Natural Language Processing I

lecture 2: topic models
aka making sense of text collections

Ivan Titov

Institute for Logic, Language and Computation
A motivation for topic modeling
A simple topic model (PLSA)
A refresher for EM
EM for PLSA
Problems with PLSA => LDA
MCMC methods => Gibbs sampling

• By the end of this lecture you will know how to construct a topic model
• A modeling framework used within many solutions in the industry (e.g., analysis of news / reviews /... )
• A building block for many interesting models
NLP Problems:
- Doc. classification
- Topic analysis
- Shallow synt. parsing /tagging
- Syntactic parsing
- Relation extraction
- Semantic parsing
- Models of inference
- Machine translation
- Question answering
- Opinion analysis
- Summarization
- Dialogue systems

Types of structures:
- Sequences / Chains
- Bags
- Spanning trees
- Hierarchical trees
- DAGs
- Bipartite graphs

Models/Views:
- Naive Bayes
- Topic models
- HMMs
- History- / transition-based models
- PCFGs
- DOP
- Global scoring (e.g., MST)
- "IBM" models

Set-ups:
- Supervised estimation
- Unsupervised
- Partially/semi-supervised
- Representation learning (factorizations / NNs)

Modeling frameworks:
- Generative ML
- Generative Bayes
- Discriminative
- Discriminative Bayes

...
Problem set-up

- **Given:** a collection of documents
- **Want:**
  - Detect key ‘topics’ discussed in the collection
  - For each document detect which ‘topics’ are discussed there
- **Requirements:**
  - No supervision (documents are not labeled)
  - Probabilistic methods
- The models we will describe generalize beyond these goals
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
Latent Semantic Analysis [Deerwester et al., 1990]

- Decomposition (SVD) of the co-occurrence matrix $X = U\Sigma V^T$
- Approximate the co-occurrence matrix $X \approx \hat{X} = U_k \Sigma_k V_k^T$

- **Hope**: terms having common meaning are mapped to the same direction
- **Hope**: documents discussing similar topics have similar representation
- Non-zero inner products between documents with non-overlapping terms
Latent Semantic Analysis [Deerwester et al., 1990]

\[
X \approx \hat{X} = U_k \Sigma_k V_k^T
\]

- Optimal rank \(k\) approximation (Frobenius norm):

\[
\hat{X} = \arg \min_{X' : \text{rank}(X') = k} \|X - X'\|
\]
Latent Semantic Analysis [Deerwester et al., 1990]

\[ X \approx \hat{X} = U_k \Sigma_k V_k^T \]

- Not motivated probabilistically (no clean underlying probability model)
- No obvious interpretation of directions
Lecture

- A motivation for topic modeling
- A simple topic model (PLSA)
- A refresher for EM
- EM for PLSA
- Problems with PLSA => LDA
- MCMC methods => Gibbs sampling
- ...
Probabilistic LSA  [Hofmann, 99]

- **Parameters:**
  - Distributions of topics in document $P(z|i)$, for every $i$
  - Distribution of words for every topic, $P(w|z)$, $z \in \{1, \ldots, K\}$

- **Generative story:**
  - For each document $i$
    - For each word occurrence $j$ in document $i$
      - Select topic $z_j$ for the word from $P(z_j | i)$
      - Generate word $w_j$ from $P(w_j | z_j)$

- **Note:**
  - Words in the same documents can be generated from different topics
Probabilistic LSA: Example

Generative story: Given parameters generate text

Eruption $P(w \mid z = 1)$
- delays 0.003
- volcanic 0.002
- volcano 0.001
- ash 0.001
- cloud 0.001
- ...

Sport $P(w \mid z = 2)$
- football 0.004
- teams 0.002
- ball 0.002
- preparations 0.001
- Formula 1 0.001
- ...

Politics $P(w \mid z = 3)$
- Obama 0.005
- Merkel 0.001
- ceremony 0.001
- attend 0.0001
- party 0.0001
- ...

Document 1
... Delays due to the volcanic ash cloud will affect Formula 1 teams' preparations ...

Document 2
... Obama will not attend the ceremony due to delays caused by eruption ...

- Eruption
- Sport
- Politics
Probabilistic LSA: Example

... Delays due to the volcanic ash cloud will affect Formula1 teams preparations ...

... Obama will not attend the ceremony due to delays caused by eruption ...

**Eruption P(w | z = 1)**
- delays: 0.003
- volcanic: 0.002
- volcano: 0.001
- ash: 0.001
- cloud: 0.001
- ...
- ...

**Sport P(w | z = 2)**
- football: 0.004
- teams: 0.002
- ball: 0.002
- preparations: 0.001
- Formula1: 0.001
- ...
- ...

**Politics P(w | z = 3)**
- Obama: 0.005
- Merkel: 0.001
- ceremony: 0.001
- attend: 0.0001
- party: 0.0001
- ...
- ...
Probabilistic LSA : Example

\[ \text{Document 1} \]

... Delays due to the volcanic ash cloud will affect Formula 1 teams preparations ...

\[ \text{Document 2} \]

... Obama will not attend the ceremony due to delays caused by eruption ...

<table>
<thead>
<tr>
<th>Eruption $P(w \mid z = 1)$</th>
<th>Sport $P(w \mid z = 2)$</th>
<th>Politics $P(w \mid z = 3)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>delays 0.003</td>
<td>football 0.004</td>
<td>Obama 0.005</td>
</tr>
<tr>
<td>volcanic 0.002</td>
<td>teams 0.002</td>
<td>Merkel 0.001</td>
</tr>
<tr>
<td>volcano 0.001</td>
<td>ball 0.002</td>
<td>ceremony 0.001</td>
</tr>
<tr>
<td>ash 0.001</td>
<td>preparations 0.001</td>
<td>attend 0.0001</td>
</tr>
<tr>
<td>cloud 0.001</td>
<td>Formula 1 0.001</td>
<td>party 0.0001</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Probabilistic LSA: Example

**Document 1**

... *Delays* due to the volcanic ash cloud will affect Formula 1 teams preparations ...

**Document 2**

... Obama will not attend the ceremony due to delays caused by eruption ...

| Eruption $P(w | z = 1)$ | Sport $P(w | z = 2)$ | Politics $P(w | z = 3)$ |
|-------------------------|----------------------|------------------------|
| delays 0.003            | football 0.004       | Obama 0.005            |
| volcanic 0.002          | teams 0.002          | Merkel 0.001           |
| volcano 0.001           | ball 0.002           | ceremony 0.001         |
| ash 0.001               | preparations 0.001   | attend 0.0001          |
| cloud 0.001             | Formula 1 0.001      | party 0.0001           |
| ...                     | ...                  | ...                    |
Probabilistic LSA: Example

Document 1

... Delays due to the volcanic ash cloud will affect Formula 1 teams preparations ...

Document 2

... Obama will not attend the ceremony due to delays caused by eruption ...

Eruption $P(w | z = 1)$
- delays 0.003
- volcanic 0.002
- volcano 0.001
- ash 0.001
- cloud 0.001

Sport $P(w | z = 2)$
- football 0.004
- teams 0.002
- ball 0.002
- preparations 0.001
- Formula 1 0.001

Politics $P(w | z = 3)$
- Obama 0.005
- Merkel 0.001
- ceremony 0.001
- attend 0.0001
- party 0.0001
Probabilistic LSA : Example

... Delays due to the *volcanic* ash cloud will affect Formula1 teams preparations ...

... Obama will not attend the ceremony due to delays caused by eruption ...

**Eruption P(w | z = 1)**
- delays 0.003
- **volcanic** 0.002
- volcano 0.001
- ash 0.001
- cloud 0.001

**Sport P(w | z = 2)**
- football 0.004
- teams 0.002
- ball 0.002
- preparations 0.001
- Formula1 0.001

**Politics P(w | z = 3)**
- Obama 0.005
- Merkel 0.001
- ceremony 0.001
- attend 0.0001
- party 0.0001
... Delays due to the **volcanic ash** cloud will affect Formula 1 teams preparations ...

... Obama will not attend the ceremony due to delays caused by eruption ...

| Eruption P(w | z = 1) | Sport P(w | z = 2) | Politics P(w | z = 3) |
|-----------------------|-------------------|----------------------|
| delays                | football          | Obama                |
| volcanic              | teams             | 0.005                |
| volcano               | ball              | Merkel               |
| ash                   | preparations      | ceremony             |
| cloud                 | Formula 1         | attend               |
| ...                   | ...               | party                |
| ...                   | ...               | 0.0001               |
| ...                   | ...               | ...                  |
Probabilistic LSA: Example

**Document 1**

... Delays due to the **volcanic ash cloud** will affect Formula 1 teams' preparations ...

**Document 2**

... Obama will not attend the ceremony due to delays caused by eruption ...

| Eruption P(w | z = 1)          | Sport P(w | z = 2)          | Politics P(w | z = 3)          |
|-------------------------------|----------------------------|-------------------------------|
| delays 0.003                  | football 0.004             | Obama 0.005                   |
| volcanic 0.002                | teams 0.002                | Merkel 0.001                  |
| volcano 0.001                 | ball 0.002                 | ceremony 0.001               |
| ash 0.001                     | preparations 0.001         | attend 0.0001                 |
| cloud 0.001                   | Formula 1 0.001            | party 0.0001                  |
| ...                           | ...                        | ...                           |
Probabilistic LSA: Example

**Document 1**

... Delays due to the volcanic ash cloud will affect Formula 1 teams' preparations ...

**Document 2**

... Obama will not attend the ceremony due to delays caused by eruption ...

---

**Eruption P(w | z = 1)**

- delays: 0.003
- volcanic: 0.002
- volcano: 0.001
- ash: 0.001
- cloud: 0.001
- ...

**Sport P(w | z = 2)**

- football: 0.004
- teams: 0.002
- ball: 0.002
- preparations: 0.001
- Formula 1: 0.001
- ...

**Politics P(w | z = 3)**

- Obama: 0.005
- Merkel: 0.001
- ceremony: 0.001
- attend: 0.0001
- party: 0.0001
- ...

---
Probabilistic LSA: Example

... Delays due to the volcanic ash cloud will affect Formula1 teams preparations ...

... Obama will not attend the ceremony due to delays caused by eruption ...

<table>
<thead>
<tr>
<th>Eruption $P(w \mid z = 1)$</th>
<th>Sport $P(w \mid z = 2)$</th>
<th>Politics $P(w \mid z = 3)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>delays 0.003</td>
<td>football 0.004</td>
<td>Obama 0.005</td>
</tr>
<tr>
<td>volcanic 0.002</td>
<td>teams 0.002</td>
<td>Merkel 0.001</td>
</tr>
<tr>
<td>volcano 0.001</td>
<td>ball 0.002</td>
<td>ceremony 0.001</td>
</tr>
<tr>
<td>ash 0.001</td>
<td>preparations 0.001</td>
<td>attend 0.0001</td>
</tr>
<tr>
<td>cloud 0.001</td>
<td>Formula1 0.001</td>
<td>party 0.0001</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>


Probabilistic LSA : Example

... Delays due to the volcanic ash cloud will affect Formula 1 teams’ preparations ...

... Obama will not attend the ceremony due to delays caused by eruption ...

| Eruption P(w | z = 1)          | Sport P(w | z = 2)                | Politics P(w | z = 3)            |
|-------------------------------|---------------------------------|--------------------------------|
| delays: 0.003                 | football: 0.004                 | Obama: 0.005                   |
| volcanic: 0.002               | teams: 0.002                    | Merkel: 0.001                  |
| volcano: 0.001                | ball: 0.002                     | ceremony: 0.001                |
| ash: 0.001                    | preparations: 0.001             | attend: 0.0001                  |
| cloud: 0.001                  | Formula 1: 0.001                | party: 0.0001                   |
| ...                           | ...                             | ...                            |
Probabilistic LSA: Example

... Delays due to the volcanic ash cloud will affect Formula 1 teams' preparations

... Obama will not attend the ceremony due to delays caused by eruption ...

Eruption $P(w \mid z = 1)$
- delays: 0.003
- volcanic: 0.002
- volcano: 0.001
- ash: 0.001
- cloud: 0.001

Sport $P(w \mid z = 2)$
- football: 0.004
- teams: 0.002
- ball: 0.002
- preparations: 0.001
- Formula 1: 0.001

Politics $P(w \mid z = 3)$
- Obama: 0.005
- Merkel: 0.001
- ceremony: 0.001
- attend: 0.0001
- party: 0.0001
Probabilistic LSA: Example

... Delays due to the volcanic ash cloud will affect Formula 1 teams' preparations ...

... Obama will not attend the ceremony due to delays caused by eruption ...

<table>
<thead>
<tr>
<th>Eruption $P(w \mid z = 1)$</th>
<th>Sport $P(w \mid z = 2)$</th>
<th>Politics $P(w \mid z = 3)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>delays 0.003</td>
<td>football 0.004</td>
<td>Obama 0.005</td>
</tr>
<tr>
<td>volcanic 0.002</td>
<td>teams 0.002</td>
<td>Merkel 0.001</td>
</tr>
<tr>
<td>volcano 0.001</td>
<td>ball 0.002</td>
<td>ceremony 0.001</td>
</tr>
<tr>
<td>ash 0.001</td>
<td>preparations 0.001</td>
<td>attend 0.0001</td>
</tr>
<tr>
<td>cloud 0.001</td>
<td>Formula 1 0.001</td>
<td>party 0.0001</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Probabilistic LSA: Example

Document 1

... Delays due to the volcanic ash cloud will affect Formula 1 teams’ preparations ...

Document 2

... Obama will not attend the ceremony due to delays caused by eruption ...

Eruption $P(w \mid z = 1)$

- delays: 0.003
- volcanic: 0.002
- volcano: 0.001
- ash: 0.001
- cloud: 0.001
- ...

Sport $P(w \mid z = 2)$

- football: 0.004
- teams: 0.002
- ball: 0.002
- preparations: 0.001
- Formula 1: 0.001
- ...

Politics $P(w \mid z = 3)$

- Obama: 0.005
- Merkel: 0.001
- ceremony: 0.001
- attend: 0.0001
- party: 0.0001
- ...

...
PLSA: Example

- **Note:**
  - Words in the same documents can be generated from different topics
  - Does not take into account order, the following to texts are guaranteed to have the same probability under the model

```plaintext
delays due to the volcanic ash cloud will affect Formula 1 teams’ preparations
```

```plaintext
to ash due Formula 1 affect volcanic the will teams’ cloud preparations delays
```
We considered:

| Eruption $P(w | z = 1)$ | Sport $P(w | z = 2)$ | Politics $P(w | z = 3)$ |
|------------------------|----------------------|------------------------|
| delays 0.003           | football 0.004       | Obama 0.005            |
| volcanic 0.002         | teams 0.002          | Merkel 0.001           |
| volcano 0.001          | ball 0.002           | ceremony 0.001         |
| ash 0.001              | preparations 0.001   | attend 0.0001          |
| cloud 0.001            | Formula 1 0.001      | party 0.0001           |
| …                      | …                    | …                      |

Document 1

Document 2
In fact we are solving reverse problem:

Document 1

… Delays due to the volcanic ash cloud will affect Formula 1 teams’ preparations …

Document 2

… We are going to talk about learning in a moment

Eruption $P(w | z = 1)$

- delays 0.003
- volcanic 0.002
- volcano 0.001
- ash 0.001
- cloud 0.001

Sport $P(w | z = 2)$

- football 0.004
- teams 0.002
- ball 0.002
- preparations 0.001
- Formula 1 0.001

Politics $P(w | z = 3)$

- Obama 0.005
- Merkel 0.001
- ceremony 0.001
- attend 0.001
- party 0.0001

... ... ...
Directed Graphical Models

- Roughly, arrows denote conditional dependencies
- A generative story corresponds to a topological order on the graph
Plate Notation

\[
\begin{array}{c}
\text{Doc 1} \\
\text{Doc N}
\end{array}
\]
Plate Notation

\[ z \xrightarrow{\text{Generated from } P(z | d)} w \xrightarrow{\text{Generate from } P(w | z)} \]

\[ MN \]
Distributions are also variables...

\[ \theta_d = P(z|d) \]

\[ \varphi_k = P(w|z_k) \]
Summary so far

- We defined a generative model of document collections: \( P(W, Z|\varphi, \theta) \)

- Now, first we will consider how to do Maximum likelihood estimation, i.e.
  \[
  (\hat{\varphi}, \hat{\theta}) = \arg \max_{\varphi, \theta} P(W, Z|\varphi, \theta)
  \]

- Then we will consider Bayesian methods

- First, a refresher on EM for the discrete domain

We will discuss slightly later today why calling it a generative model of a collection is a bit misleading
Lecture

- A motivation for topic modeling
- A simple topic model (PLSA)
- A refresher for EM
- EM for PLSA
- Problems with PLSA => LDA
- MCMC methods => Gibbs sampling
- ...

33
EM is a class of algorithms that is used to estimate parameters in the presence of missing attributes.

Non-convex optimization, therefore:

- converges to a local maximum of the maximum likelihood function (or posterior distribution if we incorporate the prior distribution).
- can be very sensitive to the starting point

In PLSA we do not observe from each topic \( z \) a word is generated.
Three coin example

- We observe a series of coin tosses generated in the following way:
  - A person has three coins.
    - Coin 0: probability of Head is $q_1$ of Tail $q_2 = 1 - q_1$
    - Coin 1: probability of Head $p_1$
    - Coin 2: probability of Head $p_2$
  - Consider the following coin-tossing scenario
Three coin example

- Scenario:
  - Toss coin 0 (do not show it to anyone!).
  - If Head – toss coin 1 \( M \) times;
  - Else -- toss coin 2 \( M \) times.
  - Only the series of tosses are observed
    - HHHT, HTHT, HHHT, HTTH
  - What are the parameters of the coins? \( q_1, p_1, p_2 \)
  - There is no closed-form solution

Though check-out spectral methods / momentum methods if interested
Three coin example

- Scenario:
  - Toss coin 0 (do not show it to anyone!).
  - If Head – toss coin 1  M times;
  - Else -- toss coin2  M times.
  - Only the series of tosses are observed
    - HHHH, HTHT, HHHH, HTTH
  - What are the parameters of the coins  ?  \( q_1, p_1, p_2 \)
  - There is no closed-form solution

Though check-out spectral methods / momentum methods if interested
Key Intuition

- If we knew which of the data points (HHHT), (HTHT), (HTTH) came from Coin1 and which from Coin2, that would be trivial.

- Assume that you a \((p, 1-p)\) coin \(m\) times and get \(k\) heads and \(m-k\) tails.

- What would be the maximum likelihood estimate for \(p\)?
Key Intuition

- If we knew which of the data points (HHHT), (HTHT), (HTTH) came from Coin1 and which from Coin2, that would be trivial.

- Assume that you a \((p, 1-p)\) coin \(m\) times and get \(k\) heads and \(m-k\) tails.

- What would be the maximum likelihood estimate for \(p\)?

Derivation (was on the board):

\[
\arg \max_{p} L(p) = \arg \max_{p} \log(p^k (1-p)^{m-k}) = \arg \max_{p} k \log p + (m - k) \log(1 - p)
\]

\[
\frac{dL(p)}{dp} = \frac{k}{p} - \frac{(m-k)}{(1-p)} = 0
\]

Follows \(p = \frac{k}{m}\)

A general lesson learned:

(even if \(A, B \in \mathbb{R}^+\), not \(\mathbb{N}^+\))

\[
\arg \max_{p} (A \log p + B \log(1 - p)) = \frac{A}{A + B}
\]
Instead, use an iterative approach for estimating the parameters:

- **Guess the probability** that a data point came from Coin 1/2
- **Generate fictional labels**, weighted according to this probability.
- **Re-estimate** the initial parameter setting: set them to maximize the likelihood of these augmented data.

This process can be iterated and can be shown to converge to a local maximum of the likelihood function.
EM-algorithm: coins (E-step)

- We will assume (for a minute) that we know the parameters and use it to estimate which Coin a series came from.
- Then, we will use the prediction to estimate the most likely parameters and so on...
- What is the probability that the $i$th data point came from Coin 1?

$D_i$ - is the $i$-th head-tail sequence

\[ \mu_i(z_i) = p(z_i|D_i) = \frac{P(z_i)P(D_i|z_i)}{\sum_{z'=1}^{2} P(z')P(D_i|z')} \]

\[ = \frac{q_z p_{z_i} h_i (1 - p_{z_i})^{(M-h_i)}}{\sum_{z'=1}^{2} q_z' p_{z_i} h_i (1 - p_{z_i})^{(M-h_i)}} \]

The number of heads in series $i$
EM-algorithm: coins (M-step)

- If we would observe the coin 0, the likelihood would look like:
  \[
  L_c(p, q) = \sum_{i=1}^{N} \log P(D_i, z_i|p, q) = \sum_{i=1}^{N} \log P(z_i)P(D_i|z_i, p, q)
  \]

- We do not observe it, so we want to maximize incomplete likelihood
  \[
  L_I(p, q) = \sum_{i=1}^{N} \log \sum_{z=1}^{2} P(D_i, z|p, q) = \sum_{i=1}^{N} \log \sum_{z=1}^{2} P(z)P(D_i|z, p, q)
  \]

- This hard to optimize directly (no closed form solution)
  
  - Instead on each step we maximize expectation of the likelihood
    \( L_c \) over the coin name:
    \[
    E_\mu[L_c] = E_\mu[\sum_{i=1}^{N} \log P(D_i, z|p, q)] = \sum_{i=1}^{N} E_\mu[\log P(D_i, z|p, q)] \\
    = \sum_{i=1}^{N} \sum_{z=1}^{2} \mu_i(z) \log P(D_i, z|p, q)
    \]
EM-algorithm: coins (M-step) - contd

$$E_\mu[L_c] = \sum_{i=1}^N E_\mu[\log P(D_i, z|p, q)]$$

$$= \sum_{i=1}^N \sum_{z=1}^2 \mu_i(z) \log P(D_i, z|p, q) = \sum_{i=1}^N \sum_{z=1}^2 \mu_i(z) \log q(z)p_z^{h_i}(1 - p)^{(M-h_i)}$$

$$= \sum_{z=1}^2 \left( \sum_{i=1}^N \mu_i(z)h_i \log p_z + \mu_i(z)(M - h_i) \log (1 - p_z) \right) + \left( \sum_{i=1}^N \mu_i(1) \right) \log q_1 + \left( \sum_{i=1}^N \mu_i(2) \right) \log(1 - q_1)$$

- **Maximize these 2 terms to get** $p_1$ **and** $p_2$
- **Maximize these term to get** $q_1$ **and** $q_2 = 1 - q_1$

Recall: $\arg \max_p (A \log p + B \log(1 - p)) = \frac{A}{A + B}$

$$p_z = \frac{\sum_{i=1}^N \mu_i(z)h_i}{\sum_{i=1}^N \mu_i(z)h_i + \sum_{i=1}^N \mu_i(z)(M - h_i)} = \frac{\sum_{i=1}^N \mu_i(z)h_i}{\sum_{i=1}^N \mu_i(z)M}$$

$$q_1 = \frac{\sum_{i=1}^N \mu_i(1)}{\sum_{i=1}^N \mu_i(1) + \sum_{i=1}^N \mu_i(2)} = \frac{\sum_{i=1}^N \mu_i(1)}{N}$$
Now say we have two sets $\mathcal{X}$ and $\mathcal{Y}$ and a joint distribution $P(X, Y|\theta)$.

If we have fully observable data, i.e. pairs $(X_i, Y_i)$, then

$$L(\theta) = \sum_i \log P(X_i, Y_i|\theta)$$

If we have partially observable data, i.e. only $X_i$:

$$L(\theta) = \sum_i \log P(X_i|\theta) = \sum_i \log \sum_{Y \in \mathcal{Y}} P(X_i, Y|\theta)$$

The EM algorithm is the method for finding:

$$\hat{\theta} = \arg \max_{\theta} \sum_i \log \sum_{Y \in \mathcal{Y}} P(X_i, Y|\theta)$$
Iterative procedure is defined as \( \theta^t = \arg \max_{\theta} Q(\theta, \theta^{t-1}) \), where

\[
Q(\theta, \theta^{t-1}) = \sum_i \sum_{Y \in \mathcal{Y}} P(Y|X_i, \theta^{t-1}) \log P(X_i, Y|\theta)
\]

Intuition:
- Fill hidden variables according to \( P(Y|X_i, \theta^{t-1}) \)
- EM is guaranteed to converge to a local minimum (or a saddle point) of the likelihood
- If \( L(\theta) = \sum_i \log P(X_i, Y_i|\theta) \) has a closed-form solution, then
  \[
  \arg \max_{\theta} \sum_i \sum_{Y \in \mathcal{Y}} P(Y|X_i, \theta^{t-1}) \log P(X_i, Y|\theta) \text{ has a closed-form solution as well}
  \]
EM for PLSA

\[
(\theta_d = P(z|d))
\]

\[
(\varphi_k = P(w|z_k))
\]

**Sum over positions in the documents**

**Vocabulary size**

**The number of times word \( w \) appears in document \( i \)**

**E-step:**

\[
P(z|i, w) = \frac{P(z|i)P(w|z)}{\sum_{z' = 1}^{K} P(z'|i)P(w|z')} = \frac{\theta_i(z)\varphi_z(w)}{\sum_{z' = 1}^{K} \theta_i(z')\varphi_{z'}(w)} \propto \theta_i(z)\varphi_z(w)
\]

**M-step:**

\[
\varphi_z(w) \propto \sum_{i=1}^{N} C(i, w)P(z|i, w)
\]

\[
\theta_i(z) \propto \sum_{w=1}^{V} C(i, w)P(z|i, w)
\]
Summary so far

- We defined a generative model of document collections: \( P(W, Z|\varphi, \theta) \)
- We considered how to do ML estimation, i.e.

\[
(\hat{\varphi}, \hat{\theta}) = \arg \max_{\varphi, \theta} P(W, Z|\varphi, \theta)
\]

- Now we could visualize collections \( P(w|z) = \hat{\varphi}_z \)
- Estimate topic distributions in each document \( P(z|i) = \hat{\theta}_i \)
Example (Science collection)

- Top 10 words for 10 topics out of 128 (ordered by $P(w|z) = \hat{\phi}_z$)

<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>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topic</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<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>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topic</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<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>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topic</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<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>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topic</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<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>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topic</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<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>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topic</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>male</td>
<td>0.0558</td>
</tr>
<tr>
<td>females</td>
<td>0.0541</td>
</tr>
<tr>
<td>female</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>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topic</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<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>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topic</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<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>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topic</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<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>
Example: topics of a document

\[ P(z|i) = \hat{\theta}_i \]

Colored according to
\[
\arg \max_z P(z|i, w) = \arg \max_z \hat{\theta}_i(z) \hat{\phi}_z(w)
\]

... Delays due to the volcanic ash cloud will affect Formula 1 teams' preparations ...
Problems of PLSA

- Not really a probabilistic model of document collections
  - How do we compute the probability of an unseen document?

- It overfits
  - There are techniques to deal with it (e.g., temporal annealing) but they are not necessary pretty

- It does provide a direct way to encode some prior knowledge, e.g.:
  - There are only a few topics per document
  - Words have only few senses
  - ...

What can we do about it?
Lecture

- A motivation for topic modeling
- A simple topic model (PLSA)
- A refresher for EM
- EM for PLSA
- Problems with PLSA => LDA
- MCMC methods => Gibbs sampling
- ...
You saw the cab #199 from the train in a small Finnish town. Assuming that taxis are numbered sequentially, how many taxis are in the town?