Text Mining
Statistical Modeling of Textual Data
Lecture 3

Måns Magnusson, Mattias Villani

Division of Statistics
Dept. of Computer and Information Science
Linköping University

Overview Lecture 3

- Distributional semantics and word embeddings
- Topic models
- Demo of **topicmodels package** in R

The distributional semantics hypothesis

- Semantics - the **meaning** of words

The distributional semantics hypothesis

- Semantics - the **meaning** of words
- Distributional semantics - categorizing meaning by the distributional properties in **large samples**

The distributional semantics hypothesis

- Semantics - the **meaning** of words
- Distributional semantics - categorizing meaning by the distributional properties in **large samples**
- The distributional semantics hypothesis

  "a word is characterized by the company it keeps"
  Firth (1957)

Distributional semantics

- Word meaning comes from textual context

  “cold”

Distributional semantics

- Word meaning comes from textual context

  "cold"

  "It's cold outside."

Distributional semantics

- Word meaning comes from textual context

  "cold"

  "It's cold outside."

  "I'm having a cold"

Distributional semantics

- Word meaning comes from textual context

"cold"

"It's cold outside."

"I'm having a cold"

"I'm cold"

Distributional semantics

- Word meaning comes from textual context

  "cold"

  "It's cold outside."

  "I'm having a cold"

  "I'm cold"

- Different contexts (sentence, word windows, documents)

Distributional semantics

- Word meaning comes from textual context

  "cold"

  "It's cold outside."

  "I'm having a cold"

  "I'm cold"

- Different contexts (sentence, word windows, documents)
- Different context size - different properties
  - Short distance context, syntagmatic similarities
  - Long distance context, topical similarities

Co-occurance matrix

A friend in need is a friend indeed.
She is my friend indeed.

	Doc 1	Doc 2
a	2	0
friend	2	1
in	1	0
indeed	1	1
is	1	1
my	0	1
need	1	0
she	0	1

Co-occurance matrix II

A friend in need is a friend indeed.
She is my friend indeed.

- Context window of one step

	a	friend	in	indeed	is	my	need	she
a	2	2	0	0	1	0	0	0
friend	2	3	1	2	0	1	0	0
in	0	1	1	0	0	0	1	0
indeed	0	2	0	2	0	0	0	0
is	1	0	0	0	2	1	1	1
my	0	1	0	0	1	1	0	0
need	0	0	1	0	1	0	1	0
she	0	0	0	0	1	0	0	1

Word embedding

- Reduce co-occurance matrix to a **lower dimension**

Word embedding

- Reduce co-occurance matrix to a **lower dimension**
- Often **part** of more complex models

Word embedding

- Reduce co-occurance matrix to a **lower dimension**
- Often **part** of more complex models
- Popular approaches (word-word)
  - Random indexing
  - Word2Vec

Word embedding

- Reduce co-occurance matrix to a **lower dimension**
- Often **part** of more complex models
- Popular approaches (word-word)
  - Random indexing
  - Word2Vec
- Popular approaches (word-doc)
  - Latent Semantic analysis (SVD decomposition)
  - Topic models

Topic models

- Models for **unsupervised learning**, but more recently also for **supervised learning**.

Topic models

- Models for **unsupervised learning**, but more recently also for **supervised learning**.
- **Probabilistic generative** models.

Topic models

- Models for **unsupervised learning**, but more recently also for **supervised learning**.
- **Probabilistic generative** models.
- **Very popular** model in applications and research. > 12000 Google scholar citations in 11 years.

Topic models

- Models for **unsupervised learning**, but more recently also for **supervised learning**.
- **Probabilistic generative** models.
- **Very popular** model in applications and research. $>$ 12000 Google scholar citations in 11 years.
- **Many extensions** in recent years: nGrams, supervised, nonparametric, relational topics, correlated topics, dynamically time-varying topics etc.

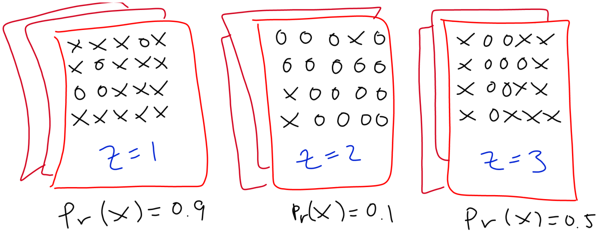
Topic models

- Models for **unsupervised learning**, but more recently also for **supervised learning**.
- **Probabilistic generative** models.
- **Very popular** model in applications and research. $>$ 12000 Google scholar citations in 11 years.
- **Many extensions** in recent years: nGrams, supervised, nonparametric, relational topics, correlated topics, dynamically time-varying topics etc.
- The basic topic models are extensions of the bag-of-words (unigram) model.
- **Unigram** model: each word is assumed to be drawn from the same word (term) distribution.

$$\hat{P}(w) = \frac{\#w}{N}$$

Mixture of unigrams

- **Mixture of unigrams**:
  1. Draw a *topic* z_d for the dth document from a topic distribution $\theta = (\theta_1, ..., \theta_K)$.
  2. Conditional on the drawn topic z_d draw words from a word distribution for that topic.



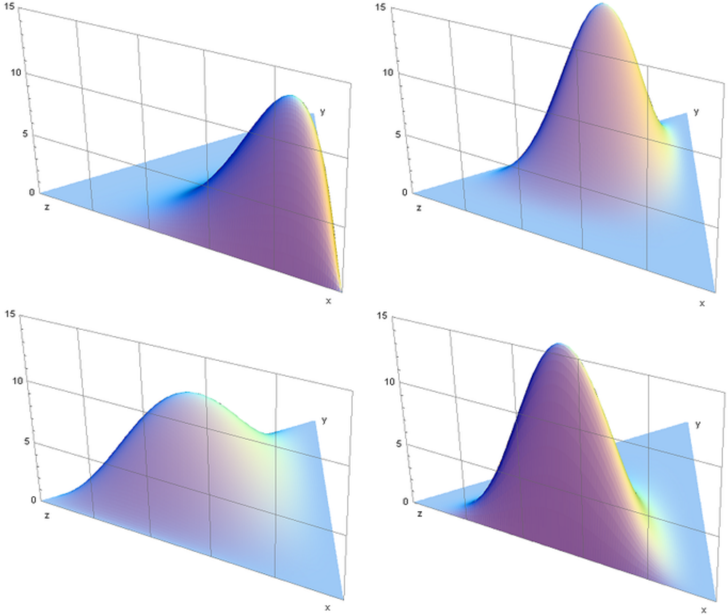
- Topic models are **mixed-membership models**: each document can belong to **several topics simultaneously**.

Multinomial and Dirichlet distributions

- **Multinomial distribution**: random discrete variable $X \in \{1, 2, ..., K\}$ that can assume exactly one of K (unordered) values.
  - $Pr(X = k) = \theta_k$
  - Parameters $\theta = (\theta_1, ..., \theta_K)$.

Multinomial and Dirichlet distributions

- **Multinomial distribution**: random discrete variable $X \in \{1, 2, ..., K\}$ that can assume exactly one of K (unordered) values.
  - $Pr(X = k) = \theta_k$
  - Parameters $\theta = (\theta_1, ..., \theta_K)$.
- **Dirichlet distribution**: random **vector** $X = (X_1, ..., X_K)$ satisfying the constraint $X_1 + X_2 + ... + X_K = 1$.
  - Unit simplex
  - Parameters: $\alpha = (\alpha_1, ..., \alpha_K)$
  - Uniform distribution: $\alpha = (1, 1, ..., 1)$
  - Small variance (informative) when the α's are large.
  - "Bathtub shape" when $\alpha_k < 1$ for all k.


15
10
5
0
z
y
x
15
10
5
0
z
y
x
15
10
5
0
z
y
x
15
10
5
0
z
y
x

Generating a corpus from a topic model

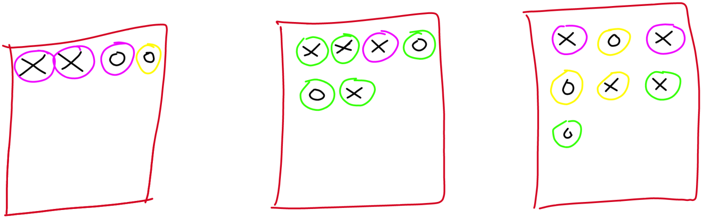
- Assume that we have:
  - A fixed vocabulary V
  - D documents
  - N words in each document
  - K topics

Generating a corpus from a topic model

- Assume that we have:
  - A fixed vocabulary V
  - D documents
  - N words in each document
  - K topics

1. **For each topic** $(k = 1, ..., K)$:
   A. Draw a distribution over the words $\beta_k \sim Dir(\eta, \eta, ..., \eta)$
2. **For each document** $(d = 1, ..., D)$:
   A. Draw a vector of topic proportions $\theta_d \sim Dir(\alpha_1, ..., \alpha_K)$
   B. **For each word** $(n = 1, ..., N)$:
      I. Draw a topic assignment $z_{d,n} \sim Multinomial(\theta_d)$
      II. Draw a word $w_{d,n} \sim Multinomial(\beta_{z_{d,n}})$

(HORRIBLE PICTURE OF A) TOPIC MODEL



$\theta_1 = (0.9\ \ 0.1\ \ 0)$

$\theta_2 = (0.1\ \ 0.1\ \ 0.8)$

$\theta_3 = (0.3\ \ 0.4\ \ 0.3)$

$\beta_1 = (0.9\ \ 0.1)$

$\beta_2 = (0.1\ \ 0.9)$

$\beta_3 = (0.5\ \ 0.5)$

Example - simulation from two topics

Topic	Word distr.	probability	dna	gene	data	distribution
1	β_1	0.5	0.1	0.0	0.2	0.2
2	β_2	0.0	0.5	0.4	0.1	0.0

Doc 1	$\theta_1 = (0.2, 0.8)$		
	Word 1:	Topic=2	Word='gene'
	Word 2:	Topic=2	Word='gene'
	Word 3:	Topic=1	Word='data'

Doc 2	$\theta_2 = (0.9, 0.1)$		
	Word 1:	Topic=1	Word='probability'
	Word 2:	Topic=1	Word='data'
	Word 3:	Topic=1	Word='probability'

Doc 3	$\theta_2 = (0.5, 0.5)$

Learning/inference in topic models

- What do we know?
  - The words in the documents: $\mathbf{w}_{1:D}$

Learning/inference in topic models

- What do we know?
  - The words in the documents: $\mathbf{w}_{1:D}$
- What do we not know?
  - Topic proportions for each document: $\theta_{1:D}$
  - Topic assignments for each word in each document: $z_{1:D}$
  - Word distributions for each topic: $\beta_{1:K}$

Learning/inference in topic models

- What do we know?
  - The words in the documents: $\mathbf{w}_{1:D}$
- What do we not know?
  - Topic proportions for each document: $\theta_{1:D}$
  - Topic assignments for each word in each document: $z_{1:D}$
  - Word distributions for each topic: $\beta_{1:K}$
- Do the Bayes dance: Posterior distribution

$$p(\theta_{1:D}, z_{1:D}, \beta_{1:K} | w_{1:D})$$

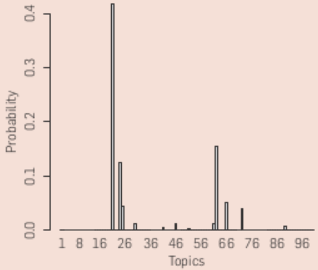
Learning/inference in topic models

- What do we know?
  - The words in the documents: $\mathbf{w}_{1:D}$
- What do we not know?
  - Topic proportions for each document: $\theta_{1:D}$
  - Topic assignments for each word in each document: $z_{1:D}$
  - Word distributions for each topic: $\beta_{1:K}$
- Do the Bayes dance: Posterior distribution

$$p(\theta_{1:D}, z_{1:D}, \beta_{1:K} | w_{1:D})$$

- The posterior is mathematically untractable. **Solutions**:
  - Gibbs sampling (MCMC) [Correct, but can be slow]
  - Variational Bayes [Crude approximation of the posterior *distribution*, but typically rather accurate about posterior mode (MAP)]

Figure 2. Real inference with LDA. We fit a 100-topic LDA model to 17,000 articles from the journal *Science*. At left are the inferred topic proportions for the example article in Figure 1. At right are the top 15 most frequent words from the most frequent topics found in this article.



"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

From Blei et al. (2010)

Gibbs sampler

Bayes theorem

$$p(B|A) = \frac{p(A|B) \cdot p(B)}{p(B)}$$

For the topic model

$$\begin{aligned} p(\mathbf{z}, \Theta, \Phi|\mathbf{w}) &= \frac{p(\mathbf{z}, \Theta, \Phi|\mathbf{w}) \cdot p(\mathbf{z}, \Theta, \Phi)}{p(\mathbf{w})} \\ &\propto p(\mathbf{z}, \Theta, \Phi|\mathbf{w}) \cdot p(\mathbf{z}, \Theta, \Phi) \end{aligned}$$

Gibbs sampler II

Integrating out (collapsing) Θ and Φ (?):

$$p(\mathbf{z}|\mathbf{w}) \;=\; \int\int p(\mathbf{z},\Theta,\Phi|\mathbf{w})\cdot p(\mathbf{z},\Theta,\Phi)\,d\Phi d\Theta$$

will result in the following gibbs sampler

$$p(z_i = k|w_i,\mathbf{z}_{\neg i}) = \underbrace{\frac{n^{(w)}_{k,v_i}+\beta}{n^{(w)}_{k,\cdot}+V\beta}}_{type-topic\ (\Phi)} \cdot \underbrace{(n^{(d)}_{k,d_i}+\alpha)}_{topic-doc\ (\Theta)}$$

where $n^{(w)}$ and $n^{(d)}$ are count matrices of size $D\times K$ and $K\times V$.

EXAMPLE OF $n^{(w)}$ AND $n^{(d)}$

$\mathbf{w}_1$	boat	shore	bank		
$\mathbf{z}_1$	1	1	1		
$\mathbf{w}_2$	Zlatan	boat	shore	money	bank
$\mathbf{z}_2$	2	1	1	3	3
$\mathbf{w}_3$	money	bank	soccer	money	
$\mathbf{z}_3$	3	3	2	3	

Example of $n^{(w)}$ and $n^{(d)}$

$$
\begin{array}{llllll}
\mathbf{w}_1 & \text{boat} & \text{shore} & \text{bank} & & \\
\mathbf{z}_1 & 1 & 1 & 1 & & \\
\mathbf{w}_2 & \text{Zlatan} & \text{boat} & \text{shore} & \text{money} & \text{bank} \\
\mathbf{z}_2 & 2 & 1 & 1 & 3 & 3 \\
\mathbf{w}_3 & \text{money} & \text{bank} & \text{soccer} & \text{money} & \\
\mathbf{z}_3 & 3 & 3 & 2 & 3 &
\end{array}
$$

$$
n^{(w)} = \begin{array}{cccccc}
\text{boat} & \text{shore} & \text{soccer} & \text{Zlatan} & \text{bank} & \text{money} \\
2 & 2 & 0 & 0 & 1 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 2 & 2
\end{array}
$$

Example of $n^{(w)}$ and $n^{(d)}$

$\mathbf{w}_1$	boat	shore	bank		
$\mathbf{z}_1$	1	1	1		
$\mathbf{w}_2$	Zlatan	boat	shore	money	bank
$\mathbf{z}_2$	2	1	1	3	3
$\mathbf{w}_3$	money	bank	soccer	money	
$\mathbf{z}_3$	3	3	2	3	

$$n^{(w)} = \begin{array}{cccccc} \text{boat} & \text{shore} & \text{soccer} & \text{Zlatan} & \text{bank} & \text{money} \\ 2 & 2 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 2 \end{array}$$

$$n^{(d)} = \begin{bmatrix} 3 & 0 & 0 \\ 2 & 1 & 3 \\ 0 & 2 & 3 \end{bmatrix}$$

(Naive) algorithm

```
# Initialization
Sample all topic indicators randomly
Calculate n^(w) and n^(d)


# Gibbs sampler
for each gibbs iteration do
    for each token w_i do
        remove z_i from n^(w) and n^(d)
        for each k in 1 to K do
            prob_k[k] = 
```
$$\frac{n^{(w)}_{k,v_i}+\beta}{n^{(w)}_{k,\cdot}+V\beta}\cdot\left(n^{(d)}_{k,d_i}+\alpha\right)$$
```
        end for
        z_i <- draw multinomial(prob_k)
        add z_i to n^(w) and n^(d)
    end for
end for
return n^(w), n^(d)
```

(Naive) algorithm II

- Estimation of Φ and Θ

$$
\begin{aligned}
\hat{\phi}_{k,v} &= \frac{n_{k,v}^{(w)} + \beta}{n_{k,\cdot}^{(w)} + V\beta} \\
\hat{\theta}_{d,k} &= \frac{n_{d,k}^{(d)} + \alpha}{n_{d,\cdot}^{(d)} + K\alpha}
\end{aligned}
$$

(Naive) algorithm II

- Estimation of Φ and Θ

$$\hat{\phi}_{k,v} = \frac{n_{k,v}^{(w)} + \beta}{n_{k,\cdot}^{(w)} + V\beta}$$

$$\hat{\theta}_{d,k} = \frac{n_{d,k}^{(d)} + \alpha}{n_{d,\cdot}^{(d)} + K\alpha}$$

- Serial
- Computational complexity is $O(K)$ for each token
- Slow for larger corpuses...

Evaluation of topic models

- Convergence:
  - Log-likelihood

Evaluation of topic models

- Convergence:
  - Log-likelihood
- Evaluating and comparing models:
  - Held-out perplexity
  - See Wallach et al. (2009)

References

Blei, D., Carin, L., Dunson, D., Nov. 2010. Probabilistic Topic Models. IEEE Signal Processing Magazine, 77–84.
URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5563111

Firth, J., 1957. A synopsis of linguistic theory 1930-1955. Studies in linguistic analysis, 1–32.

Wallach, H. M., Murray, I., Salakhutdinov, R., Mimno, D., 2009. Evaluation methods for topic models. In: Proceedings of the 26th Annual International Conference on Machine Learning. ACM, pp. 1105–1112.