Unsupervised topic identification and topic segmentation of texts

CS546 / Apr 24, 2009
Problem Setup

Problem:
- Detect a set of topics discussed in a text
- Segment a document according to the topics

Requirements:
- No supervision
- Probabilistic approaches
Motivation

- Visualization of collections:
  - what are the topics discussed?
  - which documents discuss a topic?

- Opinion mining:
  - what is sentiment towards a product aspects?
  - what are important aspects of a product?

- Dimensionality reduction:
  - for information retrieval
  - for document classification

- Summarization:
  - ensuring topic coverage in a summary

...
Outline

1. Bag-of-words Models
2. Beyond bag-of-words
   - Documents as Sequences of Windows
   - Topic N-grams
3. Topic hierarchies
4. Approximate Inference
Latent Semantics Analysis [Deerwester et al, 90]

- Decomposition of the co-occurrence table $X = U \Sigma V'$
- $X \approx \hat{X} = U_k \Sigma_k V_k'$

Optimal rank $k$ approximation (Frobenius norm)

Hope: terms having **common meaning** are mapped to the same direction

Non-zero inner products between documents with non-overlapping terms

Not motivated probabilistically, no obvious interpretation of directions
Generative model of a collection:
for each word position in collection:

- select a document $d$ with probability $P(d)$
- select a latent class $z$ with probability $P(z|d)$
- select a word with probability $P(w|z)$

$P(w|z)$ - *topic models*

$P(z|d)$ - *topic distributions*

ML estimation of parameters:

$$L = \sum_{d,w} n(d, w) \log P(d) \sum_z P(z|d) P(w|z)$$

[Hofmann, 99]
Example topics (Science)

- Top words for 10 topics out of 128 ordered by $P(w|z)$:

<table>
<thead>
<tr>
<th>Topic</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>universe</td>
<td>0.0439</td>
</tr>
<tr>
<td>galaxies</td>
<td>0.0375</td>
</tr>
<tr>
<td>clusters</td>
<td>0.0279</td>
</tr>
<tr>
<td>matter</td>
<td>0.0233</td>
</tr>
<tr>
<td>galaxy</td>
<td>0.0232</td>
</tr>
<tr>
<td>cluster</td>
<td>0.0214</td>
</tr>
<tr>
<td>cosmic</td>
<td>0.0137</td>
</tr>
<tr>
<td>dark</td>
<td>0.0131</td>
</tr>
<tr>
<td>light</td>
<td>0.0109</td>
</tr>
<tr>
<td>density</td>
<td>0.01</td>
</tr>
<tr>
<td>drug</td>
<td>0.0672</td>
</tr>
<tr>
<td>patients</td>
<td>0.0493</td>
</tr>
<tr>
<td>drugs</td>
<td>0.0444</td>
</tr>
<tr>
<td>clinical</td>
<td>0.0346</td>
</tr>
<tr>
<td>treatment</td>
<td>0.028</td>
</tr>
<tr>
<td>trials</td>
<td>0.0277</td>
</tr>
<tr>
<td>therapy</td>
<td>0.0213</td>
</tr>
<tr>
<td>trial</td>
<td>0.0164</td>
</tr>
<tr>
<td>disease</td>
<td>0.0157</td>
</tr>
<tr>
<td>medical</td>
<td>0.00997</td>
</tr>
<tr>
<td>cells</td>
<td>0.0675</td>
</tr>
<tr>
<td>stem</td>
<td>0.0478</td>
</tr>
<tr>
<td>human</td>
<td>0.0421</td>
</tr>
<tr>
<td>cell</td>
<td>0.0309</td>
</tr>
<tr>
<td>gene</td>
<td>0.025</td>
</tr>
<tr>
<td>tissue</td>
<td>0.0185</td>
</tr>
<tr>
<td>cloning</td>
<td>0.0169</td>
</tr>
<tr>
<td>transfer</td>
<td>0.0155</td>
</tr>
<tr>
<td>blood</td>
<td>0.0113</td>
</tr>
<tr>
<td>embryos</td>
<td>0.0111</td>
</tr>
<tr>
<td>sequence</td>
<td>0.0818</td>
</tr>
<tr>
<td>sequences</td>
<td>0.0493</td>
</tr>
<tr>
<td>genome</td>
<td>0.033</td>
</tr>
<tr>
<td>dna</td>
<td>0.0257</td>
</tr>
<tr>
<td>sequencing</td>
<td>0.0172</td>
</tr>
<tr>
<td>map</td>
<td>0.0123</td>
</tr>
<tr>
<td>genes</td>
<td>0.0122</td>
</tr>
<tr>
<td>chromosome</td>
<td>0.0119</td>
</tr>
<tr>
<td>regions</td>
<td>0.0119</td>
</tr>
<tr>
<td>human</td>
<td>0.0111</td>
</tr>
<tr>
<td>years</td>
<td>0.156</td>
</tr>
<tr>
<td>million</td>
<td>0.0556</td>
</tr>
<tr>
<td>ago</td>
<td>0.045</td>
</tr>
<tr>
<td>time</td>
<td>0.0317</td>
</tr>
<tr>
<td>age</td>
<td>0.0243</td>
</tr>
<tr>
<td>year</td>
<td>0.024</td>
</tr>
<tr>
<td>record</td>
<td>0.0238</td>
</tr>
<tr>
<td>early</td>
<td>0.0233</td>
</tr>
<tr>
<td>billion</td>
<td>0.0177</td>
</tr>
<tr>
<td>history</td>
<td>0.0148</td>
</tr>
<tr>
<td>bacteria</td>
<td>0.0983</td>
</tr>
<tr>
<td>bacterial</td>
<td>0.0561</td>
</tr>
<tr>
<td>resistance</td>
<td>0.0431</td>
</tr>
<tr>
<td>coli</td>
<td>0.0381</td>
</tr>
<tr>
<td>strains</td>
<td>0.025</td>
</tr>
<tr>
<td>microbiol</td>
<td>0.0214</td>
</tr>
<tr>
<td>microbial</td>
<td>0.0196</td>
</tr>
<tr>
<td>strain</td>
<td>0.0165</td>
</tr>
<tr>
<td>salmonella</td>
<td>0.0163</td>
</tr>
<tr>
<td>resistant</td>
<td>0.0145</td>
</tr>
<tr>
<td>male</td>
<td>0.0558</td>
</tr>
<tr>
<td>females</td>
<td>0.0541</td>
</tr>
<tr>
<td>male</td>
<td>0.0529</td>
</tr>
<tr>
<td>males</td>
<td>0.0477</td>
</tr>
<tr>
<td>sex</td>
<td>0.0339</td>
</tr>
<tr>
<td>reproductive</td>
<td>0.0172</td>
</tr>
<tr>
<td>offspring</td>
<td>0.0168</td>
</tr>
<tr>
<td>sexual</td>
<td>0.0166</td>
</tr>
<tr>
<td>reproduction</td>
<td>0.0143</td>
</tr>
<tr>
<td>eggs</td>
<td>0.0138</td>
</tr>
<tr>
<td>theory</td>
<td>0.0811</td>
</tr>
<tr>
<td>physics</td>
<td>0.0782</td>
</tr>
<tr>
<td>physicists</td>
<td>0.0146</td>
</tr>
<tr>
<td>einstein</td>
<td>0.0142</td>
</tr>
<tr>
<td>university</td>
<td>0.013</td>
</tr>
<tr>
<td>gravity</td>
<td>0.013</td>
</tr>
<tr>
<td>black</td>
<td>0.0127</td>
</tr>
<tr>
<td>theories</td>
<td>0.01</td>
</tr>
<tr>
<td>aps</td>
<td>0.00987</td>
</tr>
<tr>
<td>matter</td>
<td>0.00954</td>
</tr>
<tr>
<td>immune</td>
<td>0.0909</td>
</tr>
<tr>
<td>response</td>
<td>0.0375</td>
</tr>
<tr>
<td>system</td>
<td>0.0358</td>
</tr>
<tr>
<td>responses</td>
<td>0.0322</td>
</tr>
<tr>
<td>antigen</td>
<td>0.0263</td>
</tr>
<tr>
<td>antigens</td>
<td>0.0184</td>
</tr>
<tr>
<td>immunity</td>
<td>0.0176</td>
</tr>
<tr>
<td>immunology</td>
<td>0.0145</td>
</tr>
<tr>
<td>antibody</td>
<td>0.014</td>
</tr>
<tr>
<td>autoimmune</td>
<td>0.0128</td>
</tr>
<tr>
<td>stars</td>
<td>0.0524</td>
</tr>
<tr>
<td>star</td>
<td>0.0458</td>
</tr>
<tr>
<td>astrophys</td>
<td>0.0237</td>
</tr>
<tr>
<td>mass</td>
<td>0.021</td>
</tr>
<tr>
<td>disk</td>
<td>0.0173</td>
</tr>
<tr>
<td>black</td>
<td>0.0161</td>
</tr>
<tr>
<td>gas</td>
<td>0.0149</td>
</tr>
<tr>
<td>stellar</td>
<td>0.0127</td>
</tr>
<tr>
<td>astron</td>
<td>0.0125</td>
</tr>
<tr>
<td>hole</td>
<td>0.00824</td>
</tr>
</tbody>
</table>
PLSA is related to standard distributional clustering [Pereira et al, 93], but **different**

Distributional clustering: document is sampled from a single (unknown) cluster $z$:  
\[
P(w_1^d, \ldots, w_n^d | d) = \sum_z P(z | d) \prod_i P(w_i^d | z)
\]

PLSA:  
\[
P(w_1^d, \ldots, w_n^d | d) = \prod_i \sum_z P(z | d) P(w_i^d | z)
\]

Each topic $z$ in PLSA does not need to explain all the words in a document
PLSA is inherently **transductive**: representation $P(z|d)$ is defined only for documents in the collection.

**Overfits**: number of parameters in $P(z|d)$ is proportional to number of documents - regularization is needed

**Need for valid generative models of text!**
Bayesian networks: notation

- Graph denotes conditional dependence of nodes on their parents:

  ![Bayesian network diagram]

  **Shaded** circles are visible variables (value known), not shaded - latent (or hidden) variables.

- Plate notation:
We are defining a generative hierarchical Bayesian model:

Distributions will not be fixed parameters?
They will be latent variables in the model conditioned on fixed parameters

But how to define distribution of distributions?
How to sample a Bernoulli distribution?

Beta Distribution:
- defines a distribution of $x \in [0, 1]$
- $p(x|\alpha, \beta) \propto x^{\alpha-1}(1-x)^{\beta-1}$
- Mode: $\frac{\alpha-1}{\alpha+\beta-2}$ (if $\alpha, \beta > 1$), mean: $\frac{\alpha}{\alpha+\beta}$
\[ p(x|\alpha, \beta) \propto x^{\alpha-1}(1-x)^{\beta-1} \]

- Mode: \( \frac{\alpha-1}{\alpha+\beta-2} \) (if \( \alpha, \beta > 1 \)), mean: \( \frac{\alpha}{\alpha+\beta} \)

- You can consider \( \alpha - 1 \) and \( \beta - 1 \) as “pseudo-counts”: if you observed a head \( \alpha - 1 \) and a tail \( \beta - 1 \) of times your ML estimate for \( x \) will be the mode of pdf

- If you have more “pseudo-counts” the model becomes more confident, so pdf is more peaky
Sampling Distributions over K variables

- Dirichlet distribution:
  \[ p(x_1, ..., x_K | \alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1} \]  
  on \( x_i > 0 \) and \( \sum_{i=1}^{K} x_i = 1 \)

- It can be used as a prior for distributions

- Why use this prior? - it is conjugate to multinomial distribution:

  - if we used Dirichlet prior \( p(\theta) \) and multinomial distribution \( p(y | \theta) \) then posterior \( p(\theta | y) \) will also be a Dirichlet distribution
  - It helps in learning and inference
Latent Dirichlet Allocation (Blei et al, 03)

Generative model of text:

- for each topic $k$ choose distribution of words
  \[ \varphi_k \sim \text{Dir}(\beta) \]
- for each document $d$
  - choose distribution of topics
    \[ \theta_d \sim \text{Dir}(\alpha) \]
  - for each word position $i$
    - choose topic $z_{d,i} \sim \theta_d$
    - choose word $w_{d,i} \sim \varphi_{z_{d,i}}$

- $\alpha$ and $\beta$ are hyperparameters
Latent Dirichlet Allocation (Blei et al, 03)

Generative model of text:

- for each topic $k$ choose distribution of words
  $$\varphi_k \sim \text{Dir}(\beta)$$
- for each document $d$
  - choose distribution of topics
    $$\theta_d \sim \text{Dir}(\alpha)$$
  - for each word position $i$
    - choose topic $z_{d,i} \sim \theta_d$
    - choose word $w_{d,i} \sim \varphi_{z_{d,i}}$

- $\alpha$ and $\beta$ are hyperparameters
Latent Dirichlet Allocation (Blei et al, 03)

Generative model of text:

- for each topic $k$ choose distribution of words
  $\varphi_k \sim \text{Dir} (\beta)$
- for each document $d$
  - choose distribution of topics
    $\theta_d \sim \text{Dir} (\alpha)$
  - for each word position $i$
    - choose topic $z_{d,i} \sim \theta_d$
    - choose word $w_{d,i} \sim \varphi_{z_{d,i}}$

- $\alpha$ and $\beta$ are hyperparameters
Latent Dirichlet Allocation (Blei et al, 03)

Generative model of text:

- for each topic $k$ choose distribution of words
  \[ \varphi_k \sim Dir(\beta) \]
- for each document $d$
  - choose distribution of topics
    \[ \theta_d \sim Dir(\alpha) \]
  - for each word position $i$
    - choose topic $z_{d,i} \sim \theta_d$
    - choose word $w_{d,i} \sim \varphi_{z_{d,i}}$

- $\alpha$ and $\beta$ are hyperparameters
Often results of classifiers with features space induced by $p(z|d)$ are better for LDA than for PLSA.

Defines generative models for any documents.

Exact inference is intractable, but efficient approximations are available:

- variational inference [Blei, et al 03]
- collapsed Gibbs sampling [Griffiths and Steyvers, 04]
- expectation-propagation [Minka and Lafferty, 02]

It still makes bag-of-the word assumption:

- no direct segmentation of text
- some topics are difficult to reveal on the basis of bag-of-word models
  - collection of user reviews of hotels
  - would like to get topics: "location", "staff", "cleanliness",...
  - get topics: "New York hotels", "hostels", "resorts",...
Outline

1. Bag-of-words Models
2. Beyond bag-of-words
   - Documents as Sequences of Windows
   - Topic N-grams
3. Topic hierarchies
4. Approximate Inference
HMM Segmentation

- HMM segmentation [Mulbregt et al, 98]: text is divided into $L$-word blocks

Again, as in distribution clustering, **naive Bayes assumption**: $P(w_1^t, \ldots, w_L^t | z^t) = \prod_i P(w_i^t | z^t)$

**Idea**: replace naive Bayes model with an aspect model

$P(w_1^t, \ldots, w_L^t | d^t) = \prod_i \sum z_i P(z_i^t | d^t) P(w_i^t | z_i^t)$
Aspect HMM:

- words generated through an aspect model
  \[ P(w_1^t, \ldots, w_n^t) = \prod_i \sum_{z_i} P(z_i^t | d^t) P(w_i^t | z_i^t) \]
  - model transitions between distributions \( \theta^t \)

- **Intractable**

  (Blei and Moreno, 01):
  - supervised! (segmentation during learning)
  - **very** coarse approximation
  - surprise: still works well
Switching Aspect HMM [Purver et al, 06]:

Variable $r$ switches: either topic distributions $\theta^t$ and $\theta^{t+1}$ are the same or independent:

$P(\theta^{t+1} | r^{t+1}, \theta^t) = \delta(\theta^t - \theta^{t+1}), r = 0$

$P(\theta^{t+1} | r^{t+1}, \theta^t) = \text{Dir}(\alpha), r = 1$

Advantage:

- Direct modeling of topic breaks
- Collapsed Gibbs-sampling is feasible

Drawbacks:

- Transition are not smooth enough
- E.g., topic mixture “History”+ “London” is more likely to be followed by “Geography” + “London”, than “Geography” + “Moscow”
Sliding (Overlapping) Windows [Titov and McDonald, 08]

- choose topic models $\phi \sim Dir(\beta)$
- for each document $d$
  - for each sliding window $t$ choose distribution of topics $\theta_t \sim Dir(\alpha)$
  - for each sentence $s$ choose distribution across windows $\psi_t \sim Dir(\gamma)$
  - for each word position $i$ from sentence $s$
    - choose window $v_{d,i} \sim \psi_s$
    - choose topic $z_{d,i} \sim \theta_{v_{d,i}}$
    - choose word $w_{d,i} \sim \phi_{z_{d,i}}$

Advantages:
- **Fast** inference (no explicit transition modeling)
- **Smooth** transitions
- **Co-occurrence domain** is larger than one window

Drawback:
- No explicit topic breaks
Sliding (Overlapping) Windows [Titov and McDonald, 08]

- Choose topic models \( \varphi \sim \text{Dir}(\beta) \)
- For each document \( d \)
  - For each sliding window \( t \) choose distribution of topics \( \theta_t \sim \text{Dir}(\alpha) \)
  - For each sentence \( s \) choose distribution across windows \( \psi_t \sim \text{Dir}(\gamma) \)
  - For each word position \( i \) from sentence \( s \)
    - Choose window \( v_{d,i} \sim \psi_s \)
    - Choose topic \( z_{d,i} \sim \theta_{v_{d,i}} \)
    - Choose word \( w_{d,i} \sim \varphi_{z_{d,i}} \)

**Advantages:**
- Fast inference (no explicit transition modeling)
- Smooth transitions
- Co-occurrence domain is larger than one window

**Drawback:**
- No explicit topic breaks
Sliding (Overlapping) Windows [Titov and McDonald, 08]

- choose topic models $\phi \sim Dir(\beta)$
- for each document $d$
  - for each sliding window $t$ choose distribution of topics $\theta_t \sim Dir(\alpha)$
  - for each sentence $s$ choose distribution across windows $\psi_t \sim Dir(\gamma)$
  - for each word position $i$ from sentence $s$
    - choose window $v_{d,i} \sim \psi_s$
    - choose topic $z_{d,i} \sim \theta_{v_{d,i}}$
    - choose word $w_{d,i} \sim \phi z_{d,i}$

Advantages:
- **Fast** inference (no explicit transition modeling)
- **Smooth** transitions
- **Co-occurrence domain** is larger than one window

Drawback:
- No explicit topic breaks
choose topic models $\varphi \sim \text{Dir}(\beta)$

for each document $d$
  
  for each sliding window $t$ choose distribution of topics $\theta_t \sim \text{Dir}(\alpha)$
  
  for each sentence $s$ choose distribution across windows $\psi_t \sim \text{Dir}(\gamma)$
  
  for each word position $i$ from sentence $s$
    
    choose window $v_{d,i} \sim \psi_s$
    
    choose topic $z_{d,i} \sim \theta_{v_{d,i}}$
    
    choose word $w_{d,i} \sim \varphi z_{d,i}$

Advantages:
- **Fast** inference (no explicit transition modeling)
- **Smooth** transitions
- **Co-occurrence domain** is larger than one window

Drawback:
- No explicit topic breaks
Sliding (Overlapping) Windows [Titov and McDonald, 08]

- choose topic models $\varphi \sim \text{Dir}(\beta)$
- for each document $d$
  - for each sliding window $t$ choose distribution of topics $\theta_t \sim \text{Dir}(\alpha)$
  - for each sentence $s$ choose distribution across windows $\psi_t \sim \text{Dir}(\gamma)$
  - for each word position $i$ from sentence $s$
    - choose window $v_{d,i} \sim \psi_s$
    - choose topic $z_{d,i} \sim \theta_{v_{d,i}}$
    - choose word $w_{d,i} \sim \varphi_{z_{d,i}}$

**Advantages:**
- Fast inference (no explicit transition modeling)
- Smooth transitions
- Co-occurrence domain is larger than one window

**Drawback:**
- No explicit topic breaks
Advantages:
- Can be fast

Non-overlapping windows (Blei et Moreno, 01; Purver et al, 06):
- Need to know where breaks can be
- Small windows - bad topics
- Large windows - bad granularity

Overlapping windows:
- No explicit topic breaks are induced
Topic bigram model [Wallach, 06]

- choose topic models $\varphi_{w,z} \sim \text{Dir}(\beta)$
- for each document $d$
  - choose distribution of topics $\theta \sim \text{Dir}(\alpha)$
  - for each word position $t$
    - choose topic $z^t \sim \theta$
    - choose word $w^t \sim \varphi_{w^{t-1},z^t}$

- Drawback:
  - very slow
  - no topic breaks

- Similar models:
  - [Gruber et al, 07] - as Switching aspect HMM but without windows
Topic bigram model [Wallach, 06]

choose topic models $\varphi_{w,z} \sim \text{Dir}(\beta)$

for each document $d$

- choose distribution of topics $\theta \sim \text{Dir}(\alpha)$
- for each word position $t$
  - choose topic $z^t \sim \theta$
  - choose word $w^t \sim \varphi_{w^{t-1},z^t}$

Drawback:

- very slow
- no topic breaks

Similar models:

- [Gruber et al, 07] - as Switching aspect HMM but without windows
choose topic models $\varphi_{w,z} \sim \text{Dir}(\beta)$

for each document $d$
- choose distribution of topics $\theta \sim \text{Dir}(\alpha)$
- for each word position $t$
  - choose topic $z^t \sim \theta$
  - choose word $w^t \sim \varphi_{w^{t-1},z^t}$

Drawback:
- very slow
- no topic breaks

Similar models:
- [Gruber et al, 07] - as Switching aspect HMM but without windows


**Topic bigram model** [Wallach, 06]

- Choose topic models $\varphi_{w,z} \sim \text{Dir}(\beta)$

- For each document $d$
  - Choose distribution of topics $\theta \sim \text{Dir}(\alpha)$
  - For each word position $t$
    - Choose topic $z^t \sim \theta$
    - Choose word $w^t \sim \varphi_{w^{t-1},z^t}$

- Drawback:
  - *very* slow
  - No topic breaks

- Similar models:
  - [Gruber et al, 07] - as Switching aspect HMM but without windows
According to [Griffiths et al, 05]:

- only a **subset** of words exhibit **long**-range dependencies
- **all** words exhibit **short** range “syntactic” dependencies
- it explains why standard combination of models does not work well [Coccaro and Jurafsky, 98]:
  - mixtures of models: **each** word exhibits either **long or short** range dependency
  - product of models: **each** word exhibits both **long and short** range dependency
Topic “Syntax” Model [Griffiths et al, 05]

- choose topic models $\varphi \sim \text{Dir}(\beta)$
- choose topic type transition distribution $\pi \sim \text{Dir}(\gamma)$
- for each document $d$
  - choose distribution of topics $\theta \sim \text{Dir}(\alpha)$
  - for each word position $t$
    - choose topic $z^t \sim \theta$
    - choose topic type $c^t \sim \pi_{c^{t-1}}$
    - if $c^t = 1$ choose word $w^t \sim \varphi_{z^t}$, else $w^t \sim \varphi_{c^t}$

network used for images
image obtained with kernel
output described with objects
neural network trained with svm images
**Topic “Syntax” Model [Griffiths et al, 05]**

- choose topic models $\varphi \sim Dir(\beta)$
- choose topic type transition distribution $\pi \sim Dir(\gamma)$
- for each document $d$
  - choose distribution of topics $\theta \sim Dir(\alpha)$
  - for each word position $t$
    - choose topic $z_t \sim \theta$
    - choose topic type $c_t \sim \pi_{c_{t-1}}$
    - if $c_t = 1$ choose word $w_t \sim \varphi_{z_t}^{(z)}$, else $w_t \sim \varphi_{c_t}^{(c)}$

Diagram:

- 0.5 network
- 0.4 image
- 0.1 kernel
- 0.8 in
- 0.4 with
- 0.5 for
- 0.1 on
- 0.7 used
- 0.9 trained
- 0.9 obtained
- 0.7 described

Network used for images
Image obtained with kernel
Output described with objects
Neural network trained with SVM images
choose topic models $\varphi \sim \text{Dir}(\beta)$
choose topic type transition distribution $\pi \sim \text{Dir}(\gamma)$
for each document $d$
choose distribution of topics $\theta \sim \text{Dir}(\alpha)$
for each word position $t$
choose topic $z^t \sim \theta$
choose topic type $c^t \sim \pi_{c^{t-1}}$
if $c^t = 1$ choose word $w^t \sim \varphi^{(z)}_{z^t}$, else $w^t \sim \varphi^{(c)}_{c^t}$
Topic “Syntax” Model [Griffiths et al, 05]

- choose topic models $\varphi \sim \text{Dir}(\beta)$
- choose topic type transition distribution $\pi \sim \text{Dir}(\gamma)$
- for each document $d$
  - choose distribution of topics $\theta \sim \text{Dir}(\alpha)$
  - for each word position $t$
    - choose topic $z^t \sim \theta$
    - choose topic type $c^t \sim \pi_{c^{t-1}}$
    - if $c^t = 1$ choose word $w^t \sim \varphi_{z^t}^{(z)}$, else $w^t \sim \varphi_{c^t}^{(c)}$
Varying Granularity

- We can extend this idea: we can model different types of 'topics' simultaneously: **each type will capture different phenomena**

> "... public transport in London is straightforward, the tube station is about an 8 minute walk ... or you can get a bus for £1.50 .... We had a stunning view (from the floor to ceiling window) of the Tower and the Thames."

- Global topic: **London**: words London, tube, ...
- Local topics: **Location** and **View**.
- Therefore, *local topics* correspond to aspects of reviewed items, *global topics* - to types of items

[Titov and McDonald, WWW 08]
Beyond bag-of-words

- Sequences of windows:
  - fast
  - require selecting window size

- Topic N-grams:
  - slow
  - capture syntax (not always desirable)
  - not easy to find topic breaks
Outline

1. Bag-of-words Models
2. Beyond bag-of-words
   - Documents as Sequences of Windows
   - Topic N-grams
3. Topic hierarchies
4. Approximate Inference
It is natural to organize topics into hierarchies

- Computer Science
  - Software engineering
    - Algorithm design
    - Reverse engineering
    - Formal methods
  - Artificial Intelligence
    - Machine learning
    - Natural language processing
    - Robotics
    - ...

...
Hierarchical LDA

Prior over trees/paths in trees (Nested Chinese Restaurant Processes)

Generative process:

- generate tree and topic models for each node in the tree $\varphi \sim \text{Dir}(\beta)$
- for each document:
  - generate a path from the root to a leaf in the tree
  - generate proportion vector $\theta \sim \text{Dir}(\alpha)$
  - for each word $i$
    - generate $z_i \sim \theta$
    - generate $w_i \sim \phi_{z_i}$

[Blei, Griffiths, Jordan and Tenenbaum, 2003]
Hierarchical LDA

Prior over trees/paths in trees (Nested Chinese Restaurant Processes)

Generative process:

- generate tree and topic models for each node in the tree $\varphi \sim Dir(\beta)$
- for each document:
  - generate a path from the root to a leaf in the tree
  - generate proportion vector $\theta \sim Dir(\alpha)$
  - for each word $i$
    - generate $z_i \sim \theta$
    - generate $w_i \sim \phi_{z_i}$

[Blei, Griffiths, Jordan and Tenenbaum, 2003]
Hierarchical LDA

Prior over trees/paths in trees (Nested Chinese Restaurant Processes)
Generative process:

- generate tree and topic models for each node in the tree $\varphi \sim \text{Dir}(\beta)$
- for each document:
  - generate a path from the root to a leaf in the tree
  - generate proportion vector $\theta \sim \text{Dir}(\alpha)$
  - for each word $i$
    - generate $z_i \sim \theta$
    - generate $w_i \sim \phi_{z_i}$

[Blei, Griffiths, Jordan and Tenenbaum, 2003]
Hierarchical LDA

Prior over trees/paths in trees (Nested Chinese Restaurant Processes)
Generative process:

- generate tree and topic models for each node in the tree \( \varphi \sim Dir(\beta) \)
- for each document:
  - generate a path from the root to a leaf in the tree
  - generate proportion vector \( \theta \sim Dir(\alpha) \)
  - for each word \( i \)
    - generate \( z_i \sim \theta \)
    - generate \( w_i \sim \phi_{z_i} \)

[Blei, Griffiths, Jordan and Tenenbaum, 2003]
Hierarchical LDA

Prior over trees/paths in trees (Nested Chinese Restaurant Processes)

Generative process:

- generate tree and topic models for each node in the tree $\varphi \sim \text{Dir}(\beta)$
- for each document:
  - generate a path from the root to a leaf in the tree
  - generate proportion vector $\theta \sim \text{Dir}(\alpha)$
  - for each word $i$
    - generate $z_i \sim \theta$
    - generate $w_i \sim \phi_{z_i}$

[Blei, Griffiths, Jordan and Tenenbaum, 2003]
Problems with hLDA

- hLDA assumes that the document is sampled from a **single path**
- Essentially, **similar to distributional clustering rather than LDA**
- E.g.: a text about machine learning for a biomedical problem cannot be properly expressed.
Hierarchical Pachecenko Allocation Model (model 2)

Hierarchy not a tree:

Generative process:
- choose topic models for each node $\varphi \sim \text{Dir}(\beta)$
- for each document:
  - choose topic distributions $\theta^{(1)}$ and $\theta^{(2)}$
  - for each word $i$
    - choose $z^{(1)} \sim \theta^{(1)}$, if $z^{(1)} = 0$ sample $w \sim \varphi^{(0)}$
    - Otherwise, choose $z^{(2)} \sim \theta^{(2)}$, if $z^{(2)} = 0$ sample $w \sim \varphi^{(1)}_{z^{(1)}}$
    - Otherwise, sample $w \sim \varphi^{(2)}_{z^{(2)}}$

[Mimno, Li and McCallum, 2007]
Hierarchical Pacheco Allocation Model (model 2)

Hierarchy not a tree:

Generative process:

- choose topic models for each node $\varphi \sim \text{Dir}(\beta)$
- for each document:
  - choose topic distributions $\theta^{(1)}$ and $\theta^{(2)}$
  - for each word $i$
    - choose $z^{(1)} \sim \theta^{(1)}$, if $z^{(1)} = 0$ sample $w \sim \varphi^{(0)}$
    - Otherwise, choose $z^{(2)} \sim \theta^{(2)}$, if $z^{(2)} = 0$ sample $w \sim \varphi_{z^{(1)}}^{(1)}$
    - Otherwise, sample $w \sim \varphi_{z^{(2)}}^{(2)}$

[Mimno, Li and McCallum, 2007]
Hierarchical Pacheco Allocation Model (model 2)

Hierarchy not a tree:

[Diagram of a non-tree hierarchy structure]

Generative process:

- choose topic models for each node \( \varphi \sim \text{Dir}(\beta) \)
- for each document:
  - choose topic distributions \( \theta^{(1)} \) and \( \theta^{(2)} \)
  - for each word \( i \)
    - choose \( z^{(1)} \sim \theta^{(1)} \), if \( z^{(1)} = 0 \) sample \( w \sim \varphi^{(0)} \)
    - Otherwise, choose \( z^{(2)} \sim \theta^{(2)} \), if \( z^{(2)} = 0 \) sample \( w \sim \varphi_{z^{(1)}}^{(1)} \)
    - Otherwise, sample \( w \sim \varphi_{z^{(2)}}^{(2)} \)

[Mimno, Li and McCallum, 2007]
Hierarchical Pachkenko Allocation Model (model 2)

Hierarchy not a tree:

Generative process:

- choose topic models for each node $\varphi \sim Dir(\beta)$
- for each document:
  - choose topic distributions $\theta^{(1)}$ and $\theta^{(2)}$
  - for each word $i$
    - choose $z^{(1)} \sim \theta^{(1)}$, if $z^{(1)} = 0$ sample $w \sim \varphi^{(0)}$
    - Otherwise, choose $z^{(2)} \sim \theta^{(2)}$, if $z^{(2)} = 0$ sample $w \sim \varphi^{(1)}_{z^{(1)}}$
    - Otherwise, sample $w \sim \varphi^{(2)}_{z^{(2)}}$

[Mimno, Li and McCallum, 2007]
Hierarchy not a tree:

Generative process:

- choose topic models for each node $\varphi \sim Dir(\beta)$
- for each document:
  - choose topic distributions $\theta^{(1)}$ and $\theta^{(2)}$
  - for each word $i$
    - choose $z^{(1)} \sim \theta^{(1)}$, if $z^{(1)} = 0$ sample $w \sim \varphi^{(0)}$
    - Otherwise, choose $z^{(2)} \sim \theta^{(2)}$, if $z^{(2)} = 0$ sample $w \sim \varphi_{z^{(1)}}^{(1)}$
    - Otherwise, sample $w \sim \varphi_{z^{(2)}}^{(2)}$

[Mimno, Li and McCallum, 2007]
Hierarchical Pacheco Allocation Model (model 2)

Hierarchy not a tree:

- choose topic models for each node $\varphi \sim \text{Dir}(\beta)$
- for each document:
  - choose topic distributions $\theta^{(1)}$ and $\theta^{(2)}$
  - for each word $i$
    - choose $z^{(1)} \sim \theta^{(1)}$, if $z^{(1)} = 0$ sample $w \sim \varphi^{(0)}$
    - Otherwise, choose $z^{(2)} \sim \theta^{(2)}$, if $z^{(2)} = 0$ sample $w \sim \varphi_{z^{(1)}}^{(1)}$
    - Otherwise, sample $w \sim \varphi_{z^{(2)}}^{(2)}$

[Mimno, Li and McCallum, 2007]
hPAM vs. hLDA

- **hLDA:**
  - Discovers tree
  - Generates structure
  - Single path per document

- **hPAM:**
  - Does not discover hierarchy
  - Uses a given structure
  - Mixture of arbitrary topic distributions

Why not to build hLDA with a mixture of paths?
Outline

1. Bag-of-words Models
2. Beyond bag-of-words
   - Documents as Sequences of Windows
   - Topic N-grams
3. Topic hierarchies
4. Approximate Inference
Approximate Inference

- **Approaches**
  - EM (PLSA): find ML estimates for $\varphi$ and $\theta$
  - Variational approach [Blei et al, 03]
  - Collapsed Gibbs sampling [Griffiths et al, 03]

- We consider LDA, but applicable to most models
Forget about priors for simplicity

\[ L(\theta, \varphi) = \sum_{d, w} n(d, w) \log \sum_z \theta_d(z) \varphi_z(w) \]

E step:

\[ P(z|d, w) \propto \theta_d(z) \varphi_z(w) \]

M step:

\[ \varphi_z(w) \propto \sum_d n(d, w) P(z|d, w) \]

\[ \theta_d(z) \propto \sum_w n(d, w) P(z|d, w) \]

In practice: entropy regularization is used to avoid overfitting
Variational Inference [Blei et al, 2003]

- Again, point estimate for $\varphi$ but not for $\theta$, therefore, distr’s factorize over documents
- Need: $P(\theta, z|w, \alpha, \varphi)$, $z$ - vector of all assignments of words to topics, $w$ - all words
- Consider an approximate distribution
  \[ P(\theta, z|w, \alpha, \varphi) \approx Q(\theta, z) = Q(\theta|\gamma) \prod_i Q(z_i|\psi_i) \]
- Lower bound: \[ \log P(w|\alpha, \varphi) \geq E_q[\log P(\theta, z, w|\alpha, \varphi)] - E_q[\log Q(\theta, z)] \]
- Select $\gamma$ and $\psi$ to maximize the bound (= to minimize $D(Q|P(\theta, z|w, \alpha, \varphi))$
- Maximize the bound by setting $\varphi$. 
**Collapsed Sampling [Griffiths et al, 2003]**

- **Gibbs sampling**: sequentially sample a variable given assignments of all other variables - converges to the joint distribution.
- Naive approach for LDA: sample distributions \( \theta \) and \( \varphi \), and assignments \( z \) - not feasible.
- Instead, **collapsed** Gibbs sampling - \( \theta \) and \( \varphi \) integrated out analytically:

\[
P(z_i = j | z_{-i}, w) \propto \frac{n(w_i, j) + \beta n(d_i, j) + \alpha}{n(j) + W\beta n(d_i) + T\alpha},
\]

where \( W \) - vocabulary size, \( T \) number of topics, \( n \) - counts which do not take into account word \( i \).

- **First factor**: choosing word for topic \( j \) for word \( w_i \).
- **Second factor**: choosing topic \( j \) for document \( d_i \).
Approximation: review

- **EM (PLSA):**
  - overfits
  - even with regularization results are not as good as with other approximations

- **Variational approximation:**
  - not always works very well
  - more complex $Q$ may be needed

- **Collapsed sampling:**
  - Not possible to aggregate over several samples
  - Not applicable to small collections
Conclusions

- Depending on your problem **different topic models are appropriate**:
  - Segmenting - probably Switching Aspect HMM
  - Multi-aspect sentiment analysis - probably overlapping windows
  - IR - probably, usual LDA
  - ...

- You can **combine ideas**:
  - e.g., Switching Aspect HMM + hPAM

- You can **model jointly**:
  - joint modeling of images and text
  - sentiment and topic

- You can go **beyond unsupervised**:
  - define priors “manually”
  - integrate available annotated data [Blei and McAuliffe, 08; Titov and McDonald, 08]

- Is it going to converge with **grammar induction**?
The aspect models found many applications outside NLP domain:

- Collaborative filtering [Hofmann, 04]
- Discovery of object types from images [Sivic et al, 05]
- Fraud detection (user modeling) [Xing and Girolami, 07]
- Microarray analysis [Rogers et al, 05]
- …