CS598JHM: **Advanced NLP** (Spring '10)

Lecture 6: Topic Models (Latent Dirichlet Allocation)

Julia Hockenmaier
juliahmr@illinois.edu
3324 Siebel Center

http://www.cs.uiuc.edu/class/sp10/cs598jhm

Topic models

Motivating questions:

- What are the topics that a document is about?
- Given one document, can we find other documents about the same topics?
- How do topics in a field change over time?

A hierarchical Bayesian approach:

- Assume each document defines a distribution over (hidden) topics
- Assume each topic defines a distribution over words
- The posterior probability of these latent variables given a document collection determines a hidden decomposition of the collection into topics.

References

D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January 2003.

D. Blei and J. Lafferty. Topic Models. In A. Srivastava and M. Sahami, editors, *Text Mining: Theory and Applications*. Taylor and Francis, 2009.

Griffiths, T., & Steyvers, M. (2004). Finding Scientific Topics. *Proceedings of the National Academy of Sciences, 101 (suppl. 1), 5228-5235.*

Steyvers, M. & Griffiths, T. (2006). Probabilistic topic models. In T. Landauer, D McNamara, S. Dennis, and W. Kintsch (eds), *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum

Setting up a generative model

- We have D documents using a vocabulary of V word types
- Each documents contains (up to) N word tokens.
- We assume K topics.
- Each document has a K-dimensional multinomial θ_d over topics with a common Dirichlet prior $Dir(\boldsymbol{\alpha})$
- Each topic has a V-dimensional multinomial β_k over words with a common symmetric Dirichlet prior $D(\eta)$.

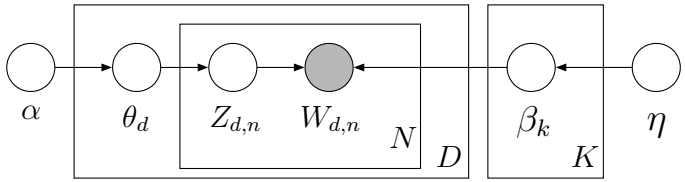
D. Blei and J. Lafferty. Topic Models. In A. Srivastava and M. Sahami, editors, *Text Mining: Theory and Applications*. Taylor and Francis, 2009.

The generative process

For each **topic** $1...k$:
- Draw a **multinomial over words** $\beta_k \sim Dir(\eta)$

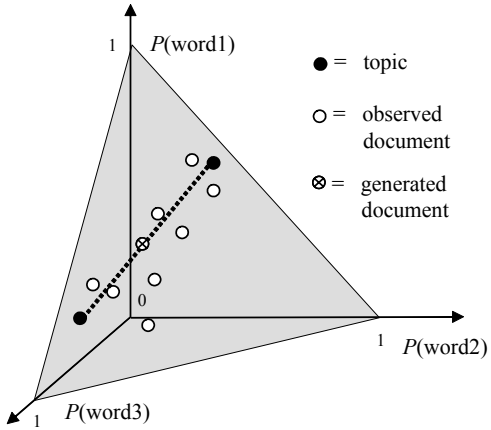
For each **document** $1...d$:
- Draw a **multinomial over topics** $\theta_d \sim Dir(\boldsymbol{\alpha})$
- For each word $w_{d,n}$:
  - Draw a topic $Z_{d,n} \sim Mult(\theta_d)$ with $Z_{d,n} \in [1..K]$
  - Draw a word $w_{d,n} \sim Mult(\beta_{Z_{d,n}})$

As a graphical model



D. Blei and J. Lafferty. Topic Models. In A. Srivastava and M. Sahami, editors, *Text Mining: Theory and Applications*. Taylor and Francis, 2009.

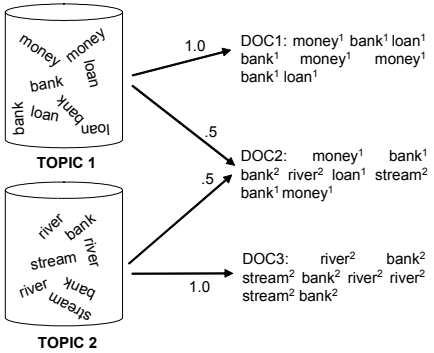
A geometric interpretation



Steyvers, M. & Griffiths, T. (2006). Probabilistic topic models. In T. Landauer et al. (eds), *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum

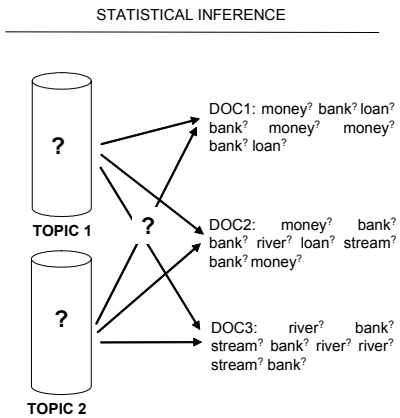
The generative process

PROBABILISTIC GENERATIVE PROCESS



Steyvers, M. & Griffiths, T. (2006). Probabilistic topic models. In T. Landauer, D McNamara, S. Dennis, and W. Kintsch (eds), *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum

The inference problem



Steyvers, M. & Griffiths, T. (2006). Probabilistic topic models. In T. Landauer, D McNamara, S. Dennis, and W. Kintsch (eds), *Latent Semantic Analysis: A Road to Meaning.* Laurence Erlbaum

The inference problem

What is the **posterior of the hidden variables** given the observed variables (and hyperparameters)?

$$p(\vec{\theta}_{1:D}, z_{1:D,1:N}, \vec{\beta}_{1:K} \mid w_{1:D,1:N}, \alpha, \eta) = \frac{p(\vec{\theta}_{1:D}, \vec{z}_{1:D}, \vec{\beta}_{1:K}\,;\, \vec{w}_{1:D}, \alpha, \eta)}{\int_{\vec{\beta}_{1:K}} \int_{\vec{\theta}_{1:D}} \sum_{\vec{z}} p(\vec{\theta}_{1:D}, \vec{z}_{1:D}, \vec{\beta}_{1:K}\,;\, \vec{w}_{1:D}, \alpha, \eta)}.$$

Problem: the integral in the denominator is intractable!

Solutions: approximate inference
- Gibbs sampling (Griffiths & Steyvers)
- Variational Inference (Blei et al.)

Gibbs sampling

Represent corpus as an array of words *w[i]*, document indices *d[i]* and topics *z[i]*.

Words *w[i]* and document indices *d[i]* are fixed.
Only the topics *z[i]* change.

States of Markov chain = topic assignments to words.

"Macrosteps": assign a new topic to all of the words
"Microsteps": assign a new topic to each word w[i].

Assigning a new topic to w_i

The probability $P(z_i = j \mid \mathbf{z}_{-i}, \mathbf{w}, \mathbf{d})$ is proportional to the probability of w_i under topic j times the probability of topic j given document d_i

Define $n_{-i,j}(w_i)$ as freq. of w_i labeled as topic j.
Define $n_{-i,j}(d_i)$ as number of words in d_i labeled as topic j.

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}, \mathbf{d}) \quad \propto \quad \underbrace{\frac{n_{-i,j}(w_i) + \eta}{V\eta + \sum_{v=1}^{V} n_{-i,j}(w_v)}}_{\text{Prob of } w_i \text{ under topic } z_i} \quad \underbrace{\frac{n_{-i,j}(d_i) + \alpha}{K\alpha \sum_{k=1}^{K} n_{-i,j}(d_i)}}_{\text{Prob of topic } z_i \text{ in document } d_i}$$

What other quantities do we need?

As Bayesians, we want to compute the expected value of the parameters given the observed data.
Our data is the set of words $w_{1:D,1:N}$.
Hence, we need to compute $E[\ldots \mid w_{1:D,1:N}]$

The **topic probability** of the word v according to topic k:
$\langle \beta_{k,v} \rangle = E[\beta_{k,v} \mid w_{1:D,1:N}]$

The **topic proportions of each document** d**:**
$\langle \theta_{d,k} \rangle = E[\theta_{d,k} \mid w_{1:D,1:N}]$

The **topic assignment of each word** $w_{d,n}$**:**
$\langle z_{d,n,k} \rangle = E[Z_{d,n} = k \mid w_{1:D,1:N}]$

What other quantities do we need?

From any single sample, we can estimate:

The **topic probability** of the word v according to topic k:
$\langle \beta_{k,v} \rangle = E[\beta_{k,v} \mid w_{1:D,1:N}]$

The **topic proportions of each document** d**:**
$\langle \theta_{d,k} \rangle = E[\theta_{d,k} \mid w_{1:D,1:N}]$

$$\langle \beta_{k,i} \rangle = \frac{n_j(w_i) + \eta}{V\eta + \sum_{v=1}^{V} n_j(w_v)}$$

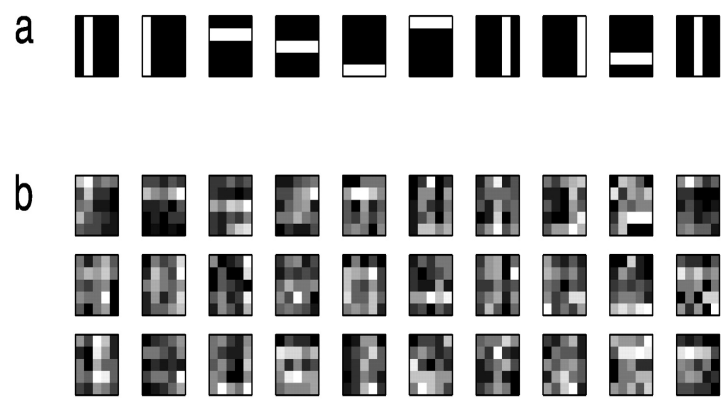
$$\langle \theta_{d,k} \rangle = \frac{n_j(d_i) + \alpha}{K\alpha + \sum_{k=1}^{K} n_j(d_i)}$$

Running the Gibbs sampler

A toy example from Griffiths, T., & Steyvers, M. (2004):

- 25 words.
- 10 predefined topics
- 2000 documents generated according to known distributions.
- Each document = 5x5 image.
  Pixel intensity = Frequency of word.

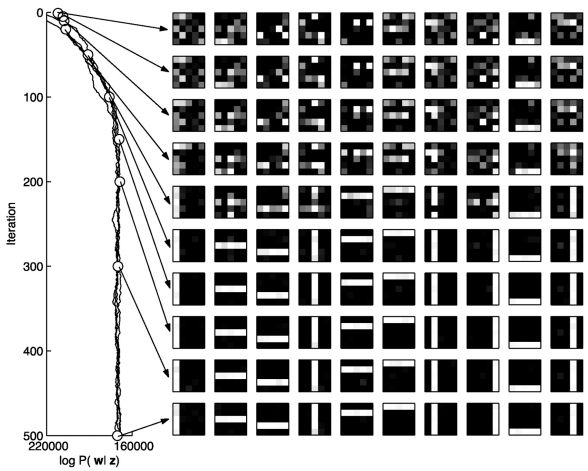
(a) Graphical representation of 10 topics, combined to produce "documents" like those shown in b, where each image is the result of 100 samples from a unique mixture of these topics



Griffiths T L, Steyvers M PNAS 2004;101:5228-5235

PNAS

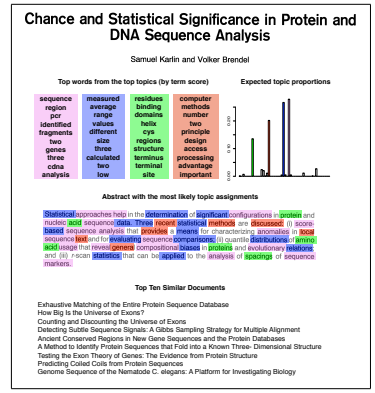
Results of running the Gibbs sampling algorithm



Griffiths T L, Steyvers M PNAS 2004;101:5228-5235

PNAS

Things you can do with LDA



D. Blei and J. Lafferty. Topic Models. In A. Srivastava and M. Sahami, editors, *Text Mining: Theory and Applications*. Taylor and Francis, 2009.

Visualizing a topic

What is the distribution of words defined by a topic?
Instead of using term probabilities directly, downweight them by how likely they are to be generated by any topic:

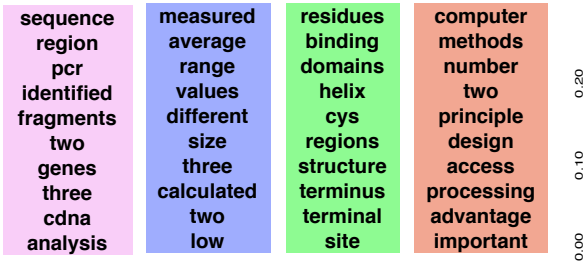
$$\text{term-score}_{k,v} = \widehat{\beta}_{k,v} \log \left(\frac{\widehat{\beta}_{k,v}}{\left(\prod_{j=1}^{K} \widehat{\beta}_{j,v} \right)^{\frac{1}{K}}} \right).$$

contractual	employment	female	markets	criminal
expectation	industrial	men	earnings	discretion
gain	local	women	investors	justice
promises	jobs	see	sec	civil
expectations	employees	sexual	research	process
breach	relations	note	structure	federal
enforcing	unfair	employer	managers	see
supra	agreement	discrimination	firm	officer
note	economic	harassment	risk	parole
perform	case	gender	large	inmates

FIGURE 3. Five topics from a 50-topic model fit to the *Yale Law Journal* from 1980–2003.

Visualizing a topic

Top words from the top topics (by term score)



D. Blei and J. Lafferty. Topic Models. In A. Srivastava and M. Sahami, editors, *Text Mining: Theory and Applications*. Taylor and Francis, 2009.

Visualizing a document

Abstract with the most likely topic assignments

Statistical approaches help in the determination of significant configurations in protein and nucleic acid sequence data. Three recent statistical methods are discussed: (i) score-based sequence analysis that provides a means for characterizing anomalies in local sequence text and for evaluating sequence comparisons; (ii) quantile distributions of amino acid usage that reveal general compositional biases in proteins and evolutionary relations; and (iii) *r*-scan statistics that can be applied to the analysis of spacings of sequence markers.

Use the posterior topic probabilities of each document and the posterior topic assignments to each word

D. Blei and J. Lafferty. Topic Models. In A. Srivastava and M. Sahami, editors, *Text Mining: Theory and Applications*. Taylor and Francis, 2009.

Document similarity

Two documents are similar if they assign similar probabilities to topics:

$$\text{document-similarity}_{d,f} = \sum_{k=1}^{K} \left(\sqrt{\widehat{\theta}_{d,k}} - \sqrt{\widehat{\theta}_{f,k}} \right)^2 .$$

Chance and Statistical Significance in Protein and DNA Sequence Analysis

Samuel Karlin and Volker Brendel

Top Ten Similar Documents

Exhaustive Matching of the Entire Protein Sequence Database
How Big Is the Universe of Exons?
Counting and Discounting the Universe of Exons
Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment
Ancient Conserved Regions in New Gene Sequences and the Protein Databases
A Method to Identify Protein Sequences that Fold into a Known Three- Dimensional Structure
Testing the Exon Theory of Genes: The Evidence from Protein Structure
Predicting Coiled Coils from Protein Sequences
Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology