Topic Modeling

Emre Kavak, 01.06.21, Natural Language Processing – Methods and Applications

Chair of Software Engineering for Business Information Systems (sebis)
Faculty of Informatics
Technische Universität München
wwwmatthes.in.tum.de

Outline

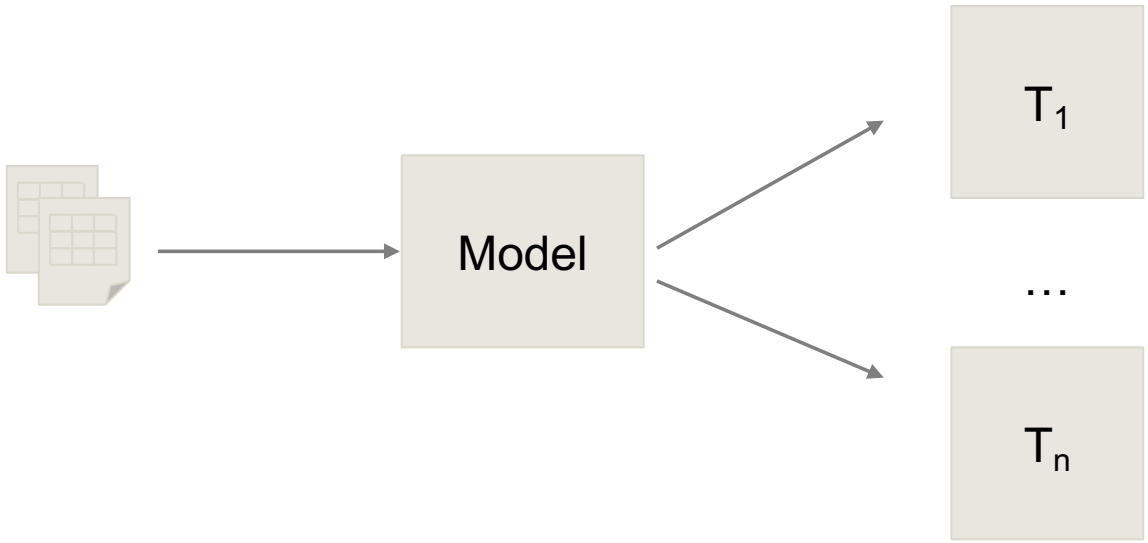
Problem Statement
- Motivation
- How to compare documents?

Topic Modeling (TM)
- From Dimensionality Reduction to Topics
- Generative Methods (pLSA, LDA)

Applications
- Data Exploration vs. Downstream Tasks
- Clustering Paper as Explanatory Example
  - TM manual tasks
  - TM Evaluation

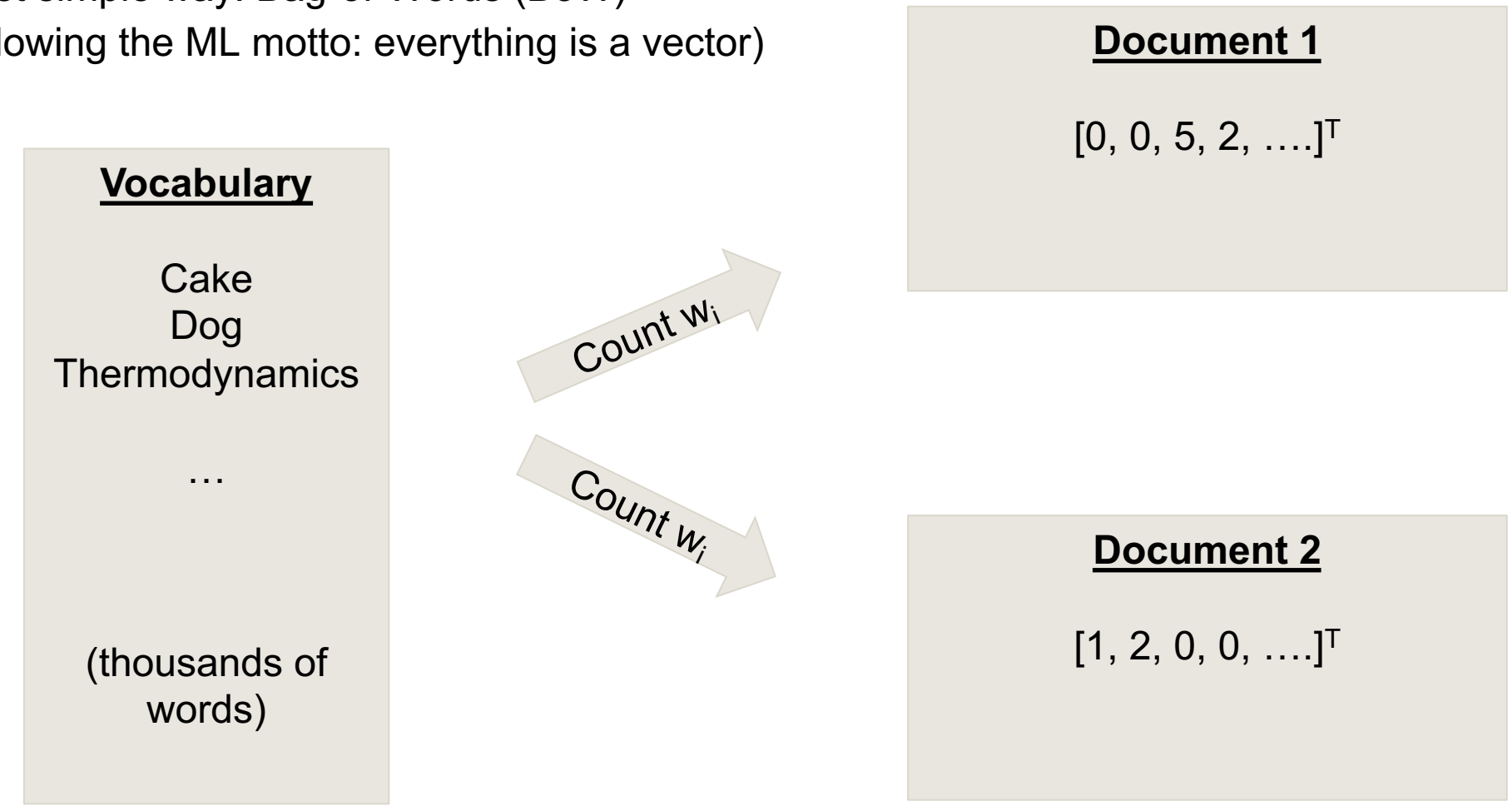
Problem Statement & Motivation

- Textual data is everywhere
- Everything is a “Document”
  - Wikipedia, books, Twitter comments, …



How To Represent Documents?

- Most simple way: Bag-of-Words (BoW)
- (following the ML motto: everything is a vector)



How To Identify Topics?

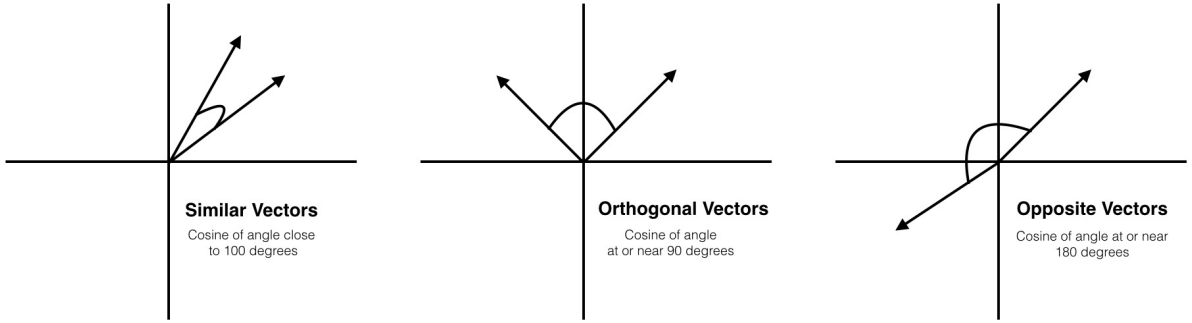
- Share **semantical** aspects
  - Texts about politics, sports, etc.
  - Higher order abstract features

- Naive variant: use similar words
  - Texts about sports may contain: referee, timeout, swim, …

Most Simple Procedure: Compare Word Occurrence

- Define own Metrics:
  - Count common words
  - Mark important words and count them
  - Etc.

- Or, Since documents are vectors:
  - **Cosine Similarity**: $< a, b > = \cos\theta \ \|a\| \ \|b\|$



https://www.oreilly.com/library/view/mastering-machine-learning/9781785283451/ba8bef27-953e-42a4-8180-cea152af8118.xhtml

Demo: Cosine Similarity

- Data taken from [1] (10kGNAD)

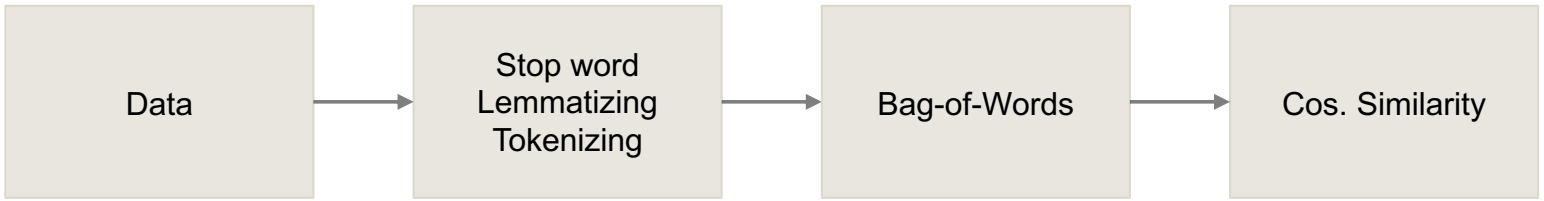
	label	text
0	Sport	21-Jähriger fällt wohl bis Saisonende aus. Wie...
1	Kultur	'Erfundene Bilder zu Filmen, die als verloren ...
2	Web	Der frischgekürte CEO Sundar Pichai setzt auf ...
3	Wirtschaft	Putin: "Einigung, dass wir Menge auf Niveau vo...
4	Inland	Estland sieht den künftigen österreichischen P...
5	Wirtschaft	Der Welser Stempelhersteller verbreitert sich ...
6	Sport	Traditionsklub setzt sich gegen den FC Utrecht...
7	Etat	Finanzausschuss tagte Montag: Konfliktthemen S...
8	International	Militär setzt Offensive an Grenze zu Afghanist...
9	Sport	Abschiedstournee für Guardiola beginnt beim HS...

```python
class MyAnalyzer(object):

    def __init__(self):
        self.tokenizer = RegexpTokenizer(r'\w+')
        self.tagger = ht.HanoverTagger('morphmodel_ger.pgz')
        self.nlp = spacy.load('de_core_news_md')
        self.german_stop_words = stopwords.words('german')

    def __call__(self, doc):
        doc = [t.lemma_ for t in self.nlp(doc)]
        doc = list(map(str.lower,doc))

        doc = [token for token in doc if not token.isnumeric()]
        doc = [token for token in doc if len(token) > 1]
        doc = [w for w in doc if not w in self.german_stop_words]

        return doc
```

```python
doc_dot_prod = docs_np@docs_np.T
```

```python
norms = np.sqrt(np.sum(docs_np**2,axis=1)).reshape(norms.shape[0],-1)
```

```python
cos_sim_matrix = (1/norms.T)*doc_dot_prod/norms
```



Demo: Cosine Similarity

Similarity matrix:

$$\begin{matrix} sim(d_1, d_1) & \dots & sim(d_1, d_n) \\ \dots & \dots & \dots \\ sim(d_n, d_1) & \dots & sim(d_n, d_n) \end{matrix}$$

Query: give docs ordered by similarity to d_{42} (most similar first)

```
comp_ex = np.argsort(cos_sim_matrix[42,:])
comp_ex[::-1]
```

```
array([ 42, 289, 169,  41, 319,  94, 224, 133,   0, 103, 258, 474, 115,
        51, 117, 436, 161, 437, 241, 269,  96, 490,  28, 370, 498, 446,
       130, 237, 145, 127, 343, 153,  10, 284, 236,  87, 412, 252, 361,
       288,  29,   9, 283,   1, 324, 431, 312, 286, 313,  50,  47, 210,
```

d_{42}

```
docs[42][:400]
```

```
'Forscher veröffentlichen sieben Fachartikel mit neuesten Erkenntnissen des Landers Philae über Tschurjumow-G
erassimenko. Göttingen/Wien – Zuletzt hatte ihm Pluto etwas die Show gestohlen, aber vergessen ist Tschurjumo
w-Gerassimenko, Zielkomet der Rosetta-Mission, keineswegs. Während Wissenschafter auf weitere Datenpakete vom
Zwergplaneten warten, die über Monate hinweg portionsweise eintreffen wer'
```

d_{289}

```
docs[289][:400]
```

```
'Britische Wissenschafter haben analysiert, ob Pflanzenproben in naturhistorischen Museen auch die richtigen
Namen tragen. Die Ergebnisse sind erschreckend. Oxford/Wien – Um die Verteilung von bestimmten Tier- oder Pfl
anzenarten rund um den Globus zu analysieren, greifen Biodiversitätsforscher gerne auf naturhistorische Samml
ungen zurück. Bei solchen Analysen gelingt es immer wieder, neue Arten zu '
```

d_{169}

```
docs[169][:400]
```

```
'Protonenstrahlen ermöglichen bis zu einer Milliarde Kollisionen pro Sekunde. Bern – Mittlerweile läuft er na
ch seiner Aufrüstung, die 27 Monate lang dauerte, bereits wieder seit einigen Wochen. Doch erst seit Mittwoch
früh wird am Large Hadron Collider (LHC) wieder richtig Physik betrieben, sprich: Es werden Daten gesammelt u
nd ausgewertet. Nach mehr als zwei Jahren kollidieren am CERN  wieder die '
```

Cosine Similarity with BoWs

- It works: we can **retrieve** similar documents given a **query** document
- But we have problems:

Synonymy

$$\cos\left(\begin{pmatrix}1\\0\end{pmatrix},\begin{pmatrix}0\\1\end{pmatrix}\right)=0$$

Let w_1=cat, w_2=feline

Polysemy

$$\cos\left(\begin{pmatrix}1\\0\end{pmatrix},\begin{pmatrix}1\\0\end{pmatrix}\right)=1$$

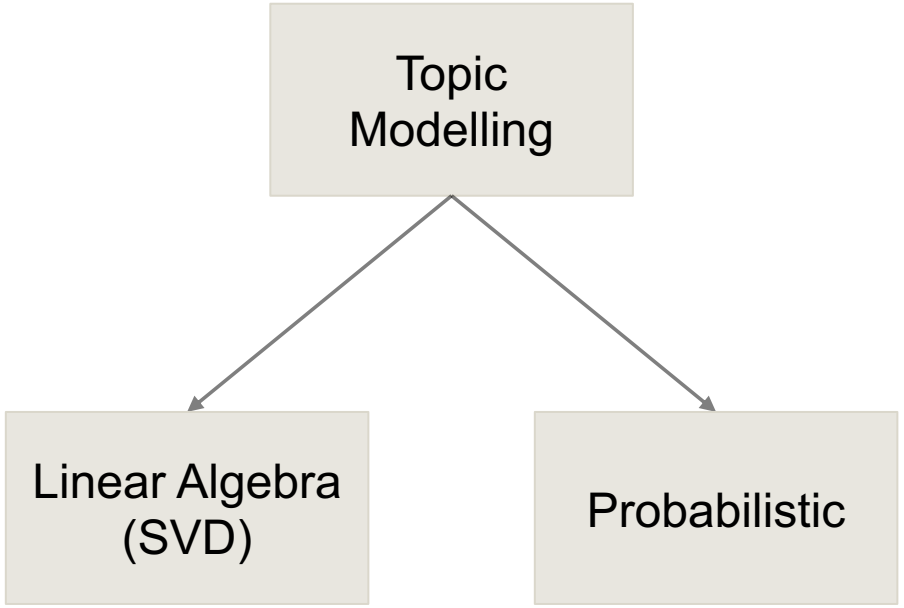
Let w_1=bank; but bank as institution vs bank building

Sparsity

$(1,0,0,\ldots,1,0,0,\ldots)^T$

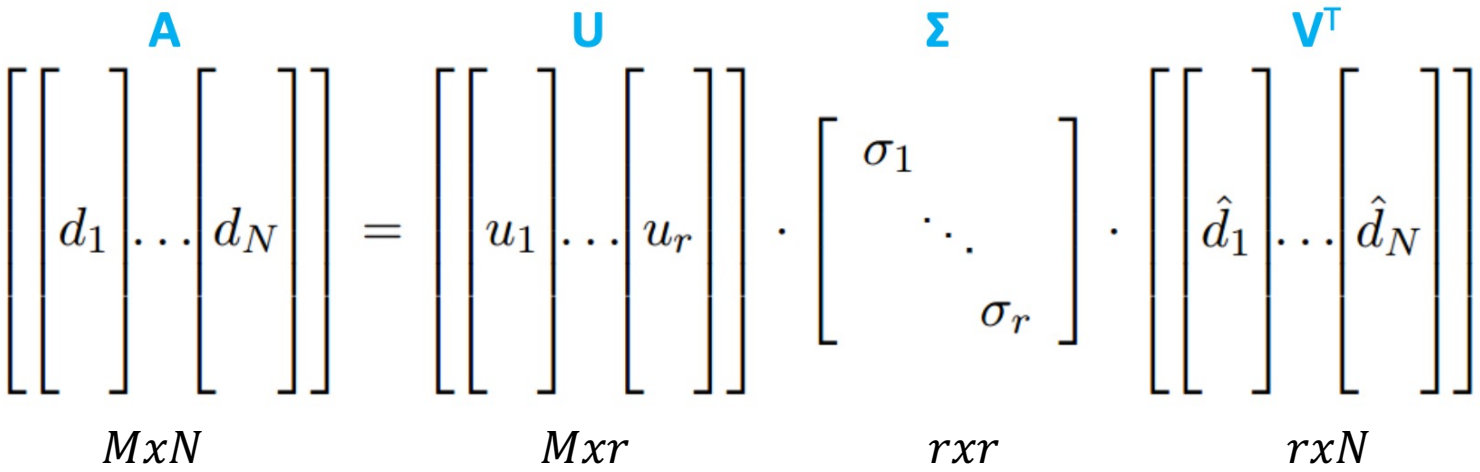
- What can we learn from this simple approach?
  - Give intuition about **Vector Space Model** (VSM) → docs as vectors
  - Can be used as a **clustering** method (pretty close to topics)
  - Show problems of **synonymy, polysemy, sparsity**
  - Coming Topic Modelling methods are in same VSM

We can do better



LSA – Latent Semantic Analysis [2]

- Based on (compact) SVD (Singular Value Decomposition)
- Every mxn-Matrix can be decomposed into:



- Exists for all complex matrices
- If matrix real: Singular Values σ_i are real and **non-negative**
- Ordered: $\sigma_1 > \sigma_2 > \cdots > \sigma_r$

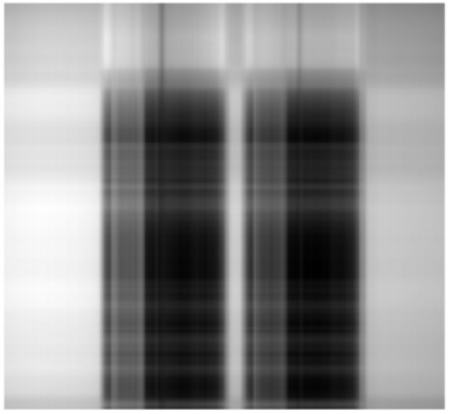
What is SVD good for?

- (Hidden) **Basis Transformations**
- Dimensionality **Reduction**
  - Keeping only bases with **highest information**/signal → keep only largest k σ_i
  - (analogue to PCA; difference in normalization constant)
- Solving linear equations and more

```
imgs = []
for k in [100, 10, 2]:
    imgs.append(U[:,:k]@np.diag(D[:k])@V[:k,:])
```

```
fig, axs = plt.subplots(1, 3,figsize=(15,15))

for i in range(3):
    axs[i].imshow(imgs[i],cmap='gray')
    axs[i].axis('off')
```

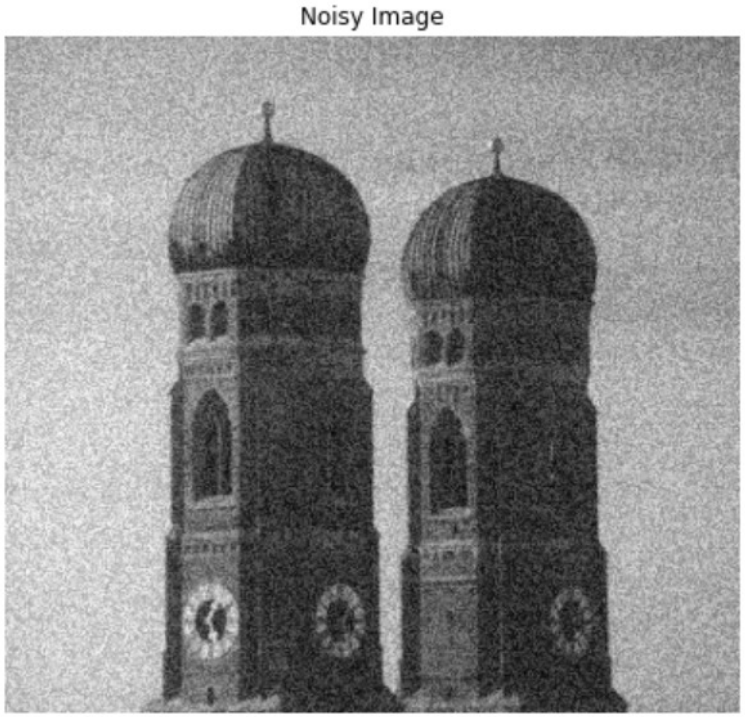




https://commons.wikimedia.org/w/index.php?search=frauenkirche&title=Special:MediaSearch&go=Go&type=image&fileres=%3C500

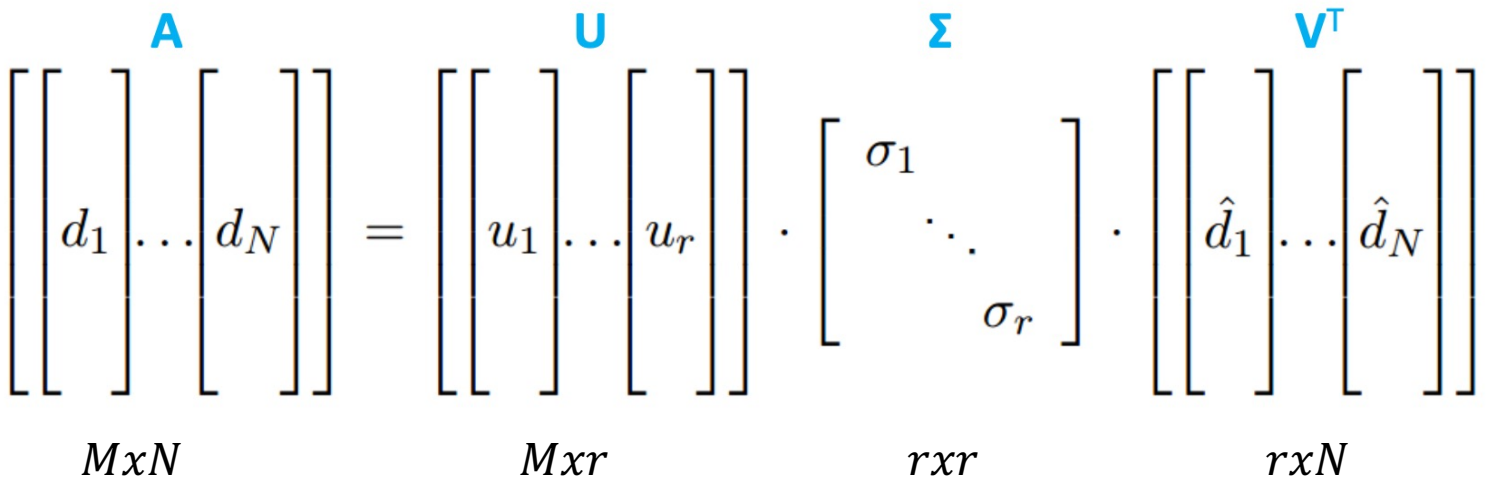
What is SVD good for?

- And Denoising:
  - Throw away less important signals (low variance)



https://commons.wikimedia.org/w/index.php?search=frauenkirche&title=Special:MediaSearch&go=Go&type=image&fileres=%3C500

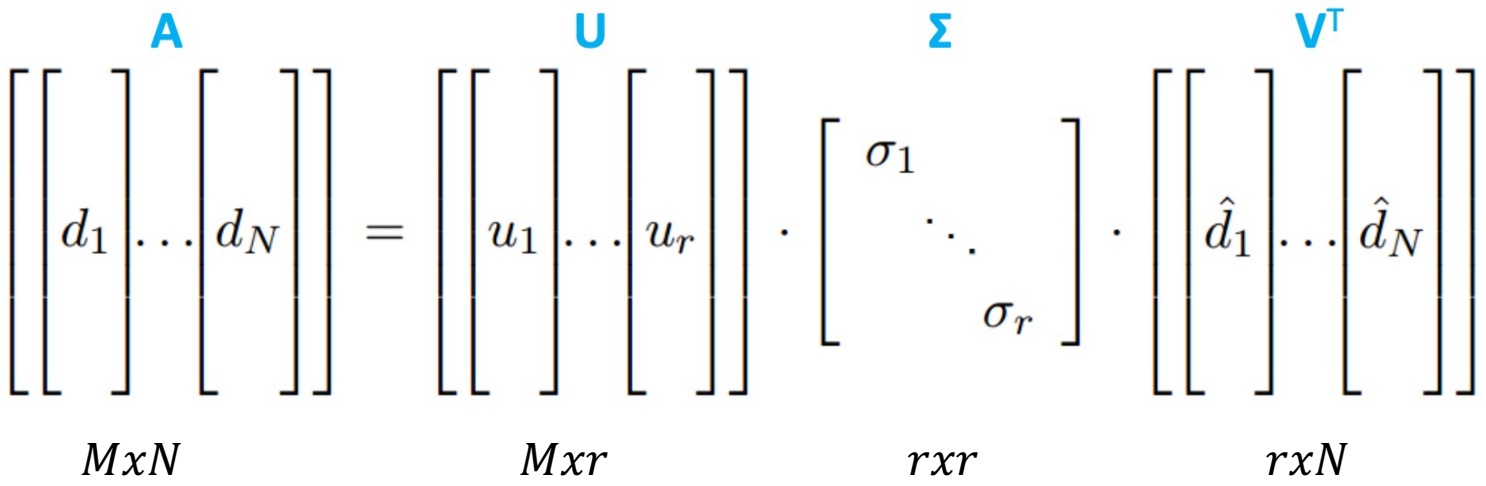
LSA – Latent Semantic Analysis [2]



- M: # terms/vocab
- N: # docs
- r: rank of A

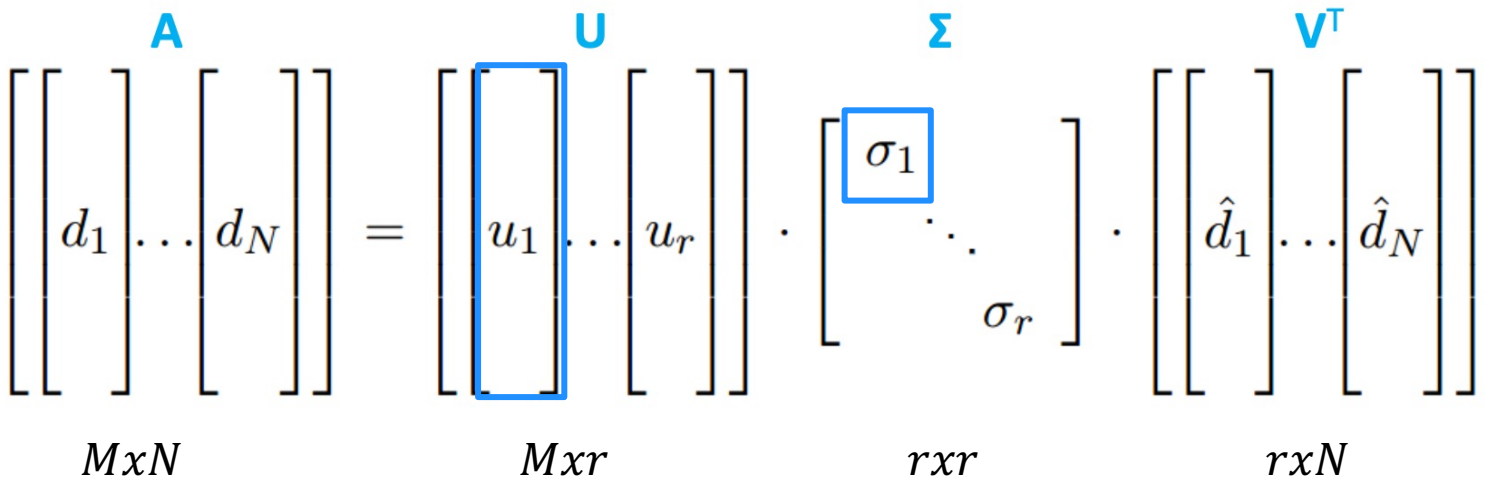
→ So far, this is only mathematics (linear algebra)

LSA – Latent Semantic Analysis [2]



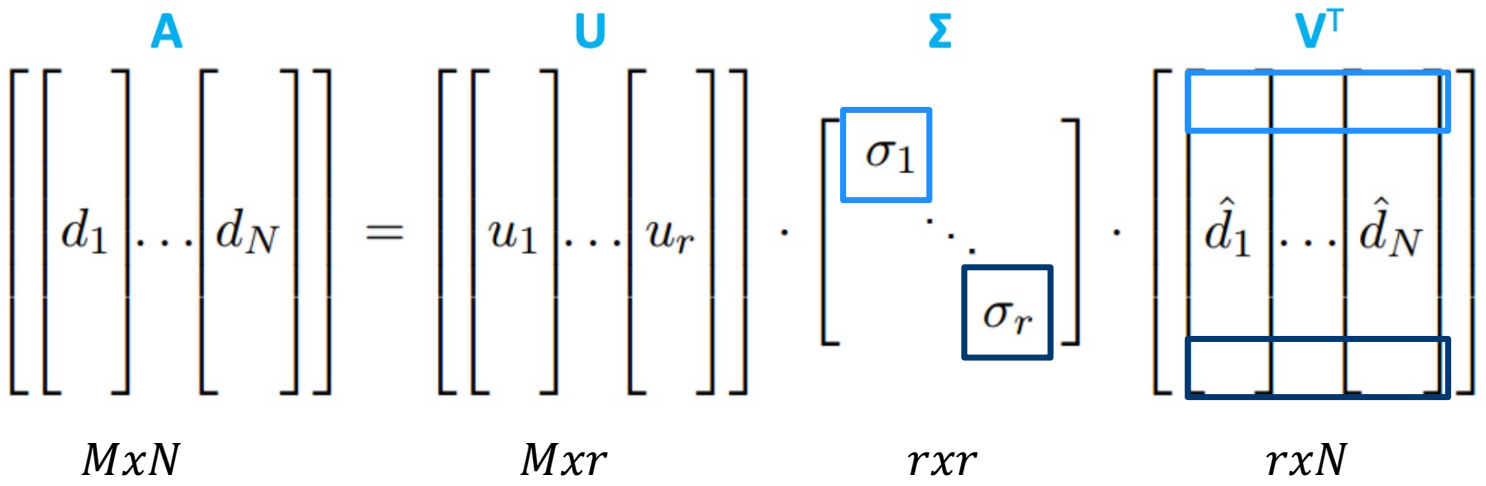
- **Possible Interpretation** of SVD from BoW:
  - Look at indices: M:#vocab, N: #docs, r: dimension latent space
  - Σ defines latent **topic** space of dimension r
    - Each σ_i is weight for hidden topic i

LSA – Latent Semantic Analysis [2]



- **Possible Interpretation** of SVD from BoW:
  - Look at indices: M:#vocab, N: #docs, r: dimension latent space
  - Σ defines latent **topic** space of dimension r
    - Each σ_i is weight for hidden topic i
  - U: σ_i is related to u_i
    - Topic i contains words in u_i → u_i is topic vector i

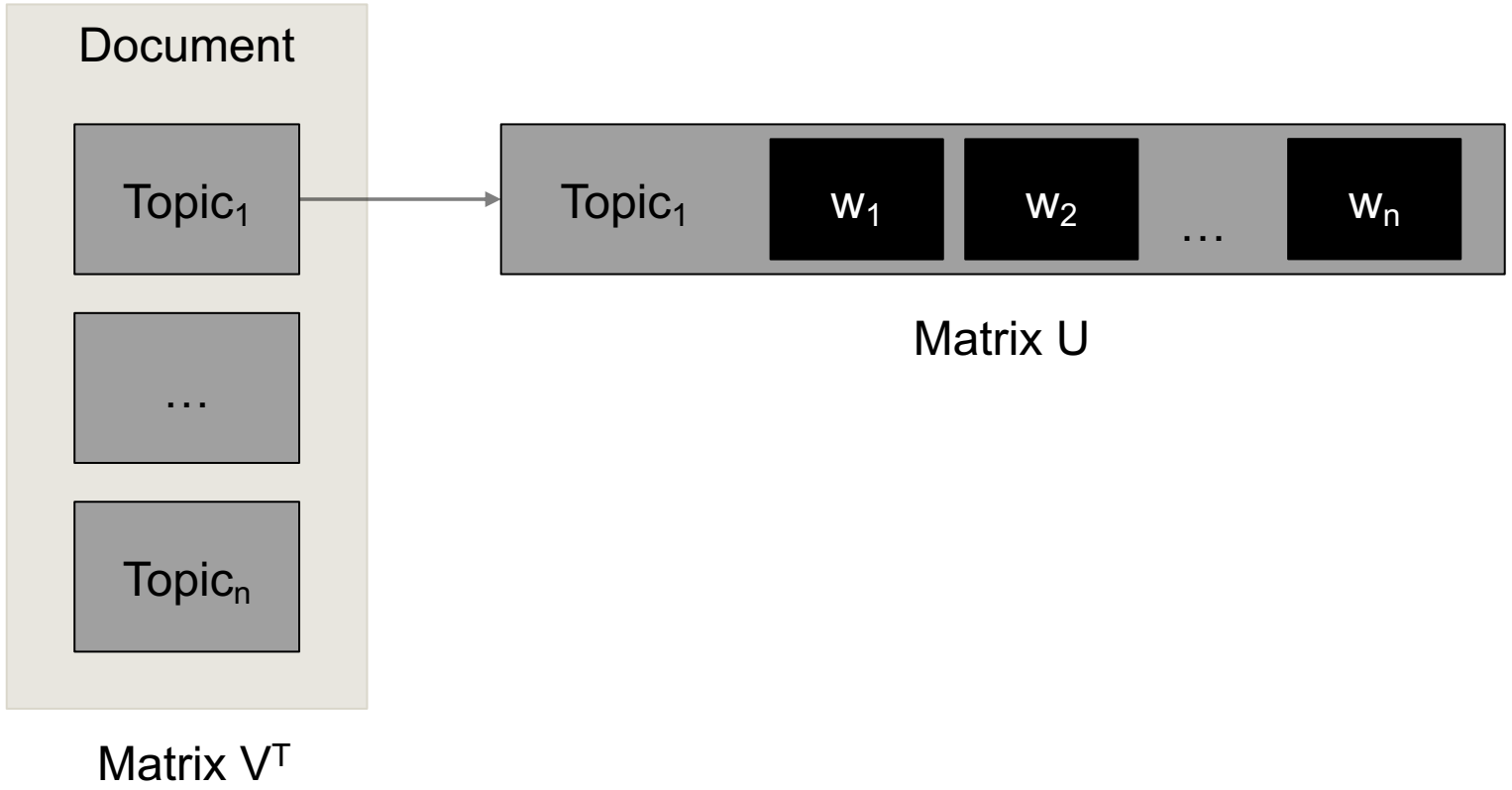
LSA – Latent Semantic Analysis [2]



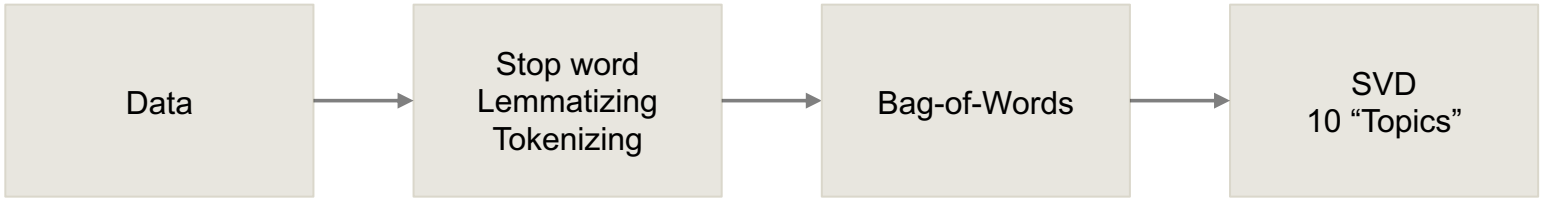
- **Possible Interpretation** of SVD from BoW:
  - Look at indices: M:#vocab, N: #docs, r: dimension latent space
  - Σ defines latent **topic** space of dimension r
    - Each σ_i is weight for hidden topic i
  - U: σ_i is related to u_i
    - Topic i contains words in u_i → u_i is topic vector i
  - V^T: $\hat{d}_i$ is related to $\sigma_1, \sigma_2, \ldots, \sigma_r$
    - Document $\hat{d}_i$ is combination of r topics

LSA – Topic Model Definition

- We implicitly got:
  - Documents are “mixtures” of topics
  - Topics are “mixtures” of words
  - # Topics (dimension of diagonal matrix) is only **hyperparameter**



LSA – Demo: Topics



	Word 0	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	
Topic 0	geben	jahr	mehr	kommen	gut	prozent	gehen	
Topic 1	prozent	euro	million	jahr	milliarde	mio.	hoch	Finance, economy
Topic 2	flüchtling	mensch	grenze	land	deutschland	polizei	laut	Refugee Crisis
Topic 3	euro	million	jahr	mio.	milliarde	waffe	laut	
Topic 4	erst	minute	tor	spiel	punkt	gut	zwei	Soccer (sports)
Topic 5	euro	flüchtling	geben	gut	spö	ja	övp	
Topic 6	spö	partei	faymann	övp	fpö	regierung	häupl	Politics
Topic 7	apple	neu	weit	neue	nutzer	gerät	etwa	Tech
Topic 8	syrien	is	geben	mensch	usa	euro	waffe	War
Topic 9	hofer	land	gut	van	bellen	apple	usa	

LSA – Demo: Document Vectors

	Word 0	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6
Topic 0	geben	jahr	mehr	kommen	gut	prozent	gehen
Topic 1	prozent	euro	million	jahr	milliarde	mio.	hoch
Topic 2	flüchtling	mensch	grenze	land	deutschland	polizei	laut
Topic 3	euro	million	jahr	mio.	milliarde	waffe	laut
Topic 4	erst	minute	tor	spiel	punkt	gut	zwei
Topic 5	euro	flüchtling	geben	gut	spö	ja	övp
Topic 6	spö	partei	faymann	övp	fpö	regierung	häupl
Topic 7	apple	neu	weit	neue	nutzer	gerät	etwa
Topic 8	syrien	is	geben	mensch	usa	euro	waffe
Topic 9	hofer	land	gut	van	bellen	apple	usa

	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
Doc 0	2.627866	-0.608375	-0.763013	0.120890	1.251656	-0.118058	-0.490042	0.627022	-0.216631	0.517732
Doc 1111	2.427030	-0.477143	-0.545306	0.080579	-0.227034	0.466506	-0.606716	3.106189	-1.244786	1.259009

d_0

```
docs_.iloc[0]["text"][:400]
```

```
'21-Jähriger fällt wohl bis Saisonende aus. Wien – Rapid muss wohl bis Saisonende auf Offensivspieler Thomas
Murg verzichten. Der im Winter aus Ried gekommene 21-Jährige erlitt beim 0:4-Heimdebakel gegen Admira Wacker
Mödling am Samstag einen Teilriss des Innenbandes im linken Knie, wie eine Magnetresonanz-Untersuchung am Don
nerstag ergab. Murg erhielt eine Schiene, muss aber nicht operiert werden.'
```

d_{1111}

```
docs_.iloc[1111]["text"][:400]
```

```
'Finale Erweiterung des Rollenspiels verspricht jede Menge Neuerungen und Verbesserungen. Blood and Wine, die
letzte Erweiterung des 2015 erschienenen Rollenspiels The Witcher 3, wird am 31. Mai in den Handel kommen. Da
s gab Hersteller CD Projekt Red in einer Aussendung bekannt. Die Entwickler versprechen mehr als 90 neue Ques
ts und mehr als 30 Stunden neue Abenteuer. Die größte Erweiterung stellt '
```

Probabilistic Methods – Generative Models

pLSA – Probabilistic LSA [3]

- Modelling Topics explicitly:
  - Every document is a mixture of topics
  - Every topic is is a mixture of words

→ $p(w,d) = p(d)p(w|d)$,

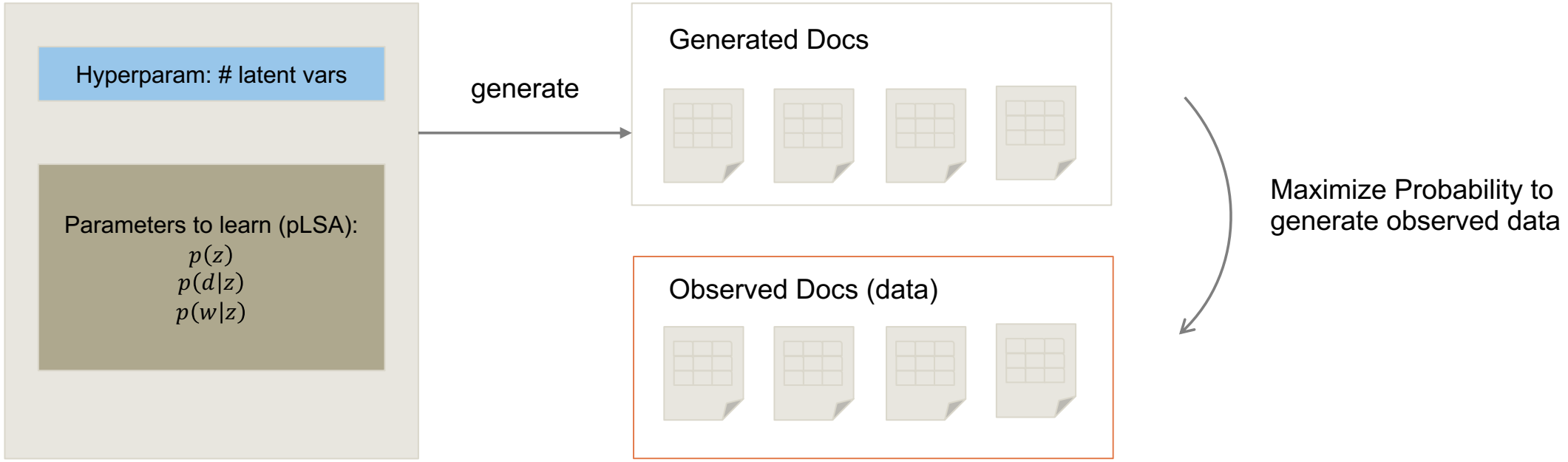
$$where\ p(w|d) = \sum_{z \in Z} p(w|z)p(z|d)$$

  - Latent Random Variable Z introduced
    - Controlled by hyperparameter k (number of topics)
    - Multinomial distribution of words

LDA – Latent Dirichlet Allocation [4]

- Similar to pLSA
- Does not model document $p(d)$ directly (good)
- Introduces (conjugate) priors to:
  - Latent variable distributions
    - Advantage: only **few topics** are touched in a document
  - Word distributions given topic
    - Advantage: only **few words** revolve around given topic
- More complex model; more flexible
- Fully Bayesian

Generative Framework – Learning Model



- pLSA: maximize **likelihood** [3] (point estimation)
  - Expectation Maximization (**EM**)
- LDA: **Bayesian** inference (integrate over entire parameter domains) [4]
  - Intractable, therefore:
    - **Variational inference** (approximate posterior by new distribution)
    - Sampling (**Gibbs**, MCMC)

LDA – Demo

	Word 0	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6
Topic 0	(0.039359946, tor)	(0.034987714, minute)	(0.030292021, salzburg)	(0.026125848, punkt)	(0.022105083, spiel)	(0.020965936, sieg)	(0.020837002, trainer)
Topic 1	(0.08016918, prozent)	(0.064049184, euro)	(0.039858047, million)	(0.019963624, million_euro)	(0.016110672, unternehmen)	(0.015092909, land)	(0.015066322, milliarde)
Topic 2	(0.042920977, spö)	(0.036138527, övp)	(0.030767983, fpö)	(0.019816885, partei)	(0.017873146, grüne)	(0.015347474, wahl)	(0.014539326, gesetz)
Topic 3	(0.09209431, eu)	(0.021308068, staat)	(0.019834965, euro)	(0.01858932, land)	(0.018140338, griechenland)	(0.018048353, milliarde)	(0.016686933, regierung)
Topic 4	(0.023142163, welt)	(0.021656757, film)	(0.021541215, buch)	(0.021300955, zeit)	(0.018437088, bild)	(0.014481175, mensch)	(0.0144333085, geschichte)
Topic 5	(0.03472652, apple)	(0.03157378, dollar)	(0.02470762, gerät)	(0.020432277, unternehmen)	(0.019971298, android)	(0.016845152, us)	(0.016498182, hersteller)
Topic 6	(0.040164616, flüchtling)	(0.03825454, mensch)	(0.020286156, polizei)	(0.012432331, leben)	(0.0123546105, deutschland)	(0.012214937, staat)	(0.010855782, angabe)
Topic 7	(0.05566558, spiel)	(0.034216553, facebook)	(0.026252259, spieler)	(0.023297368, rennen)	(0.01852375, team)	(0.01830212, woche)	(0.017015882, titel)
Topic 8	(0.026309414, orf)	(0.024537938, regierung)	(0.0200205, syrien)	(0.019832687, land)	(0.019215338, russland)	(0.01903749, us)	(0.017703969, usa)
Topic 9	(0.035649814, nutzer)	(0.032242548, microsoft)	(0.029029362, berlin)	(0.024171364, programm)	(0.024016391, datum)	(0.019385861, problem)	(0.01788145, unternehmen)

How Are The Mentioned Problems Resolved?

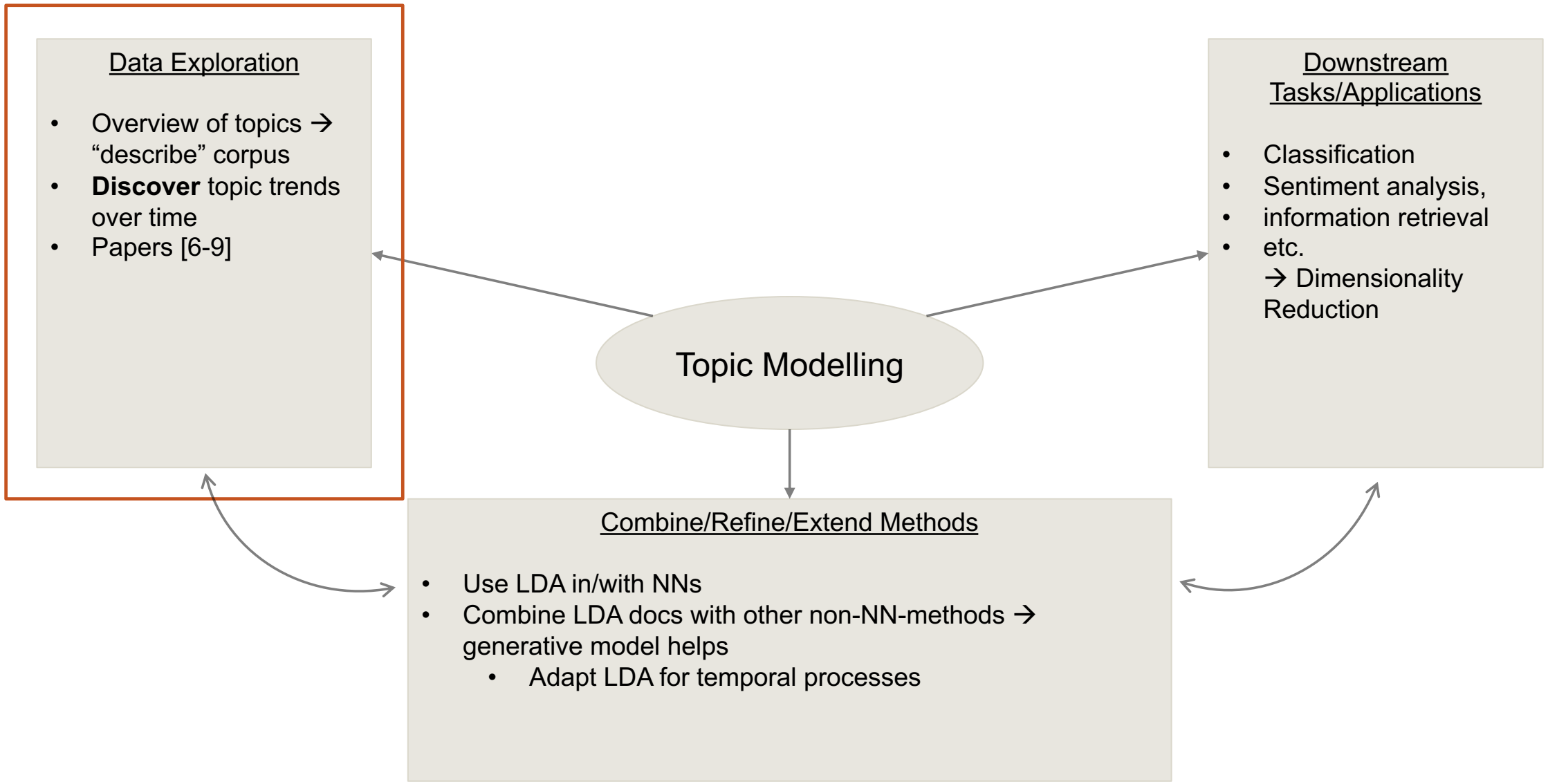
	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
Doc 0	2.627866	-0.608375	-0.763013	0.120890	1.251656	-0.118058	-0.490042	0.627022	-0.216631	0.517732
Doc 1111	2.427030	-0.477143	-0.545306	0.080579	-0.227034	0.466506	-0.606716	3.106189	-1.244786	1.259009

- Sparsity: resolved
- Synonymy: partially resolved
  - Similar words contribute to same topics
- Polysemy: partially resolved
  - Same word can contribute to different topics

  → Different topics reflecting different meanings

  → Same word associated with different meanings

Applications

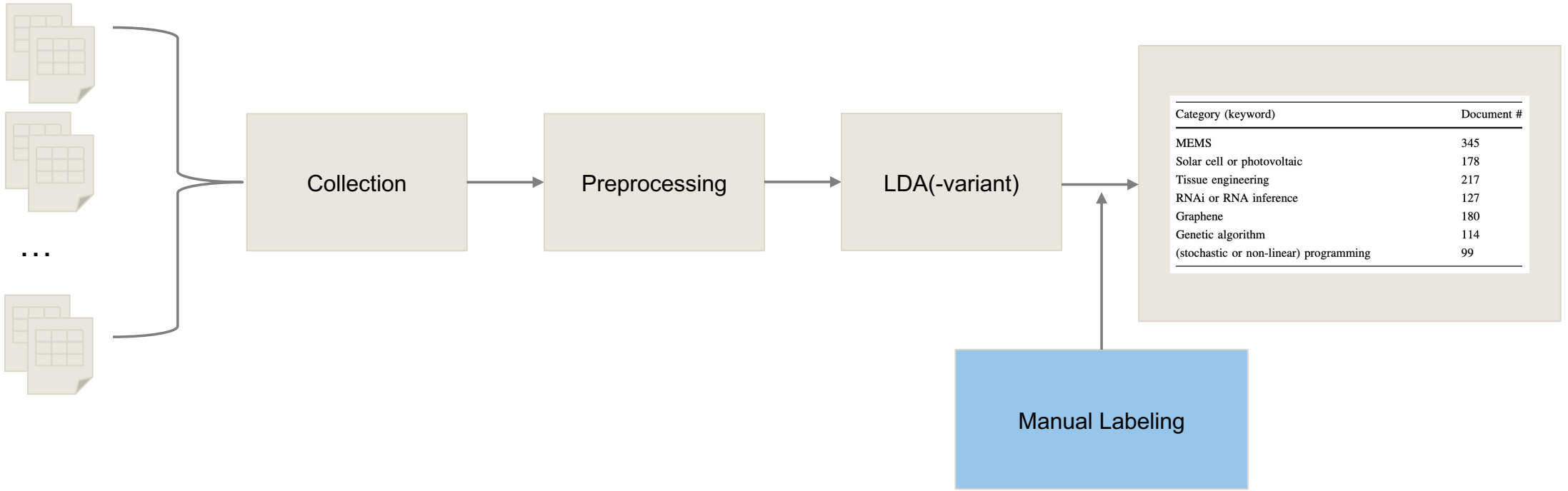


Applications

- Application fields vary
  - Bioinformatics, social/political sciences, medicine [6-10, 22]
  - Heavily used in research bibliometrics
- Short review on: Clustering scientific documents with topic modeling [21]
  - Involves exploratory aspects

Clustering scientific documents with topic modeling [21]

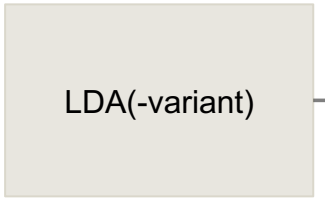
- Task
  - Cluster documents from 7 different fields
  - Test different LDA versions (LDA, and hierarchical versions)



Category (keyword)	Document #
MEMS	345
Solar cell or photovoltaic	178
Tissue engineering	217
RNAi or RNA inference	127
Graphene	180
Genetic algorithm	114
(stochastic or non-linear) programming	99

Clustering scientific documents with topic modeling [21]

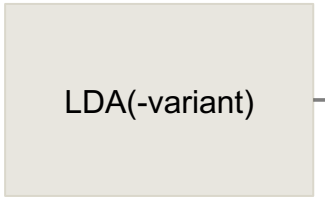
- Manual Labeling



Output topics

Topic 22 Genetic algorithm	Topic 24 Stochastic	Topic 49 Stochastic	Topic 14 Solar cell
Design	Model	Optimization	Films
Optimization	Results	Model	Thin
Genetic	Experimental	Approach	Tio
Algorithm	Analysis	Process	Silicon
Method	Based	Method	Surface

- Document to Cluster mapping
  - Take topic with p(t) > 50% as cluster (if there is none: do not assign)



Output document

Topic	Proportion (%)
22	58
24	17
49	16
14	5.4

Clustering scientific documents with topic modeling [21]

- Results of paper:
  - HDP is best variant

	F score average	*F* score st. dev.
K-means	0.66	0.30
LDA	0.71	0.04
CTM	0.72	0.07
hLDA	Not computed	Not computed
HDP	0.90	0.04

- How to evaluate Topic Models?
  - **Intrinsic**
    - Perplexity
      - ”Is generation of test data likely under model?”
    - Coherence
      - “Do word pairs in one topic appear together in docs?”
    - Human evaluation
      - TM is a human-centric approach
  - **Extrinsic**
    - Task dependent

Summary

- Topic Modelling as matrix factorization – LSA
- Model latent topics explicitly – pLSA, LDA
- Topic Models help in visualization and exploration
- Many extensions exist (NN-based, temporal extensions, hierarchical)

Sources

Data [1]:

- Subsampled: https://tblock.github.io/10kGNAD/
- Original: https://ofai.github.io/million-post-corpus/

Paper:

- [2] Indexing by Latent Semantic Analysis, Deerwester et al.
- [3] Probabilistic Latent Semantic Analysis, Hofmann
- [4] Latent Dirichlet Allocation, Blei et al.
- [5] A Survey on Bayesian Deep Learning, Wang et al.
- [6] A bibliometric study of research topics, collaboration and centrality in the Iterated Prisoner's Dilemma, Glynatsi et al.
- [7] What is the performance in public hospitals? A longitudinal analysis of performance plans through topic modeling, Noto et al.
- [8] 5335 days of Implementation Science: using natural language processing to examine publication trends and topics, Scaccia et al.
- [9] Political marketing with data analytics, Mathaisel et al.
- [11] Integrating Document Clustering and Topic Modeling, Xie et al.
- [12] Applications of Topic Models, Boyd-Graber
- [13] Neural Variational Inference for Text Processing
- [15] Improving Distributed Word Representation and Topic Model by Word-Topic Mixture Model, Fu et al.
- [14] TM-LDA: Efficient Online Modeling of the Latent Topic Transitions in Social Media, Wang et al.
- [16] Nested Variational Autoencoder for Topic Modeling on Microtexts with Word Vectors, Trinh et al.
- [10] Clustering and Topic Modelling: A New Approach for Analysis of National Cyber security Strategies, Kolini et al.
- [17] Topic Modeling with Wasserstein Autoencoders, Nan et al.
- [18] AUTOENCODING VARIATIONAL INFERENCE FOR TOPIC MODELS, Srivastava et al.
- [19] Topic Modeling in Embedding Spaces, Dieng et al. → review of topic word models and coupled new model
- [20] A network approach to topic models, Gerlach et al.
- [21] Clustering scientific documents with topic modeling, Yau et al.
- [22] An overview of topic modeling and its current applications in bioinformatics, Liu et al.

Backup

Assumptions in Presented TMs [2, 3, 4, 21]

- Documents are exchangeable in a corpus
  - Independent, etc.
- Bag-of-Words
  - Words are independent
  - No proximity and syntactical relations
- Topics are combinations of words
- Documents are combinations of topics

Applications – My Analysis

- Dimensionality Reduction: there are better methods
  - E.g. Neural doc or word embeddings & combinations [13]

- Downstream Tasks or direct incorporations: - depends on task
  - One can tailor LDA model for task→ good results
  - [12] does thorough analysis
  - But **neural methods** take over many fields

- For Corpus Analysis, Data Exploration [6-10]
  - Definitely good
  - Especially Combined Methods:
    - First cluster, then model topics [11]
    - Introduce time dependency (topics change over time) [14]

Extensions – Neural Topic Models

- Short comings of standard Topic Models
  - Work directly on BoWs → structural information is lost
  - Trouble with short documents

- Deep Learning + combined approaches [13-20]
  - Increase in Bayesian NNs (Variational Autoencoders, etc.)
  - Incorporating word vectors in topic models
    → Improve word-level information (polysemy, synonymy, homonyms)
  - Learn non-linear latent spaces