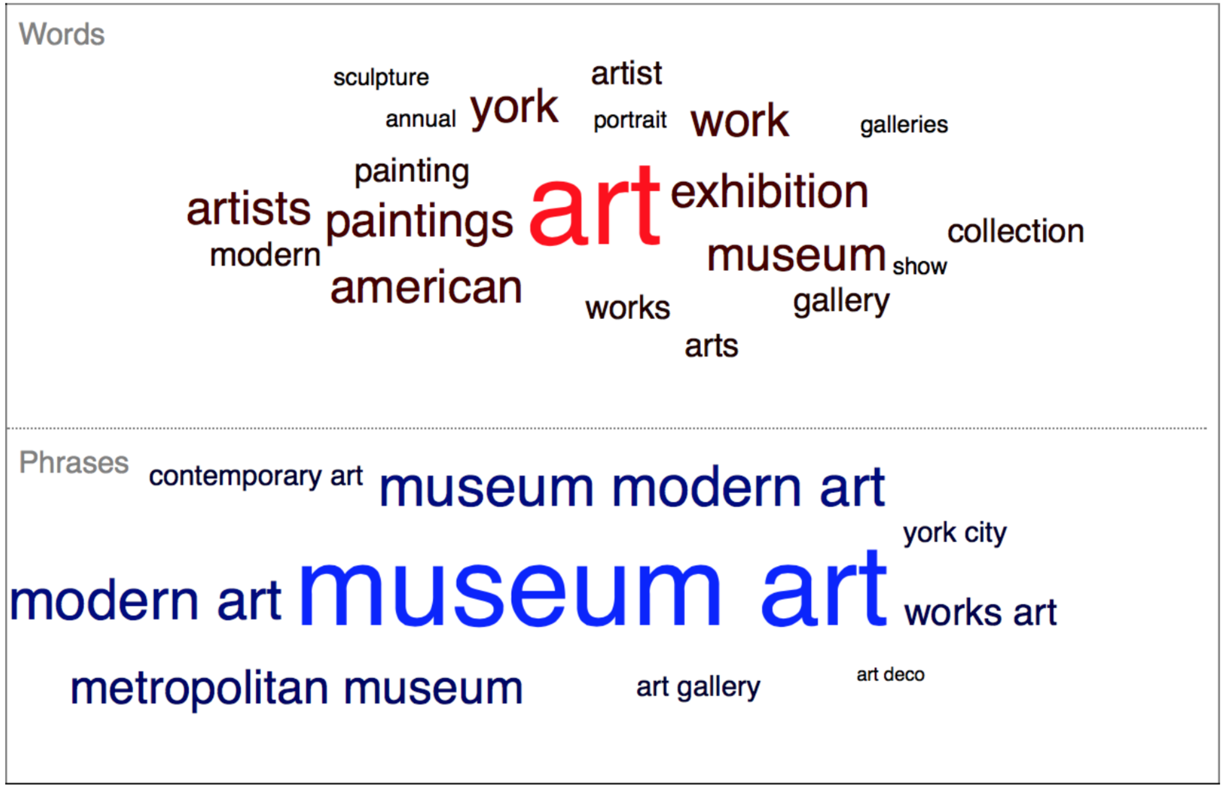


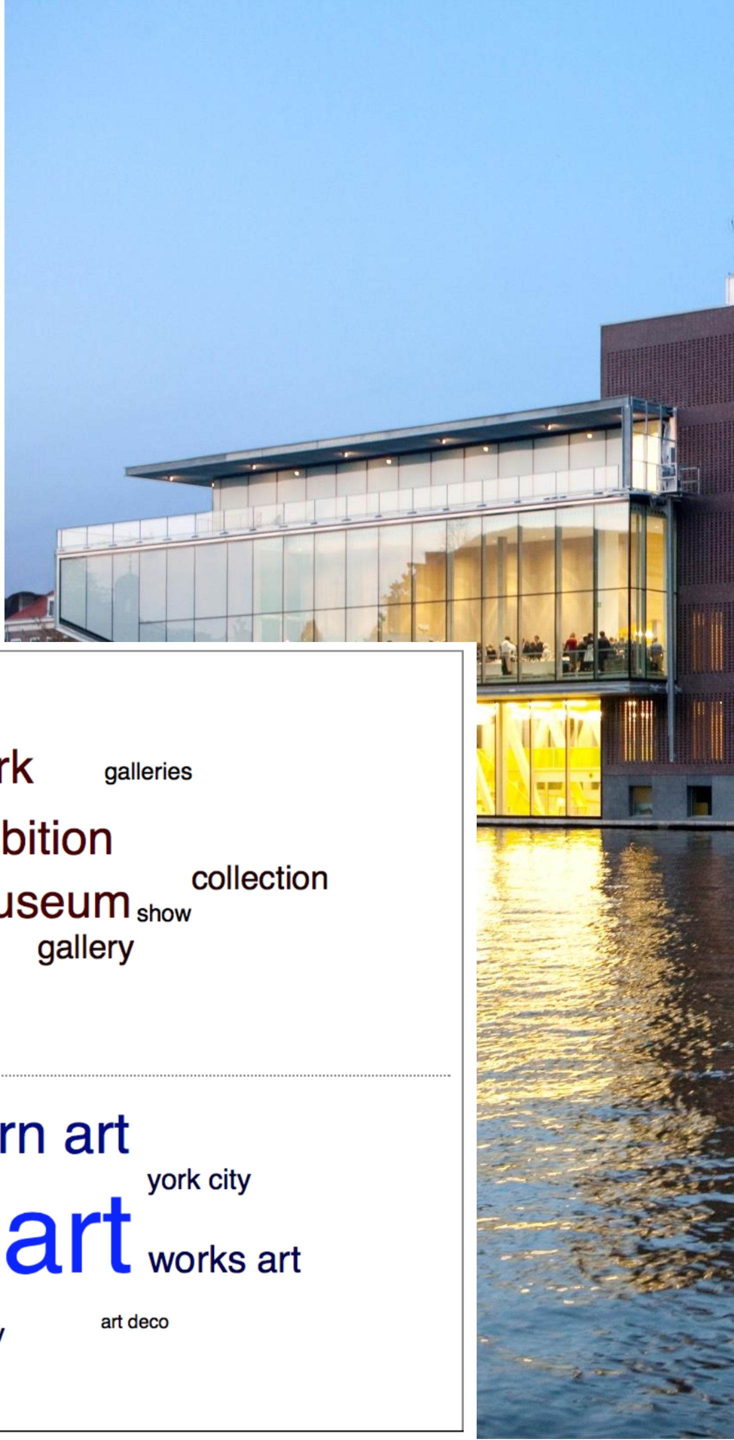
http://library.ifla.org/930/1/119-leonard-en.pdf

Topic Modeling

TOPIC

- **Weighted** list of **terms** (word / n-gram / stem …)
    - High weight = important term
    - Low weight = minor term





Topic Modeling

DOCUMENT

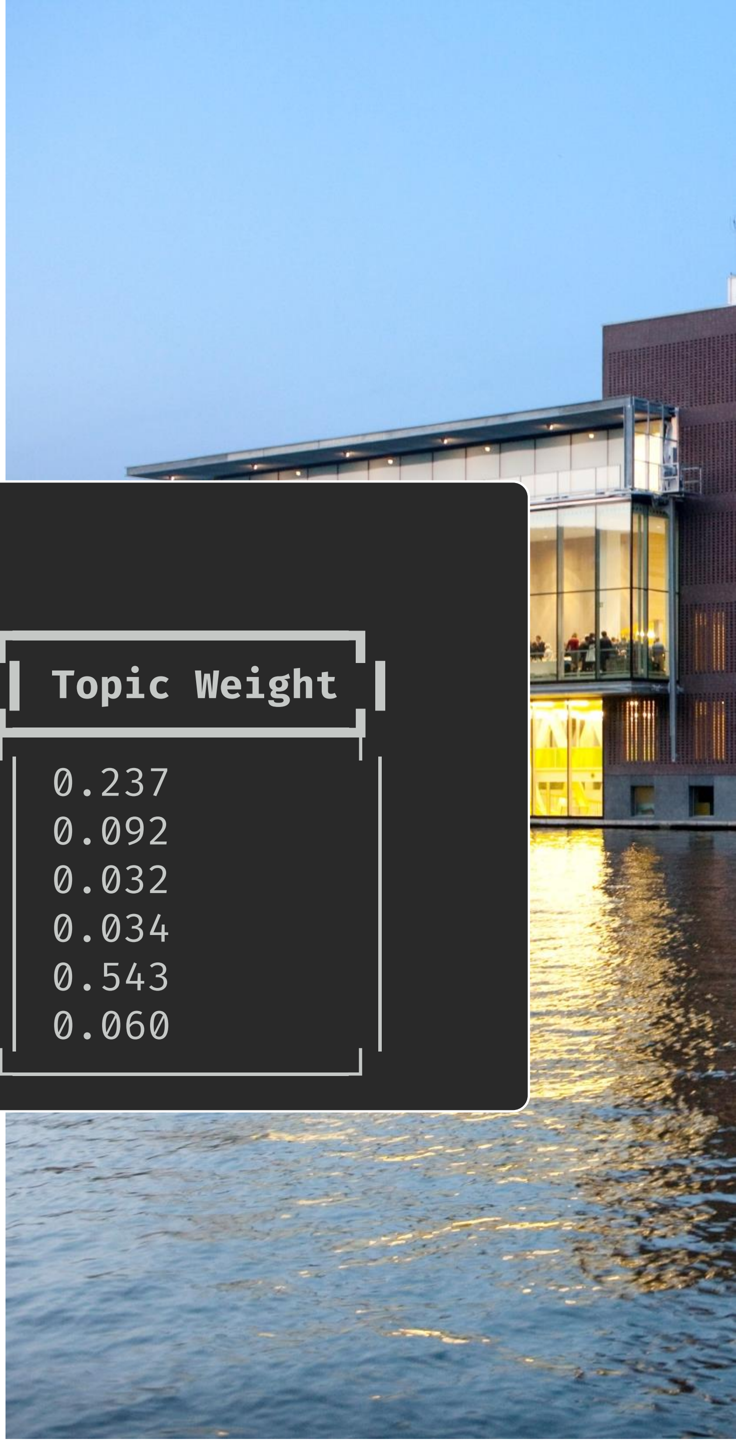
- **Weighted** list of **topics**

Topics in Document

Topics in Document

Topic ID	Topic Name	Topic Weight
00	Rugby	0.237
03	Internet	0.092
04	Politics	0.032
05	Cybersecurity	0.034
06	Mobile Devices	0.543
09	Video Technology	0.060

- Vector representation:
  - 1 dimension = 1 topic
  - Coefficient = weight of topic in document



UNIVERSITY OF AMSTERDAM
Economics & Business

Semantic Representations

Challenge

- Discover the topics from text
- We will see 1 technique:
  - Latent Dirichlet Allocation (LDA)
  - Blei et al. "Latent Dirichlet Allocation", 2003 (Journal of Machine Learning Research) PDF

LDA

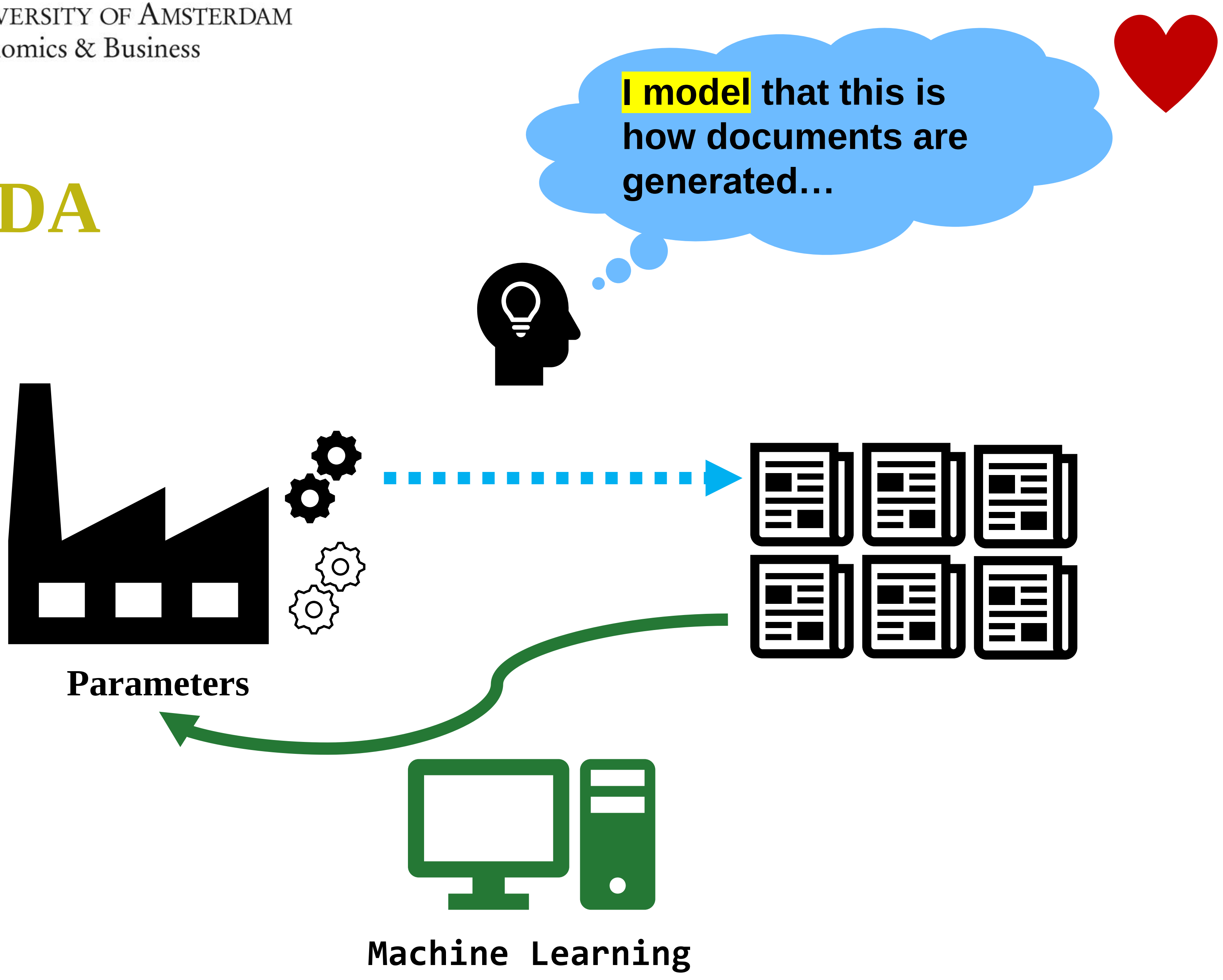
- Discover the topics from text

- Generative Model
- The Bag of Word is generated by random process

- Process:
  - 1 document = sum of weighted topics
  - 1 topic = probability distribution over dictionary

University of Amsterdam
Economics & Business

LDA



LDA – Generate New Document

- **K** (number of topics) is a **hyperparameter**
- $\boldsymbol{\alpha}$ is a **parameter**
- $\boldsymbol{\beta}$ is a **parameter**

1 Draw **N** = number of words in document

2 Draw Topic distribution
- From **Dirichlet** distribution
- With parameter α
- `[0.1, 0.4, 0.3, 0.2]`

LDA – Generate New Document

3

```
Repeat N times:
  Draw a topic (use Topic Distribution)
  Draw ONE word from this topic
        (use word weights in topic = parameter β)
```

4

READY

- We have the **Bag of Words** of the document

LDA – Generate New Document

- **K** (number of topics) is a **hyperparameter** (integer number)
- $\boldsymbol{\alpha}$ is a **parameter** (vector with K dims)
- $\boldsymbol{\beta}$ is a **parameter** (matrix $V \times K$)
    - $\beta_{i,j}$ = weight of word w_i in topic j

- Draw from **Dirichlet** distribution
    - 1 draw = K numbers $\theta_1, \theta_2, \dots, \theta_k$
    - Sum of all numbers equal 1
    - $\theta_1 + \theta_2 + \cdots + \theta_k = 1$
    - E.g. with $K = 3$: `[0.3, 0.6, 0.1]`

LDA - Generation

	Topic “SPORT”	Topic “BUSINESS”
Word 1	football (0.4)	revenue (0.6)
Word 2	ronaldo (0.2)	tax (0.2)
Word 3	goal (0.2)	benefit (0.1)
Word 4	score (0.2)	grow (0.1)
Word 5	tax (0.00001)	goal (0.001)
Word 6	revenue (0.00001)	score (0.00000001)
Word 7	grow (0.0000001)	ronaldo (0.0000001)
Word 8	benefit (0.000000001)	football (0.000000001)

LDA - Generation

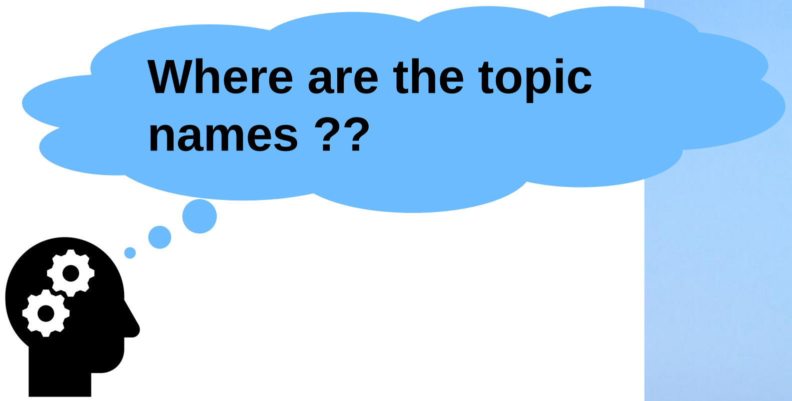
- Document
  - 6 words
  - 0.67 “Sport” + 0.33 “Business”

- Bag of Words:
  - 67% of the time we draw from “Sport”
  - 33% of the time we draw from “Business”

  - **SPORT:** *football*: 1, *ronaldo*: 1, *goal*: 1, *score:* 1
  - **BUSINESS**: *revenue*: 1, *grow*: 1

- “*Football star Ronaldo's revenue grows as he scores many goals.*”

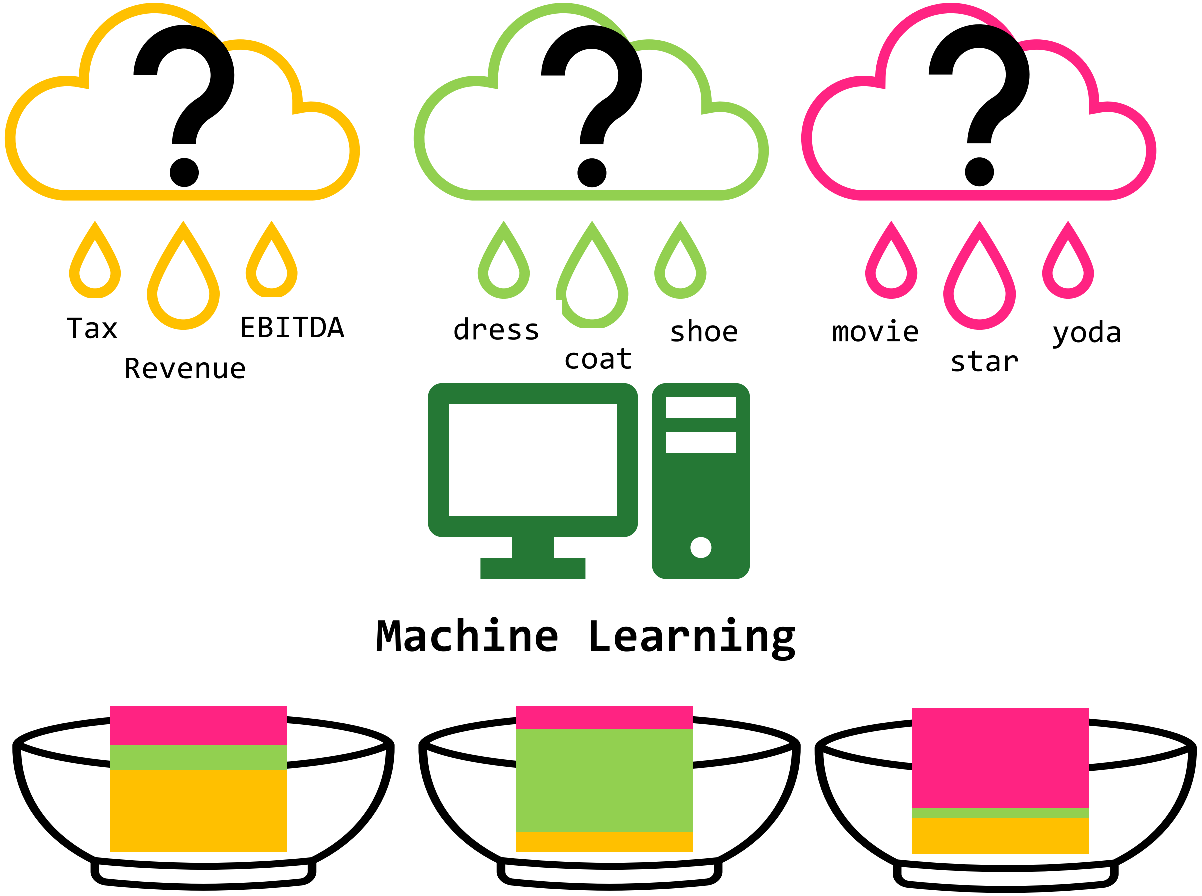
LDA – Topic Labeling



TOPICS

KEYWORDS

ARTICLES



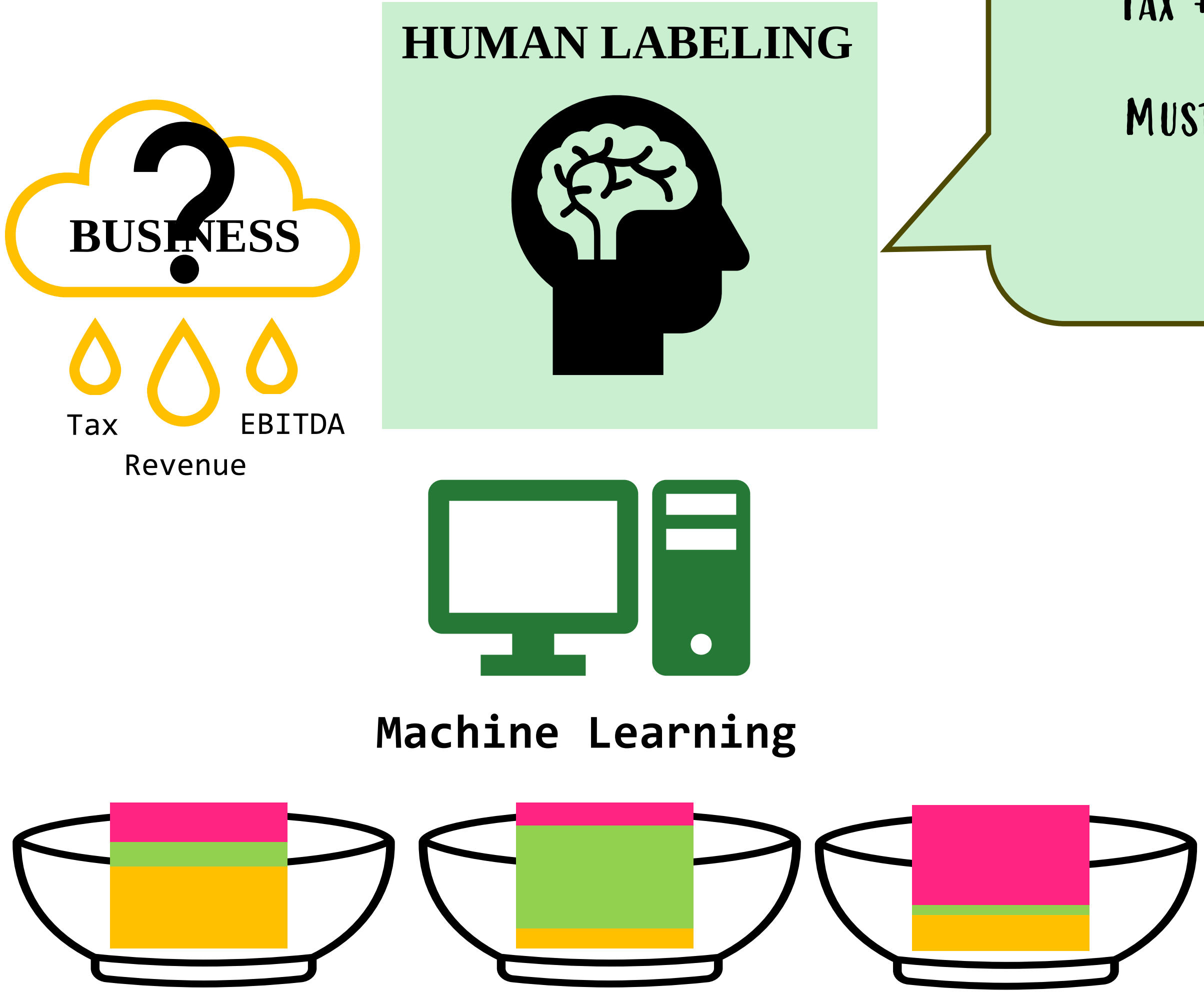
University of Amsterdam
Economics & Business

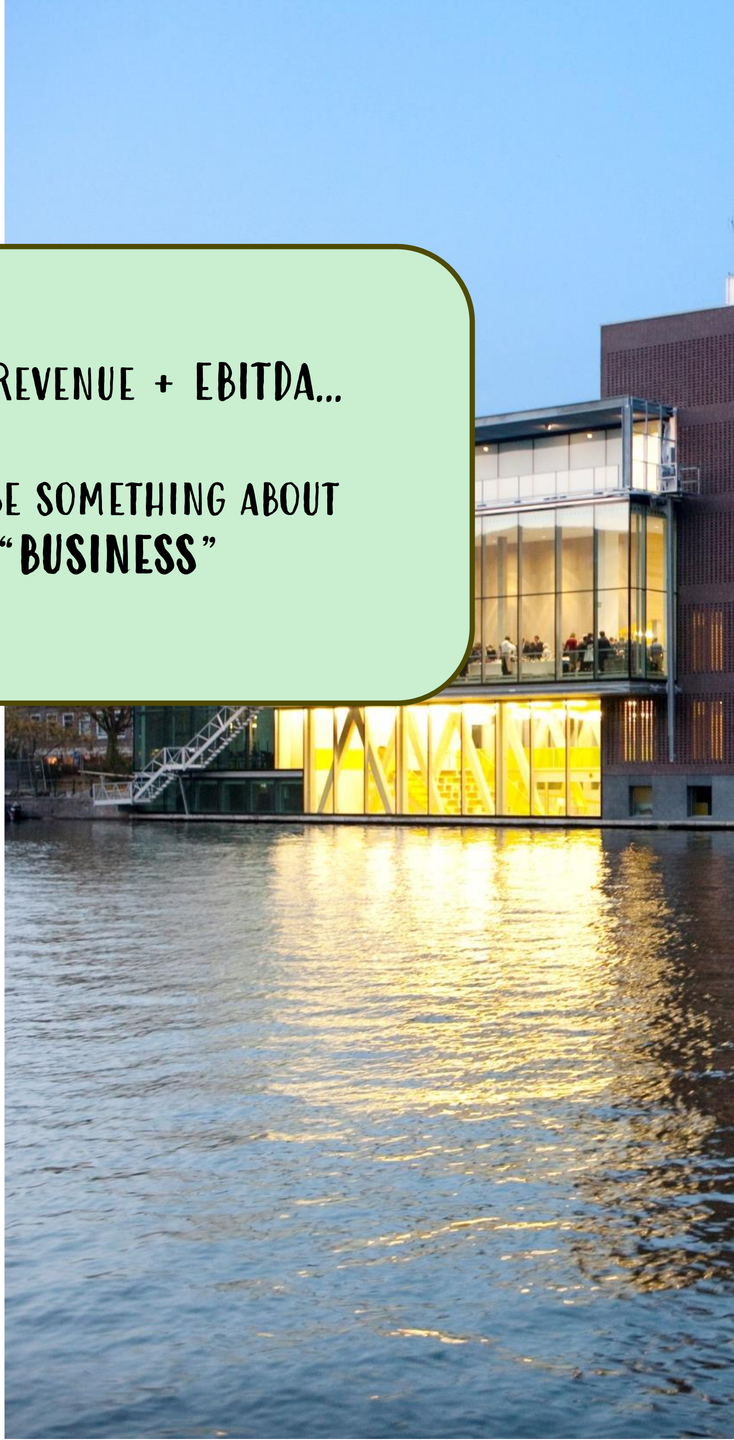
LDA – Topic Labeling

TOPICS

KEYWORDS

ARTICLES

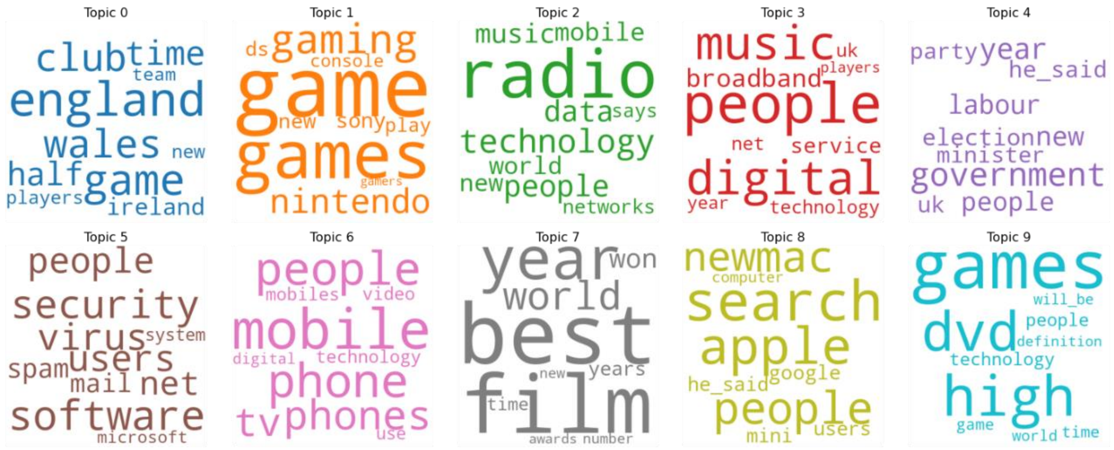




LDA - Learning

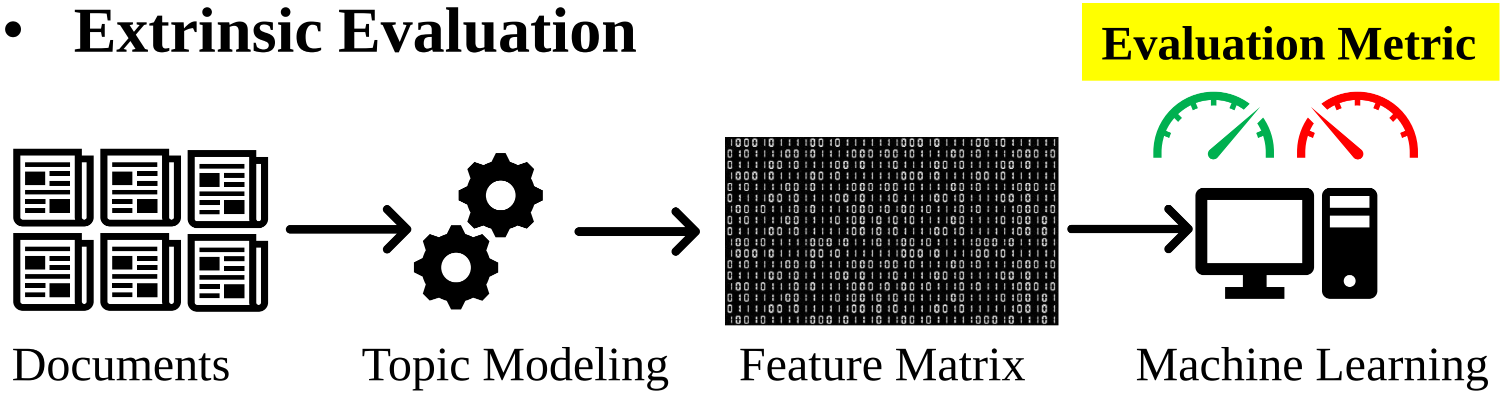
- Given a **collection of documents**
- Given a **number of topics**

- Learn the distribution of topics in documents
- Learn the distribution of words in topics



LDA – Hyperparameter Evaluation

- **Extrinsic Evaluation**

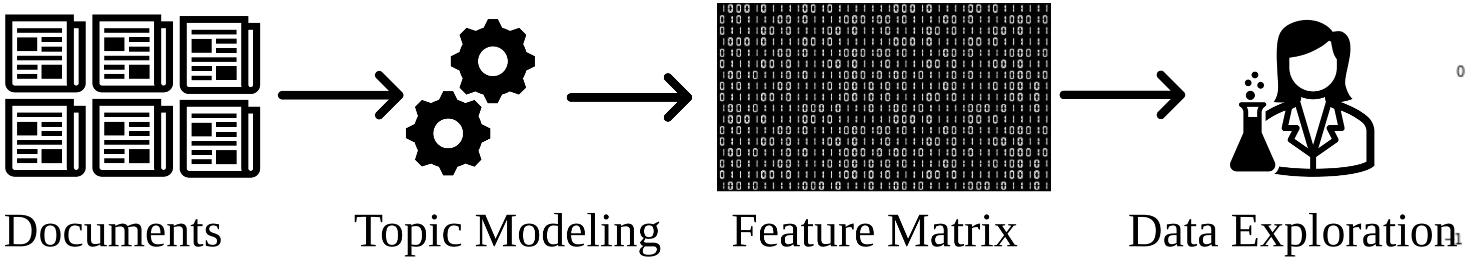


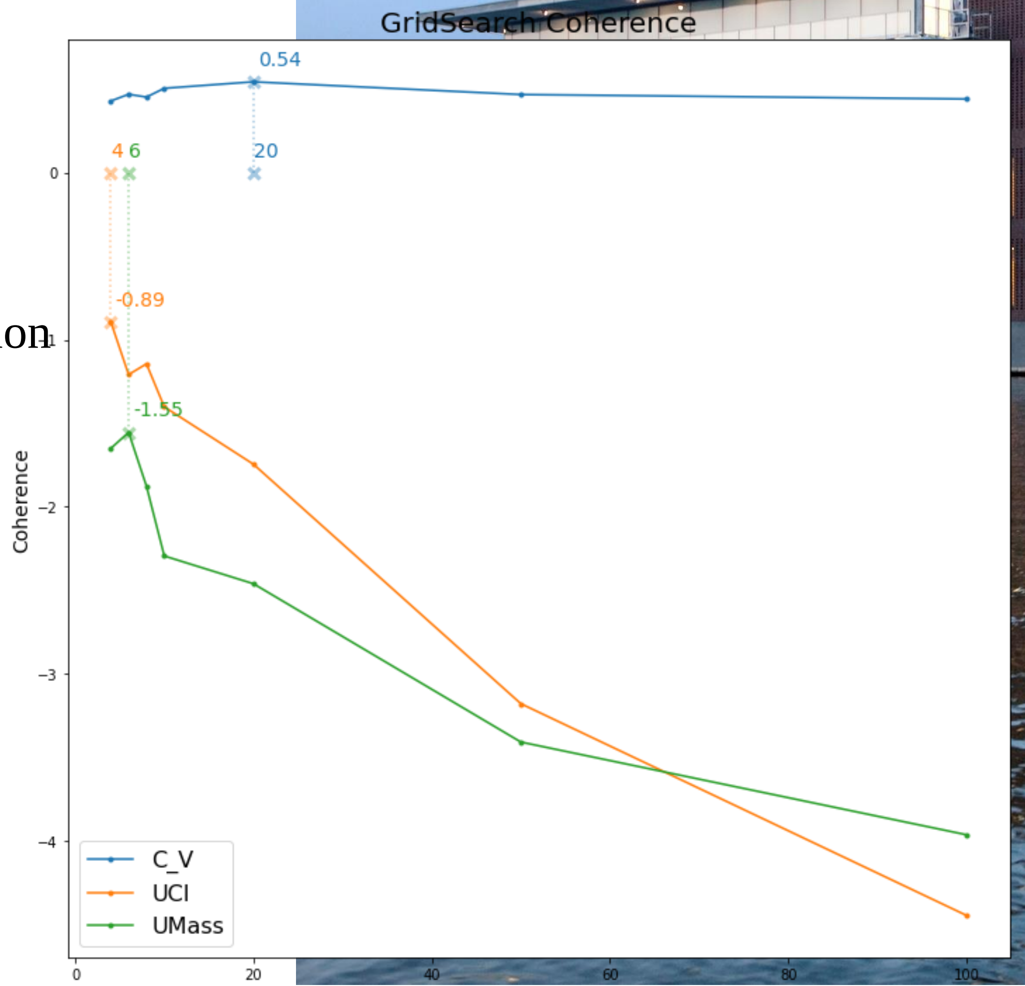
- **K** = additional Hyperparameter for GridSearchCV

UNIVERSITY OF AMSTERDAM
Economics & Business

LDA – Hyperparameter Evaluation

- **Intrinsic Evaluation**





- Topic Coherence metrics: Umass, CV, UCI
- Model Perplexity

- Try multiple values of K
- Select highest coherence (lowest perplexity)

Röder et al., "Exploring the space of topic coherence measures", WSDM 2015 PDF

LDA - Learning

- Discover Distribution
  - Of topics in documents
  - Of words in topics

- Machine Learning task
  - Python implementation in gensim or sklearn
  - Details of the learning task are out of scope

LDA

- See the notebook “LDA”


```
Topic # 3 (Internet            ): 0.135
Topic # 4 (Politics            ): 0.620
Topic # 7 (Movies              ): 0.042
Topic # 8 (Apple               ): 0.137
Topic # 9 (Video Technology    ): 0.057
-----------------------------------------
Sum                               0.992
```

Modern Topic Modeling

BERTopic

- Same Conceptual Framework
  - 1 topic = weighted words
  - 1 document = weighted topics

- Getting it through different method

Modern Topic Modeling

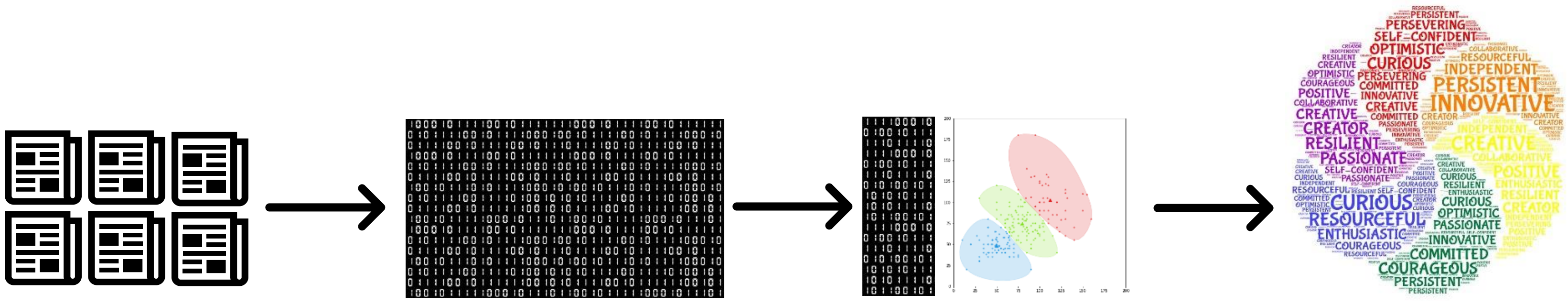
BERTopic

- https://maartengr.github.io/BERTopic/index.html
- Grootendorst "BERTopic: Neural topic modeling with a class-based TF-IDF procedure", 2022 Paper

1. Transform documents in vectors
2. Cluster vectors : 1 cluster = 1 topic
3. Word weights in topics = TF-IDF
    - 1 cluster = 1 document (concatenate all docs in cluster)
4. (optional) Use LLM to generate topic label

Modern Topic Modeling

BERTopic



Documents → Vectorizer → Dimensionality Reduction Clustering → Word Weights

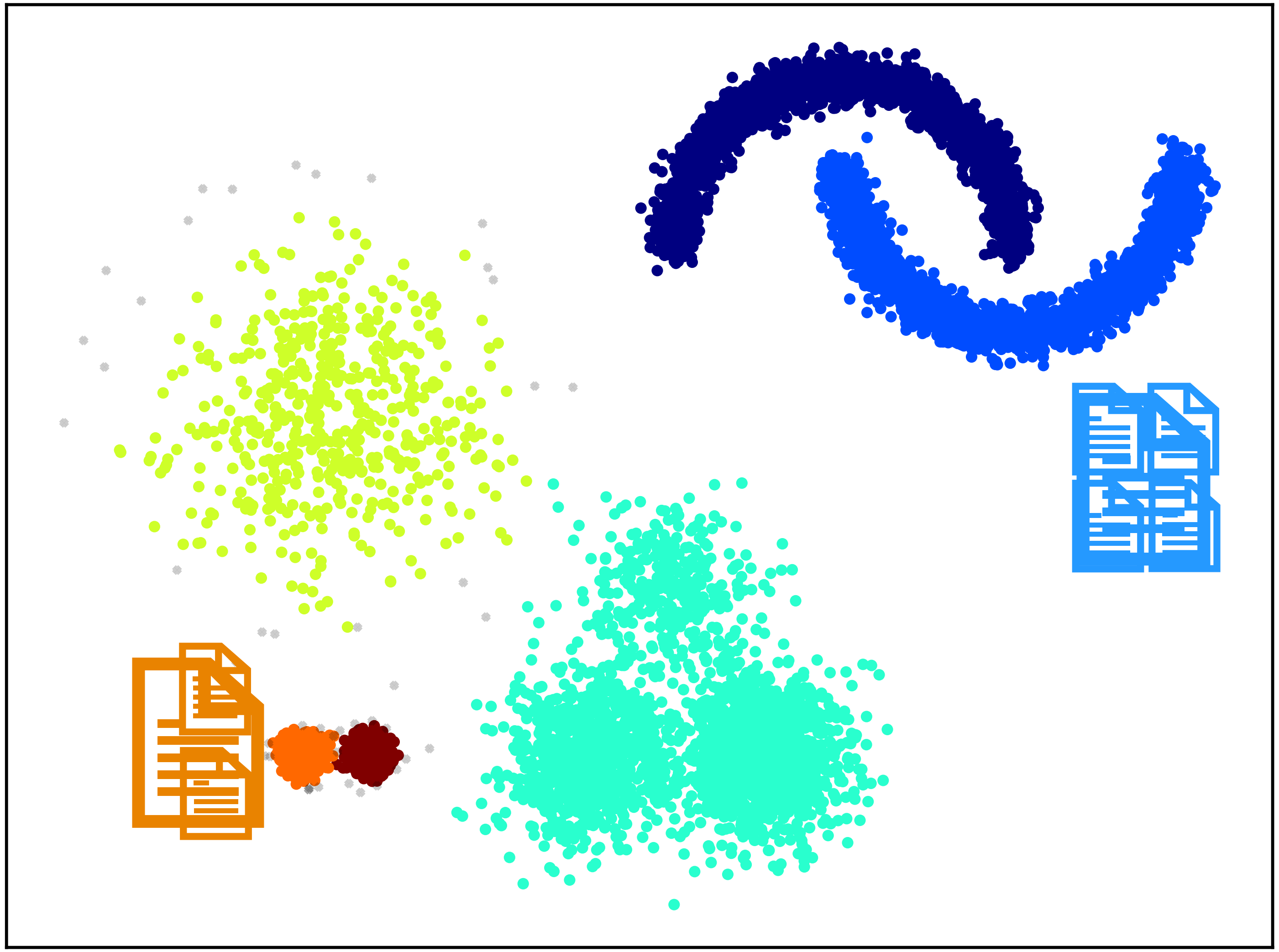
BERTopic

Clustering

- 6 clusters

Create Cluster-Based Corpus (CBC)

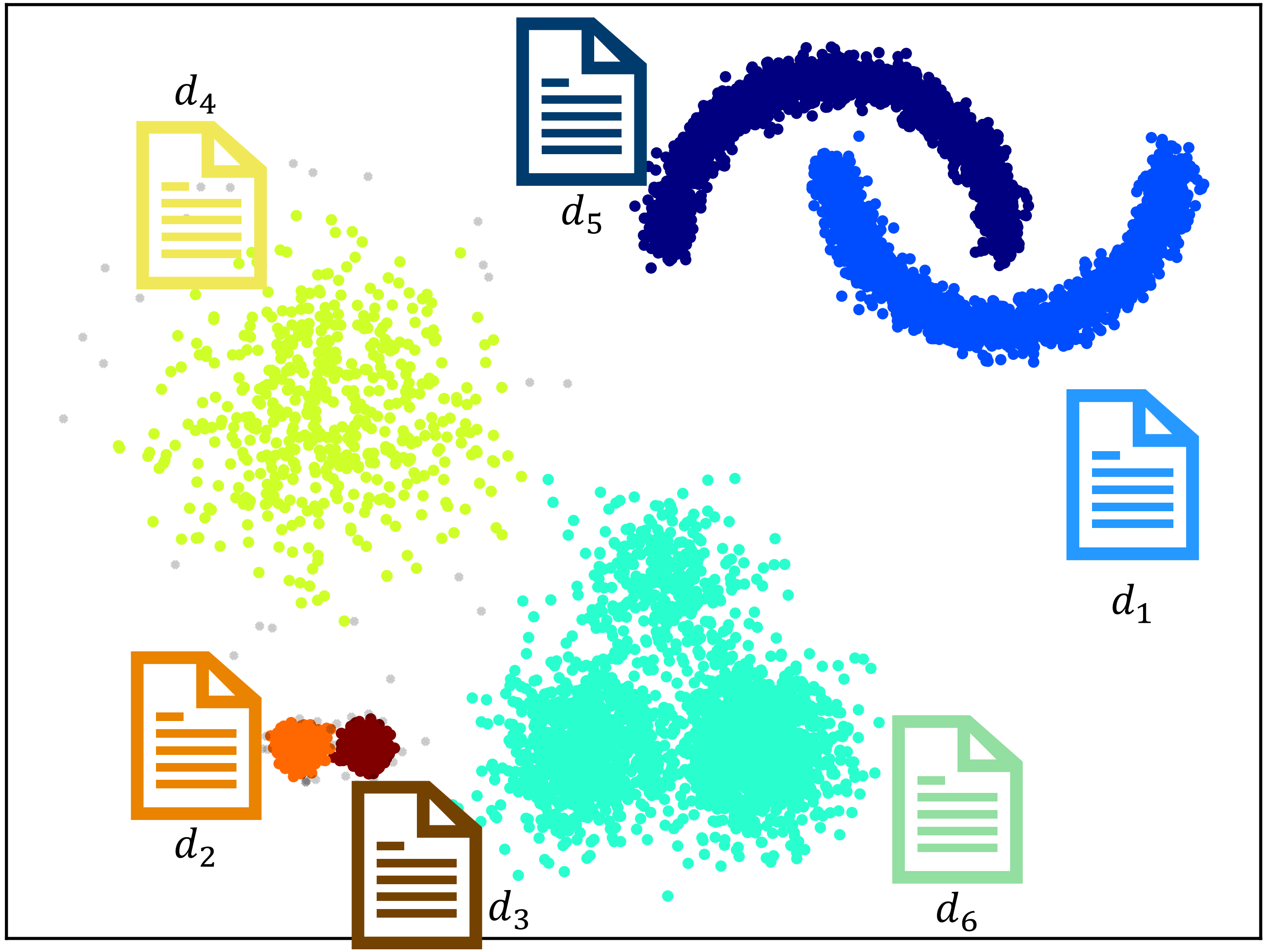
- 1 cluster = 1 doc
- CBC = 6 docs



BERTopic

Word Weights

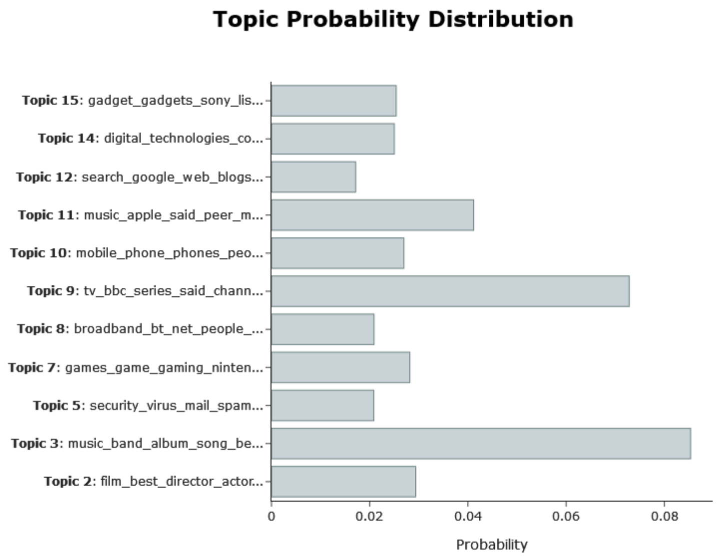
- CBC = 6 documents
- Word w / Topic k
- $w_k = \text{tfidf}(w, d_k, \text{CBC})$

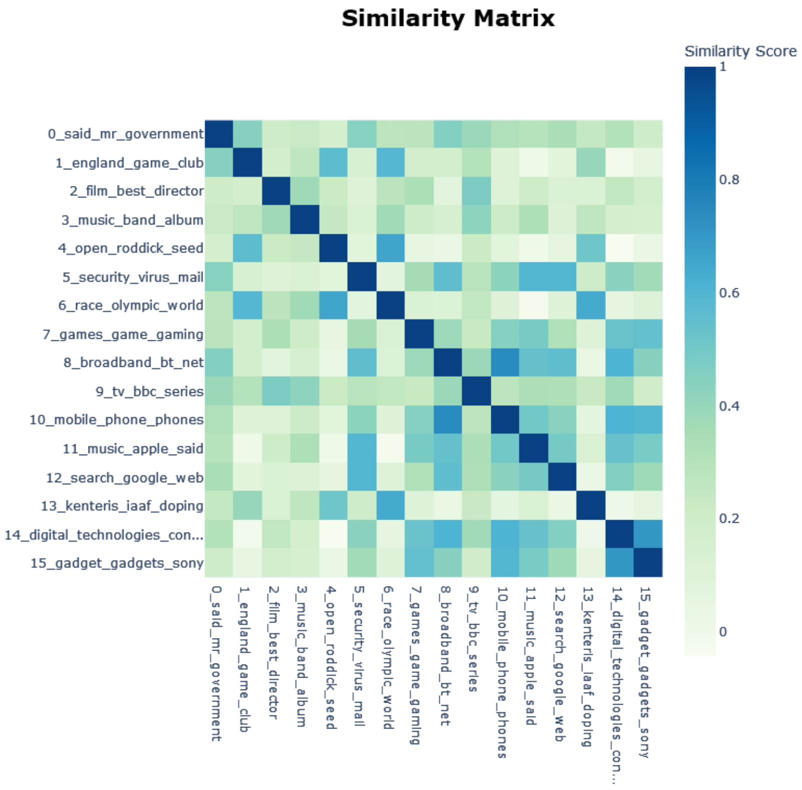


University of Amsterdam
Economics & Business

BERTopic

- See the notebook “BERTopic”





Take Away

- Document = Topics * Words
- **Topic Modeling** = learn from documents

- LDA
  - Statistical Learning
  - Human labeling

- Bertopic
  - Clustering

Prepare Tutorial

- **Read** these slides
- **Understand** the “by heart” concepts

- **Run** the attached notebooks
- **Update** your Python Env if needed