Topic Modeling

Karol Grzegorczyk
June 3, 2014

Generative probabilistic models of textual corpora

- In order to introduce automatic processing of natural language, a **language model** is needed.
- One of the main goals of language modeling is to **assign a probability to a document:** $P(D) = P(w_1,w_2,w_3,...,w_m)$
- It is assumed that documents in a corpus were **randomly generated** (it is, of course, only a theoretical assumption; in reality, in most cases, they are created by humans)
- There are two types of generative language models:
  - those that generate each word on the basis of some number of **preceding words**
  - those that generate words based on **latent topics**

n-gram language models

- **N-gram** – a contiguous sequence of *n* items from a given document
  - When *n==m*, where *m* is a total number number of words in a document:

    $P(D) = P(w_1)P(w_2|w_1)P(w_3|w_1,w_2)...P(w_m|w_1,...,w_{m-1})$
- **Unigram** – Words are independent. Each word in a document is assumed to be randomly generated in the independent way
  - $P(D) = P(w_1)P(w_2)P(w_3,)...P(w_m)$
- **Bigram** – words are generated with probability condition on the previous word
- In reality, language has long-distance dependencies
  - **Skip-gram** is one of the solutions to this problem

Document Representations in Vector Space

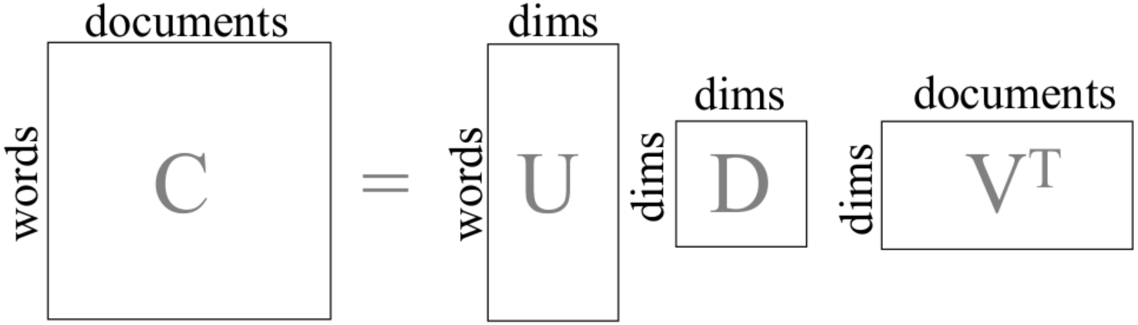
- Each document is represented by an **array of features**
- Representation types:
  - Bag-of-words (a.k.a. unigram count, term frequencies)
    - A document is represented by a **sparse vector** of the **size equals to dictionary size**
  - TF-IDF (term frequency–inverse document frequency)
    - Similar to BoW but term frequencies are weighted, to **penalize frequent words**
  - Topic Models (a.k.a. concept models, latent variable models)
    - A document is represented by a **low-rank** dense vector
- Similarity between documents (or between document and a query) can be expressed in a **cosine distance**

Topic Modeling

- Topic Modeling is a set of techniques that aim to **discover** and annotate large archives of documents with **thematic information**.
- TM is a set of methods that analyze the words (or other fine-grained features) of the original documents to **discover** the themes that run through them, how those themes are **connected** to each other, and how they **change** over time.
- Often, the **number of topics** to be discovered is predefined.
- Topic modeling can be seen as a **dimensionality reduction** technique
- Topic modeling, like clustering, **do not require any prior annotations** or labeling, but in contrast to clustering, can assign document to multiple topics.
- Semantic information can be derived from a word-document co-occurrence matrix
- Topic Model types:
  - Linear algebra based (e.g. LSA)
  - Probabilistic modeling based (e.g. pLSA, LDA, Random Projections)

Latent semantic analysis

- a.k.a. Latent Semantic Indexing
- A **low-rank** approximation of document-term matrix (typical rank 100-300)
  - In contrast, The British National Corpus (BNC) has 100-million words
- LSA downsizes the co-occurrence tables via **Singular Value Decomposition**



C - normalized co-occurrence matrix
D - a diagonal matrix, all cells except the main diagonal are zeros, elements of the main diagonal are called 'singular values'

Probabilistic Latent Semantic Analysis

pLSA models the probability of each co-occurrence as a mixture of **conditionally independent multinomial distributions**:

$$P(w, d) = \sum_{c} P(c)P(d|c)P(w|c)$$

$P(d|c)$ **relates to the V matrix** from the previous slide

$P(w|c)$ **relates to the U matrix** from the previous slide

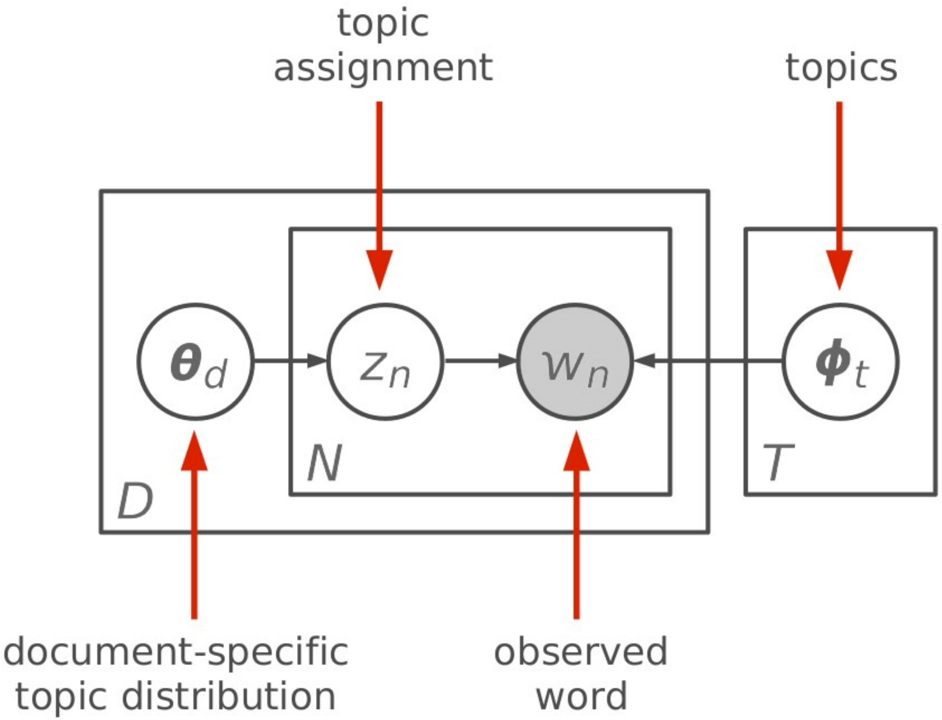
In contrast to classical LSA, words representation in topics and topic representations in documents will be **non-negative** and will **sum up to one**.

[T. Hofmann, **Probabilistic latent semantic indexing**, *SIGIR*, 1999]

pLSA – Graphical Model

- A **graphical model** is a probabilistic model for which a graph denotes the **conditional dependence** structure between random variables.
- Only a **shaded** variable is observed. All the others have to be inferred.
- We can **infer hidden variable** using maximum likelihood estimation.

- D – total number of documents
- N – total number of words in a document (it fact, it should be N_d)
- T – total number of topics

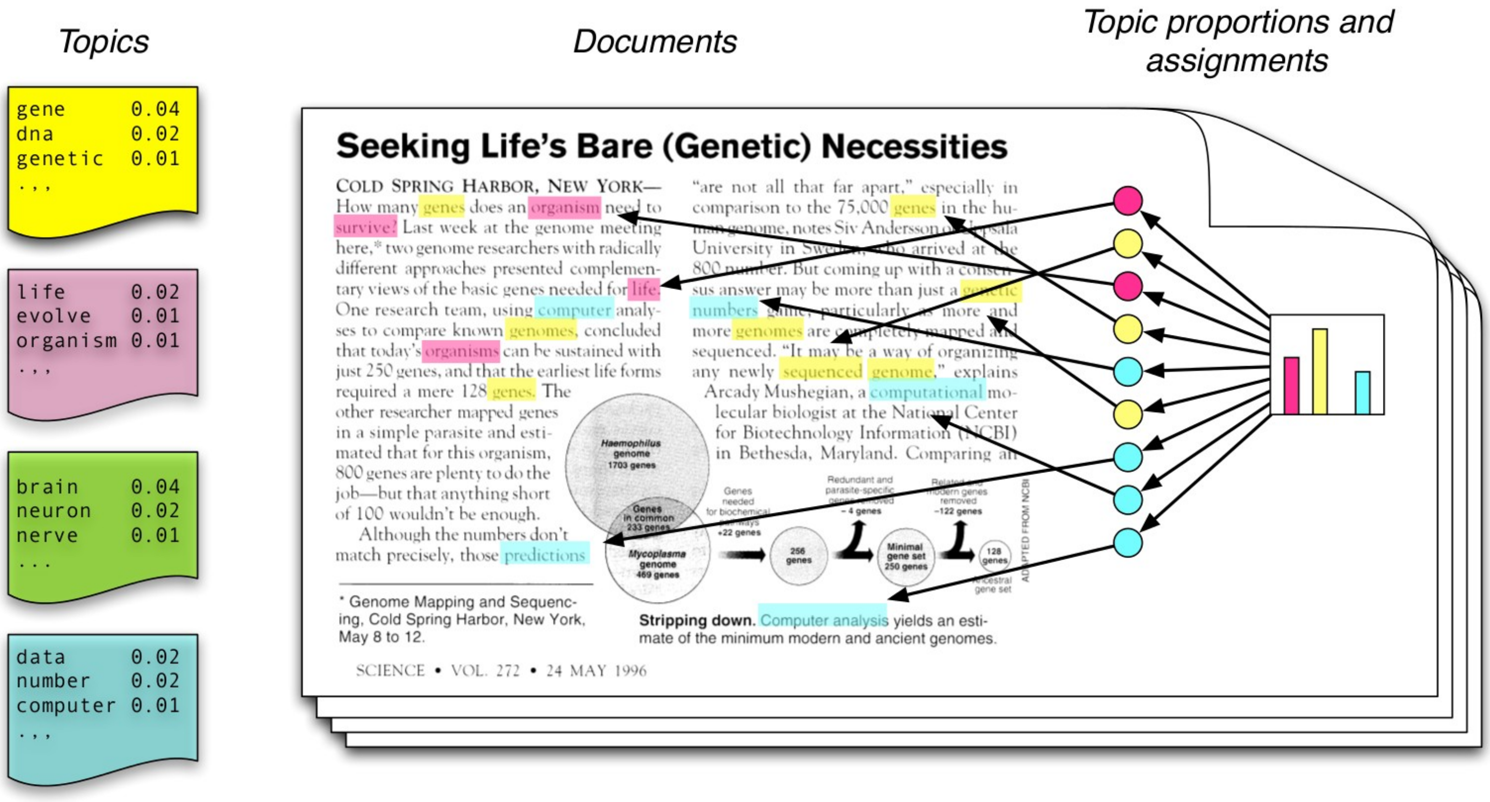


[T. Hofmann, **Probabilistic latent semantic indexing**, *SIGIR*, 1999]

Latent Dirichlet allocation

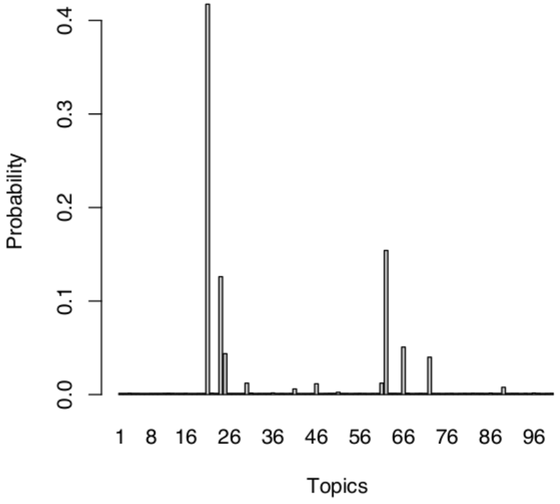
- LDA is similar to pLSA, except that in LDA the topic distribution is assumed to have a **Dirichlet prior**
- Dirichlet distribution is a **family** of continuous multivariate probability distributions
- This model assumes that documents are **generated randomly**
- Topic is a distribution over a fixed vocabulary, **each topic contains each word from the dictionary**, but some of them have very low probability
- Each word in a document is randomly selected from randomly selected topic from distribution of topics.
- Each documents exhibit multiple topics in different proportions.
  - In fact, **all the documents in the collection share the same set of topics**, but each document exhibits those topics in different proportions
- In reality, the topic structure, per-document topic distributions and the per-document per-word topic assignments are latent, and have to be inferred from observed documents.

The intuitions behind latent Dirichlet allocation



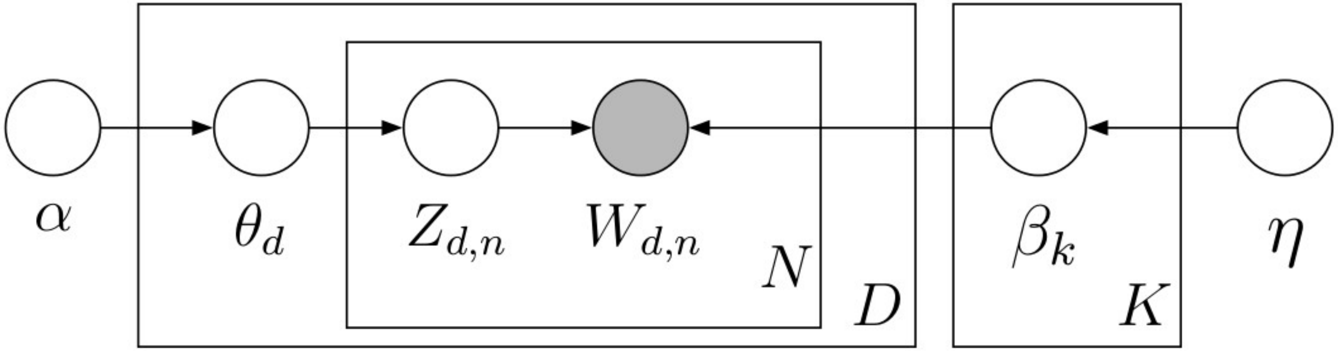
Real inference with LDA

A 100-topic LDA model was fitted to **17,000 articles from the *Science*** journal.
At right are **the top 15 most frequent words** from the most frequent topics.
At left are the **inferred topic proportions** for the example article from previous slide.



"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

The graphical model for latent Dirichlet allocation.

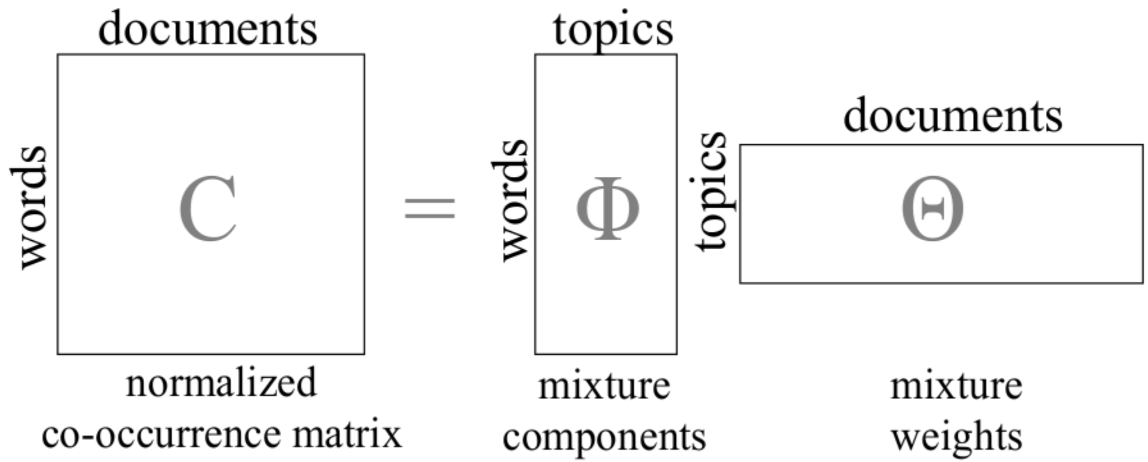


K – total number of topics
β_k – topic, a distribution over the vocabulary
D – total number of documents
Θ_d – per-document topic proportions
N – total number of words in a document (it fact, it should be N_d)
$Z_{d,n}$ – per-word topic assignment
$W_{d,n}$ – observed word
α, η – Dirichlet parameters

- Several **inference algorithms** are available (e.g. sampling based)
- A few **extensions** to LDA were created:
  - Bigram Topic Model

[D. Blei, **Probabilistic topic models**, *Communications of the ACM*, 2012]

Matrix Factorization Interpretation of LDA



Tooling

gensim: topic modeling for humans

- Free python library
- Memory independent
- Distributed computing

http://radimrehurek.com/gensim

Stanford Topic Modeling Toolbox

http://nlp.stanford.edu/software/tmt

MAchine **L**earning for **L**anguag**E** **T**oolkit (MALLET) is a Java-based package for:

- statistical natural language processing
- document classification
- Clustering
- topic modeling
- information extraction
- and other machine learning applications to text.

http://mallet.cs.umass.edu

Topic modeling applications

- Topic-based **text classification**
  - Classical text classification algorithms (e.g. perceptron, naïve bayes, k-nearest neighbor, SVM, AdaBoost, etc.) are often assuming **bag-of-words representation** of input data.
  - Topic modeling can be seen as a **pre-processing** step before applying supervised learning methods.
  - Using topic-based representation it is possible to gain ~0,039 in precision and ~0,046 in F1 score [Cai, Hofmann, 2003]
- Collaborative filtering [Hofmann, 2004]
- **Finding patterns** in genetic data, images, and social networks

[L. Cai and T. Hofmann, **Text Categorization by Boosting Automatically Extracted Concepts**, *SIGIR,* 2003]
[T. Hofmann, **Latent Semantic Models for Collaborative Filtering**, *ACM TOIS*, 2004]

Word Representations in Vector Space

- Notion of **similarity** between words
- Continuous-valued vector representation of words
- **Neural network** language model
- Prediction semantic of the word based on the **context**
- Ability to perform simple **algebraic operations on the word vectors**:

    vector("King") - *vector("Man")* + *vector("Woman")* will yield *Queen*

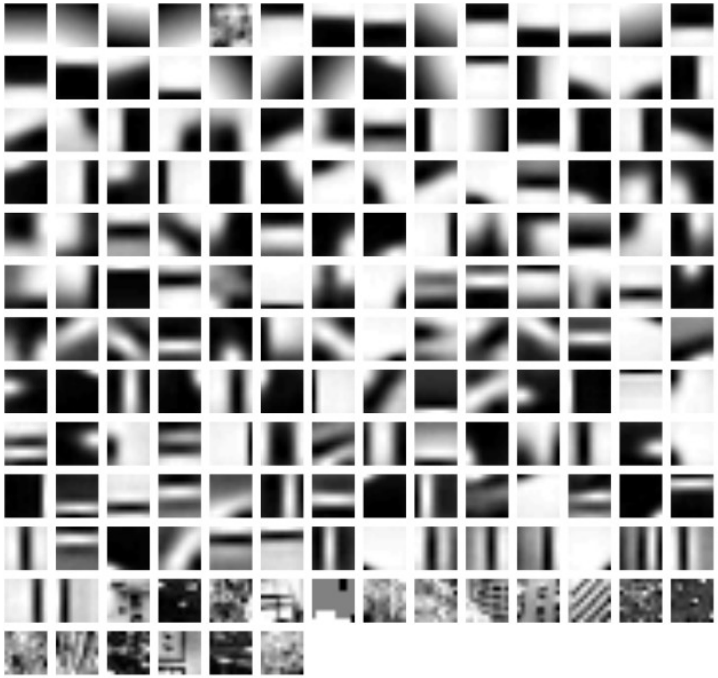
- *word2ved: https://code.google.com/p/word2vec*

Topic Modeling in Computer Vision

- Bag-of-words model in computer vision (a.k.a. bag-of-features)
  - **Codewords** (“visual words”) instead of just words
  - **Codebook** instead of dictionary
- It is assumed, that documents exhibit multiple topics and the collection of documents exhibits the same set of topics
- TM in CV has been used to:
  - **Classify** images
  - Build image **hierarchies**
  - Connecting images and **captions**
- The main advantage of this approach it its **unsupervised** training nature.

Codebook example

Obtained from 650 training examples from 13 categories of **natural scenes** (e.g. highway, inside of cities, tall buildings, forest, etc) using **k-means** clustering algorithm.



[L. Fei-Fei, P. Perona, **A bayesian hierarchical model for learning natural scene categories**, *Computer Vision and Pattern Recognition*, 2005]

Bibliography

- D. Blei, **Probabilistic topic models**, *Communications of the ACM*, 2012
- M. Steyvers, T. Griffiths, **Probabilistic Topic Models**, *Handbook of latent semantic analysis*, 2007
- S. Deerwester, et al. **Indexing by latent semantic analysis**, *JASIS*, 1990
- D. Blei, A. Ng, M. Jordan, **Latent dirichlet allocation**, *Journal of machine Learning research*, 2003
- H. Wallach, **Topic modeling: beyond bag-of-words**, *International conference on Machine learning*, 2006
- T. Mikolov, K. Chen, G. Corrado, J. Dean, **Efficient estimation of word representations in vector space**, *ICLR*, 2013
- L. Fei-Fei, P. Perona, **A Bayesian Hierarchical Model for Learning Natural Scene Categories**, *Computer Vision and Pattern Recognition*, 2005
- X. Wang, E. Grimson, **Spatial Latent Dirichlet Allocation**, *Conference on Neural Information Processing Systems*, 2007

Thank you!