Text mining and topic models

Måns Magnusson

StiMa, Linköping University

2015-09-23

Overview

Text as data

Topic models

Inference in topic models

Scaling topic models

Text as data

- Digital

Text as data

- Digital
- Abundant

Text as data

- Digital
- Abundant
- Unstructured

Text as data

- Digital
- Abundant
- Unstructured
- High-dimensional

Some definitions

“The lazy dog jumps over the fox.”

- **Tokens**

Some definitions

“The lazy dog jumps over the fox.”

- **Tokens**
- **(Word) types**

Some definitions

"The lazy dog jumps over the fox."

- **Tokens**
- **(Word) types**
- **Documents**

Some definitions

"The lazy dog jumps over the fox."

- **Tokens**
- **(Word) types**
- **Documents**
- **Corpus**

Some definitions

"The lazy dog jumps over the fox."

- **Tokens**
- **(Word) types**
- **Documents**
- **Corpus**
- **Vocabulary**

Natural language processing (NLP)

- Computers and natural language
- NLP is hard
- Computer science and computational linguistics

Examples of NLP tasks

- Machine translation
- Part-of-speach tagging
- Parse trees
- Natural language generation
- Text classification
- Text summarization
- Dialogue systems

Deep and shallow NLP

- **Deep NLP**
  - complex and language specific - do not scale
- **Shallow NLP**
  - robust, less language specific and scales

Text mining

- Shallow NLP
- Statistical approach
- **Supervised learning**
  - Text classification, Google flu trend
- **Unsupervised learning**
  - Document clustering, topic model, word2vec

The distributional semantics hypothesis

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

Distributional semantics

- Meaning comes from context

  “cold”

Distributional semantics

- Meaning comes from context

  “cold”

  “It’s cold outside.”

Distributional semantics

- Meaning comes from context

  "cold"

  "It's cold outside."

  "I'm having a cold"

Distributional semantics

- Meaning comes from context

  "cold"

  "It's cold outside."

  "I'm having a cold"

  "I'm cold"

- Different contexts (sentance, word windows, documents), different models

Topic modeling

- Unsupervised model (basic model)
- Learn “topics” or “themes” in a corpus
- Context is document
- Multiple topics per document **[example]**
- The most known model is **Latent dirichlet allocation** (Blei et al. (2003))

Topic examples

"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

Figure : Example topics from 17 000 articles in *Science* Blei et al. (2010)

The graphical model

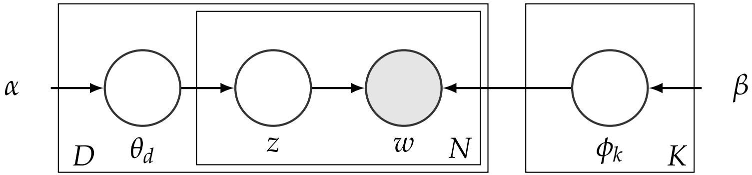


Figure : The LDA model

Θ

- Topic distribution over documents
- Size: $D \times K$

$$\Theta = \begin{array}{c} \text{Doc 1} \\ \text{Doc 2} \\ \text{Doc 3} \\ \text{Doc 4} \\ \text{Doc 5} \end{array} \left[\begin{array}{ccc} 0.3 & 0.2 & 0.5 \\ 0.1 & 0.1 & 0.8 \\ 0.1 & 0.7 & 0.2 \\ 0.8 & 0.1 & 0.1 \\ 0.2 & 0.6 & 0.2 \end{array} \right]$$

Φ

- Topic distribution over vocabulary
- Size: $K \times V$

$$\Phi = \begin{array}{lcccccc} & \text{boat} & \text{shore} & \text{soccer} & \text{Zlatan} & \text{bank} & \text{money} \\ \text{Topic 1} & 0.4 & 0.4 & 0.01 & 0.03 & 0.15 & 0.01 \\ \text{Topic 2} & 0.01 & 0.01 & 0.5 & 0.46 & 0.01 & 0.01 \\ \text{Topic 3} & 0.01 & 0.01 & 0.01 & 0.01 & 0.48 & 0.48 \end{array}$$

$\mathbf{z}$

- One topic indicator per word

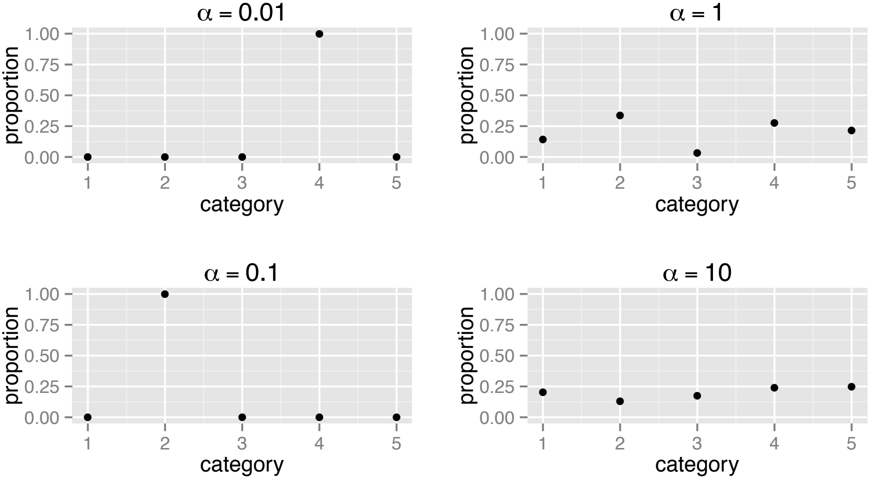
$\mathbf{w}_1$	boat	shore	bank		
$\mathbf{z}_1$	1	1	1		
$\mathbf{w}_2$	Zlatan	boat	shore	money	bank
$\mathbf{z}_2$	2	1	1	3	3
$\mathbf{w}_3$	money	bank	soccer	money	
$\mathbf{z}_3$	3	3	2	3	

Dirichlet distribution

- Distribution over the simplex
- Generalization of the beta distribution

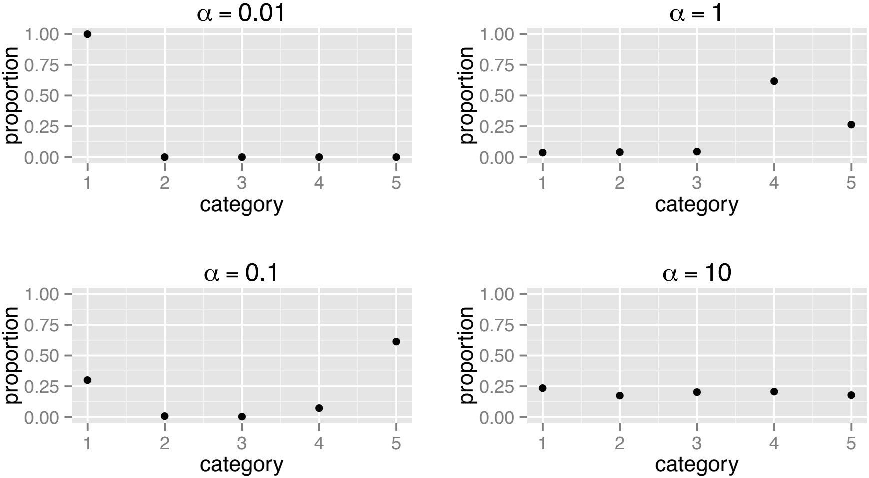
Dirichlet distribution

- Distribution over the simplex
- Generalization of the beta distribution



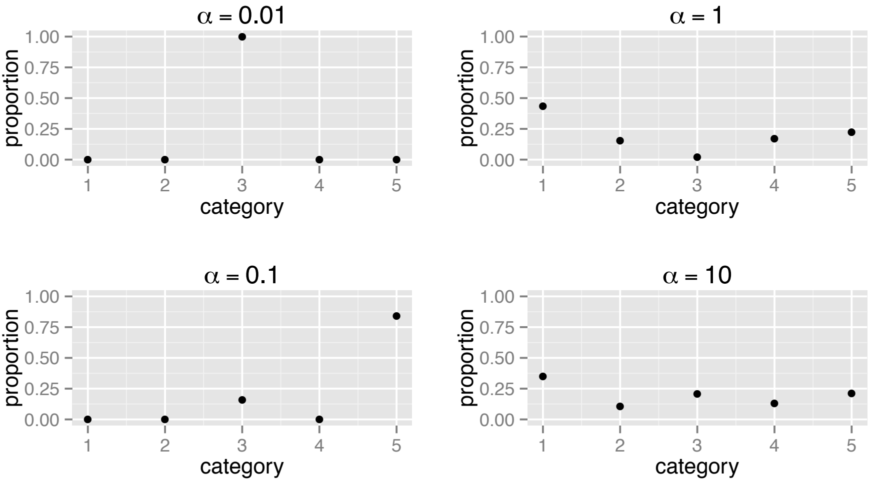
Dirichlet distribution

- Distribution over the simplex
- Generalization of the beta distribution



Dirichlet distribution

- Distribution over the simplex
- Generalization of the beta distribution



Generative model

1. For each topic k in K:
   1.1 Sample topic-word distribution $\phi_i \sim \mathrm{Dir}(\beta)$
2. For each document d in D:
   2.1 Sample the topic proportions $\theta \sim \mathrm{Dir}(\alpha)$
   2.2 For each word in document d:
      2.2.1 Sample topic indicator $z \sim \mathrm{Multinomial}(\theta)$
      2.2.2 Sample word $w \sim \mathrm{Multinomial}(\phi_z)$

Simulated documents

- We can easily simulate "documents"

$$\theta_d = (0.55894, 0.00022, 0.44084)$$

Simulated documents

- We can easily simulate "documents"

$$\theta_d = (0.55894, 0.00022, 0.44084)$$

$$\mathbf{z} = (3, 1, 1, 3, 1, 3, 1)$$

Simulated documents

- We can easily simulate "documents"

$$\theta_d = (0.55894, 0.00022, 0.44084)$$

$$\mathbf{z} = (3, 1, 1, 3, 1, 3, 1)$$

$\mathbf{w}$: boat bank shore money shore money boat

Simulated documents

- We can easily simulate "documents"

$$\theta_d = (0.01494, 0.53875, 0.44632)$$

Simulated documents

- We can easily simulate "documents"

$$\theta_d = (0.01494, 0.53875, 0.44632)$$

$$\mathbf{z} = (2, 3, 3, 3, 3, 3, 3)$$

Simulated documents

- We can easily simulate "documents"

$$\theta_d = (0.01494, 0.53875, 0.44632)$$

$$\mathbf{z} = (2, 3, 3, 3, 3, 3, 3)$$

$\mathbf{w}$: soccer money bank money money money money

Simulated documents

- We can easily simulate "documents"

$$\theta_d = (0.32533, 0.13562, 0.53905)$$

Simulated documents

- We can easily simulate "documents"

$$\theta_d = (0.32533, 0.13562, 0.53905)$$

$$\mathbf{z} = (1, 3, 3, 1, 3, 3, 1)$$

Simulated documents

- We can easily simulate "documents"

$$\theta_d = (0.32533, 0.13562, 0.53905)$$

$$\mathbf{z} = (1, 3, 3, 1, 3, 3, 1)$$

w : shore bank bank soccer bank bank shore

Inference methods for topic models

- We want to learn Φ, Θ and $\mathbf{z}$ given our observations
- "Reverse the generative model"
- **Assumption:** Bag of word

Inference methods for topic models

- We want to learn Φ, Θ and $\mathbf{z}$ given our observations
- "Reverse the generative model"
- **Assumption:** Bag of word

- Variational bayes (VB)
  Blei et al. (2003)
- Markov chain monte carlo (Gibbs sampling)
  Griffiths and Steyvers (2004)

Bayesian learning

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

For the topic model

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

Bayesian learning

Integrating out (collapsing) Θ and Φ (Griffiths and Steyvers (2004)):

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

will result in the folowing 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.

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 $n^{(d)}$

$\mathbf{w}_1$	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	

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

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	

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

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

Algorithm

```
LDA_gibbs(w)
# 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
```

$$\texttt{prob\_k[k]} = \frac{n^{(w)}_{k,v_i}+\beta}{n^{(w)}_{k,\cdot}+V\beta} \cdot \left(n^{(d)}_{k,d_i} + \alpha\right)$$

```
        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)
```

Basic algorithm

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

Big models

- Big corpuses today (Yuan et al. (2015))

Dataset	V	N	D
NYTimes	101K	99M	300K
PubMed	140K	737M	8.2M
BingWebC	1M	200B	1.2B

Big models

- Big corpuses today (Yuan et al. (2015))

Dataset	V	N	D
NYTimes	101K	99M	300K
PubMed	140K	737M	8.2M
BingWebC	1M	200B	1.2B

- How to handle **big** corpuses:
  - Paralellism
  - Improve algorithm speed
  - Subsampling

Parallel topic models

- Approximately distributed LDA
  Newman et al. (2009)
  - Ignore that the sampler is serial (not correct)

Parallel topic models

- Approximately distributed LDA
  Newman et al. (2009)
  - Ignore that the sampler is serial (not correct)
- **We want:**
  - A (correct) parallel sampler
  - Should work well when $D \to \infty$

Parallel topic models

- Approximately distributed LDA
  Newman et al. (2009)
  - Ignore that the sampler is serial (not correct)
- **We want:**
  - A (correct) parallel sampler
  - Should work well when $D \to \infty$
- **Problem:** Integrating out **both** Θ and Φ
- **But:** sampling Θ and Φ
  - is costly
  - decreases efficiency of the MCMC chain

Properties of Θ and Φ

- Φ is $K \times V$
- Θ is $D \times K$
- What happen when $D \to \infty$

Properties of Θ and Φ

- Φ is $K \times V$
- Θ is $D \times K$
- What happen when $D \to \infty$
  - $K \approx O(\log(N))$ (or less than V)
  - $D \approx O(N)$
  - $V \approx O(\sqrt{N})$
- We want to integrate out Θ that grows faster

Heaps law?

- Empirical law of language

$$V(N) = \kappa N^{\gamma}$$

where $\gamma \approx \frac{1}{2}$

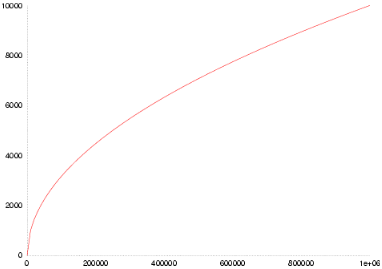


Figure : Heaps law (picture from Wikipedia)

The partially collapsed sampler

$$p(z_i = k | w_i, \mathbf{z}_{\neg i}) = \underbrace{\phi_{k,w_i}}_{type-topic} \cdot (\underbrace{n_k^{(d_i)} + \alpha)}_{topic-doc\ (\Theta)}$$

in parallel over documents, and then

$$\phi_k \sim \mathrm{Dir}(n_k^{(w)} + \beta)$$

in parallel over topics

Some extra tricks

- Walker-Alias method (seeLi et al. (2014))
- Using the sparsity in $n^{(d)}$
- Cashed Marsaglia gamma sampling (see Marsaglia and Tsang (2000))
- Job stealing

Results

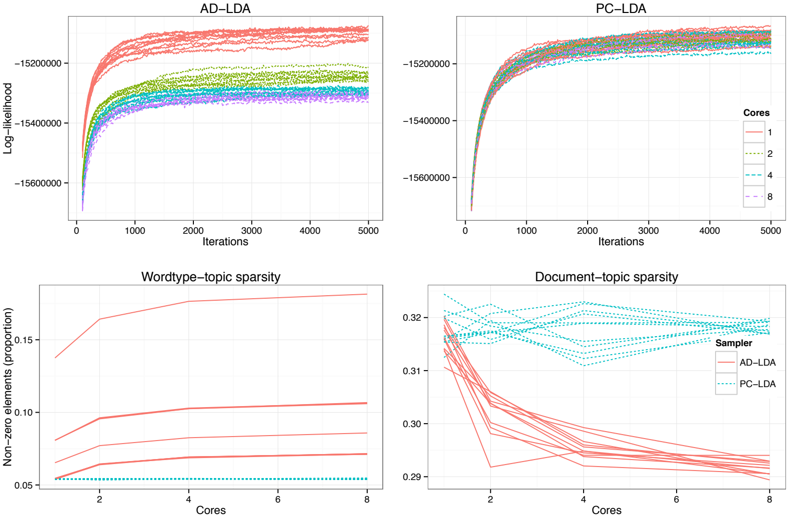


Figure : Effect of approximating MCMC Magnusson et al. (2015)

Results

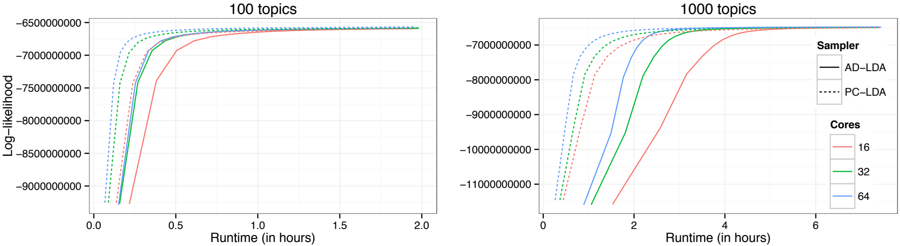


Figure : Inference in big models Magnusson et al. (2015)

Results

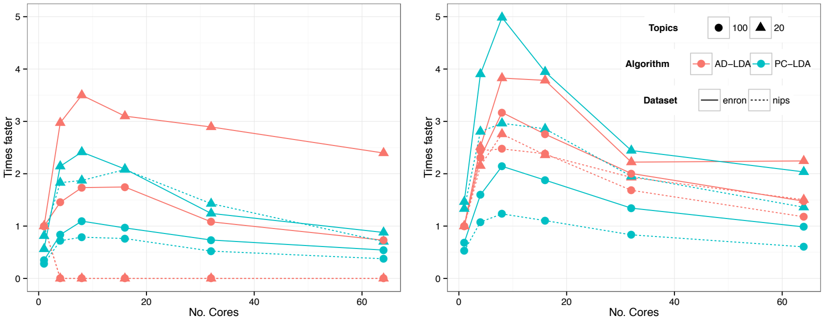


Figure : Speedup of PC-LDA and sparse AD-LDA Magnusson et al. (2015)

Summary of findings

- Approximate distributed LDA can lead to the wrong model
- Distributing topic models using partially collapsed sampling
  - can be fast (depend on K)
  - can handle big corpuses
  - can model Φ
  - is not nessecarily less effective
  - is correct
  - seems to explore the posterior better

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`

Blei, D. M., Ng, A. Y., Jordan, M. I., 2003. Latent dirichlet allocation. the Journal of machine Learning research 3, 993–1022.

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

Griffiths, T., Steyvers, M., 2004. Finding scientific topics. …academy of Sciences of the United ….
URL `http://www.pnas.org/content/101/suppl.1/5228.short`

Li, A. Q., Ahmed, A., Ravi, S., Smola, A. J., 2014. Reducing the sampling complexity of topic models. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 891–900.

Magnusson, M., Jonsson, L., Villani, M., Broman, D., 2015. Parallelizing lda using partially collapsed gibbs sampling. arXiv preprint arXiv:1506.03784.