HSLU Lucerne University of Applied Sciences and Arts

Topic Modeling

NLP
Andreas Marfurt

Information Technology
27.02.2025

Overview

- Latent semantic analysis (LSA)
- Latent Dirichlet allocation (LDA)
- Topic analysis
- Applications

What is topic modeling?

- Goal: Clustering (= unsupervised) of documents into topically similar groups

- Hypothesis
  - A document talks about one or more (generally few) topics
  - Each topic has a different distribution of words it uses
    - Sports vs. movies vs. politics
  - We can determine the topics in a document from its words

Word-Document Matrix

- Our word-document matrix is large and sparse

	Document 1	Document 2	Document 3	Document 4
Word 1	0.05	0	0.33	0
Word 2	0	0	0.04	0
Word 3	0	0	0	0.2
Word 4	0.001	0	0.1	0
Word 5	0	0	0	0
...				

- How can we make this more efficient for processing?

Word-Document Matrix

- Our word-document matrix is large and sparse

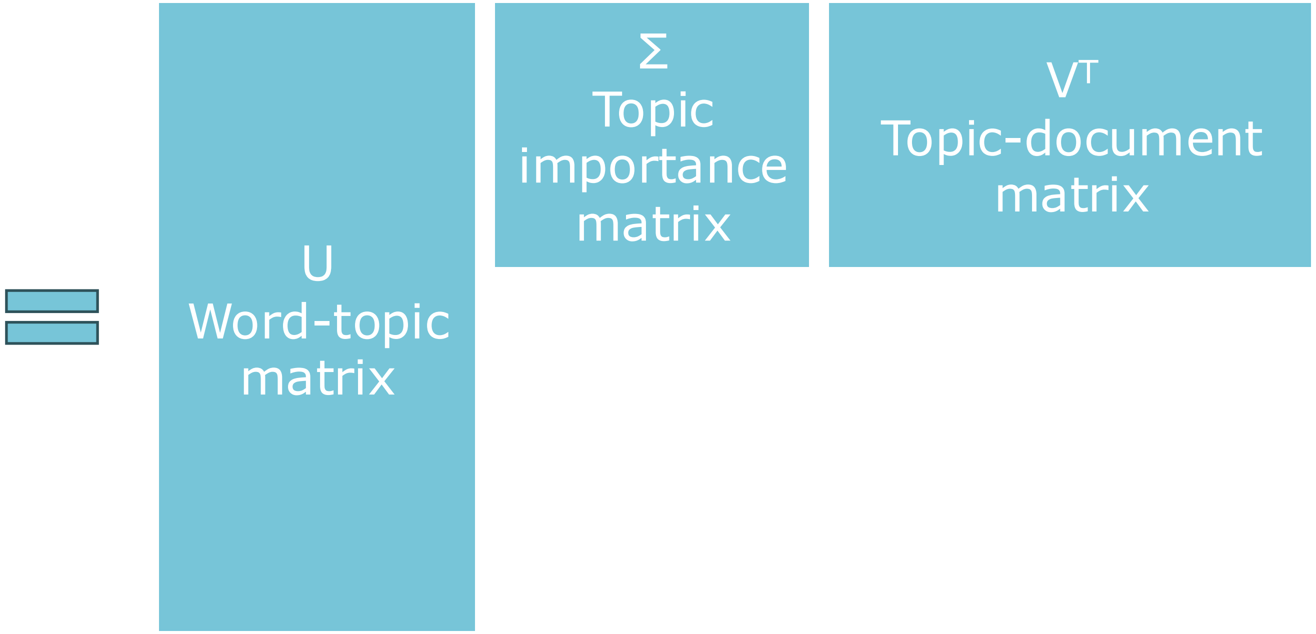
	Document 1	Document 2	Document 3	Document 4
Word 1	0.05	0	0.33	0
Word 2	0	0	0.04	0
Word 3	0	0	0	0.2
Word 4	0.001	0	0.1	0
Word 5	0	0	0	0
…				

- How can we make this more efficient for processing?
  → Dimensionality reduction

Latent Semantic Analysis (LSA)

- Deerwester et al., 1990
- LSA uses SVD on the word-document matrix: $X = U\Sigma V^T$
- Variants: original SVD vs. compact SVD vs. truncated SVD

	Document 1	Document 2	Document 3
Word 1	0.05	0	0.33
Word 2	0	0	0.04
Word 3	0	0	0
Word 4	0.001	0	0.1
Word 5	0	0	0
...			



HSLU

Latent Semantic Analysis (LSA)

- Word-topic matrix U

	sports	technology	nature	politics
bike	0.6	0.2	0.3	0.1
plane	0.05	0.5	0	0
grass	0.1	0	0.8	0
wheel	0.2	0.3	0	0
tree	0	0.4	0.7	0

Latent Semantic Analysis (LSA)

- Topic importance matrix Σ

	sports	technology	nature	politics
sports	0.6	0	0	0
technology	0	0.5	0	0
nature	0	0	0.35	0
politics	0	0	0	0.15

Latent Semantic Analysis (LSA)

- Topic-document matrix V^T

	Bike Magazine	Doping report	Rain forest	AI paper	My last hike
sports	0.6	0.2	0	0	0.7
technology	0.05	0	0	0.8	0
nature	0.3	0	0.9	0	0.8
politics	0	0.3	0.1	0.05	0

Latent Dirichlet Allocation (LDA)

- Blei et al., 2003
- Generative model for documents
  - "What is the statistical process that generated this document?"
- Idea
  - Document contains one or more topics
  - Each topic has its own distribution over words
    - More frequent words: "football" for sports topic
    - Less frequent words: "semiconductor" for sports topic

Latent Dirichlet Allocation (LDA)

- w: word
- z: topic for word in document
- N: #words in a document
- θ: topic distribution for document
- M: #documents

- α: prior topic distribution for documents
- β: prior distribution for words

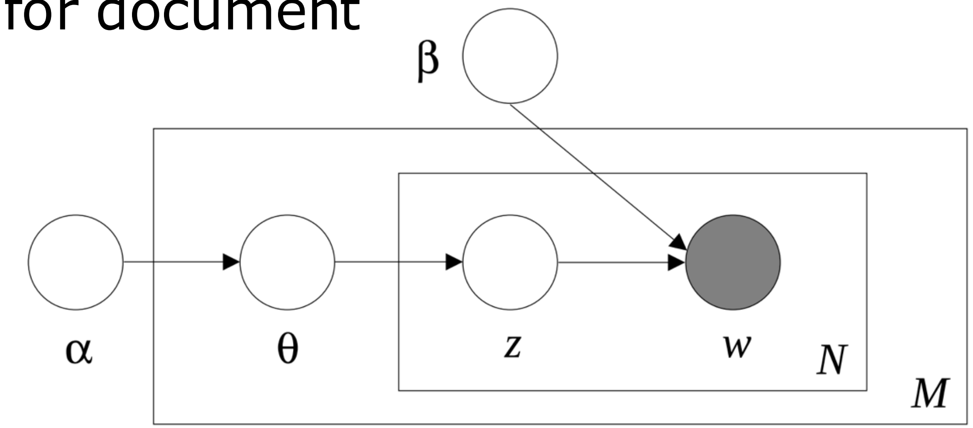


Figure 1: Graphical model representation of LDA. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

Latent Dirichlet Allocation (LDA)

- How do we know what the topic distributions should be?
- Given a document, we want to know its topics z and topic distribution θ:

$$p(\theta, z | w, \alpha, \beta) = \frac{p(\theta, z, w | \alpha, \beta)}{p(w | \alpha, \beta)}$$

- This is intractable to compute
    - Have to marginalize over all possible assignments of θ and z
- Use an approximation technique: variational inference

Latent Dirichlet Allocation (LDA)

- LDA is the standard approach to topic modeling
- LDA was the de-facto standard for document analysis before deep learning
  - Still used today (mostly in social science) for its interpretability
  - First, NN-based methods like BERT have started to replace topic modeling
  - Now, LLMs are used, but less interpretable and reproducible results

Advantages over the vector space model

- Handling synonyms (= multiple words with the same meaning)
  - "car" and "automobile" will appear in the same topics in LSA/LDA
    - There was no way to detect these as synonyms in the vector space model
- Handling polysemy (= one word with multiple meanings)
  - "bank" (park, river, money) will appear in several very different topics
    - We can figure out the relevant meaning from the document's topics

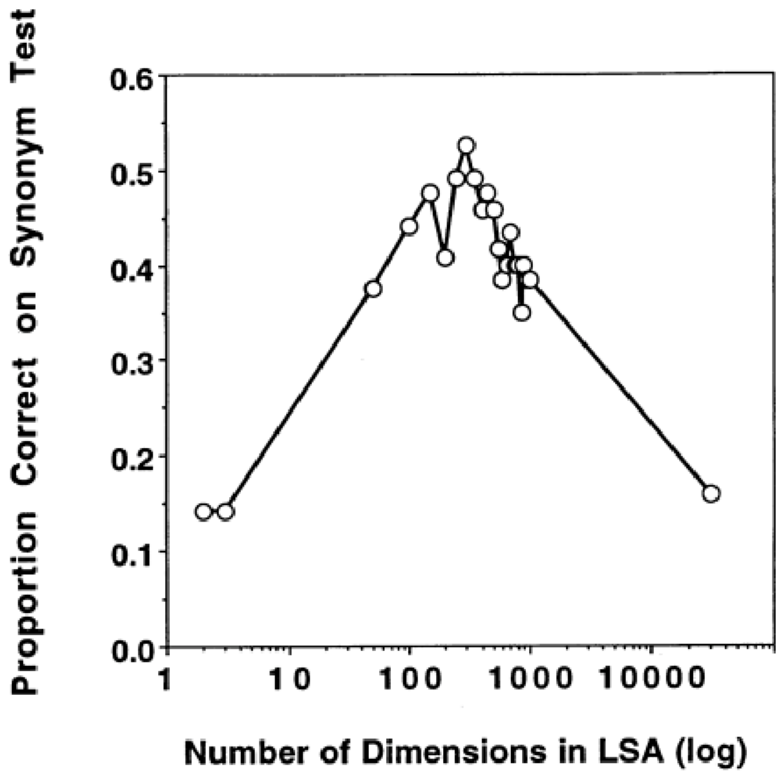
How many topics?

- Number needs to be predefined (=hyperparameter)
- Same question as we know from clustering
- Rule of thumb: Keep 80%-90% of total energy (sum of squares of singular values)
- Use gensim's topic coherence model

Landauer and Dumais, 1997

How many topics?

- Number needs to be predefined (=hyperparameter)
- Same question as we know from clustering
- Rule of thumb: Keep 80%-90% of total energy (sum of squares of singular values)
- Use gensim's topic coherence model
- Empirical: compare performance on a downstream task



 Landauer and Dumais, 1997

Topic Analysis

- Done manually: Look at words in a topic and assign it a label

?	?	?	?
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Topic Analysis

- Done manually: Look at words in a topic and assign it a label

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Topic Analysis

- Done manually: Look at words in a topic and assign it a label

"Arts" **"Budgets"** **"Children"** **"Education"**

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figure 8: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.

Applications

- Discover the topics in a collection of documents
- Automatic news categorization
- Retrieval: find similar documents
- Recommendation: find topics a user is interested in
- Social sciences: clustering/categorization of tweets

Exercise: Topic Modeling