Topic Models and Its Applications

Zhaochun Ren
University of Amsterdam
Outline

- Review of topic models
- Topic models for IR
- Other applications of topic models
- Conclusion & outlook
Outline

- Review of topic models
- Topic models for IR
- Other applications of topic models
- Conclusion & outlook
Model Document as a sample of mixed topics

From *cmo44*, in Jun 18, 2010 8:40 AM
Yep. Same here. Safari crashes many times per day on all sorts of websites. And in particularly on websites that contain embedded videos such as YouTube videos. But then Safari has always been unreliable - it's as bad on the iPhone.

From *bearduk*, in Jun 22, 2010 3:11 PM
Same here for me. It seems to happen on pages that have a number of images on them for me. Safari just cuts out and I get a blank screen for a second and then the iPad home screen. Sites like do:while that have a number of images seem to crash regularly.
Overview of topic models

× Idea?
  × find low-dimensional descriptions of high-dimensional text

× Topic models enable text document:
  × (1) Summarization: find concise restatement
  × (2) Similarity: evaluate closeness of text
Overview of topic models

- Application of topic models
  - Document Summarization
  - Facilitate navigation/browsing
  - Information retrieval model
  - Segment documents
  - Social media topic tracking
  - Event detection
  - etc.
Introduction to topic models

- Expectation Maximization
- Probabilistic LSA
- Latent Dirichlet Allocation
  - Variational EM
- Gibbs Sampling
Introduction to topic models

- Expectation Maximization
- Probabilistic LSA
- Latent Dirichlet Allocation
  - Variational EM
- Gibbs Sampling
The general EM algorithm

× Given a joint distribution $p(W, Z|\theta)$, the goal of EM algorithm is to maximal the likelihood of $p(W|\theta)$ with respect of $\theta$.

Iterate:

**E-step:** estimate $E(z)$ for each $z$, given $\theta^{old}$

**M-step:** estimate $\theta^{new}$ maximizing $\sum_Z p(Z|W, \theta^{old}) \ln p(W, Z|\theta)$
Introduction to topic models

- Expectation Maximization
- Probabilistic LSA
- Latent Dirichlet Allocation
  - Variational EM
- Gibbs Sampling
Probabilistic LSA

\[ p(d, w_n) = p(d) \sum_z p(w_n | z)p(z | d) \]

- Generative process:
  - select a document \( d \) with \( p(d) \)
  - select class \( z \) with \( p(z | d) \)
  - select word \( w \) with \( p(w | d) \)
### Probabilistic LSA

<table>
<thead>
<tr>
<th>universe</th>
<th>0.0439</th>
<th>drug</th>
<th>0.0672</th>
<th>cells</th>
<th>0.0675</th>
<th>sequence</th>
<th>0.0818</th>
<th>years</th>
<th>0.156</th>
</tr>
</thead>
<tbody>
<tr>
<td>galaxies</td>
<td>0.0375</td>
<td>patients</td>
<td>0.0493</td>
<td>stem</td>
<td>0.0478</td>
<td>sequences</td>
<td>0.0493</td>
<td>million</td>
<td>0.0556</td>
</tr>
<tr>
<td>clusters</td>
<td>0.0279</td>
<td>drugs</td>
<td>0.0444</td>
<td>human</td>
<td>0.0421</td>
<td>genome</td>
<td>0.033</td>
<td>ago</td>
<td>0.045</td>
</tr>
<tr>
<td>matter</td>
<td>0.0233</td>
<td>clinical</td>
<td>0.0346</td>
<td>cell</td>
<td>0.0309</td>
<td>dna</td>
<td>0.0257</td>
<td>time</td>
<td>0.0317</td>
</tr>
<tr>
<td>galaxy</td>
<td>0.0232</td>
<td>treatment</td>
<td>0.028</td>
<td>gene</td>
<td>0.025</td>
<td>sequencing</td>
<td>0.0172</td>
<td>age</td>
<td>0.0243</td>
</tr>
<tr>
<td>cluster</td>
<td>0.0214</td>
<td>trials</td>
<td>0.0277</td>
<td>tissue</td>
<td>0.0185</td>
<td>map</td>
<td>0.0123</td>
<td>year</td>
<td>0.024</td>
</tr>
<tr>
<td>cosmic</td>
<td>0.0137</td>
<td>therapy</td>
<td>0.0213</td>
<td>cloning</td>
<td>0.0169</td>
<td>genes</td>
<td>0.0122</td>
<td>record</td>
<td>0.0238</td>
</tr>
<tr>
<td>dark</td>
<td>0.0131</td>
<td>trial</td>
<td>0.0164</td>
<td>transfer</td>
<td>0.0155</td>
<td>transfer</td>
<td>0.0119</td>
<td>early</td>
<td>0.0233</td>
</tr>
<tr>
<td>light</td>
<td>0.0109</td>
<td>disease</td>
<td>0.0115</td>
<td>blood</td>
<td>0.0113</td>
<td>regions</td>
<td>0.0119</td>
<td>billion</td>
<td>0.0177</td>
</tr>
<tr>
<td>density</td>
<td>0.01</td>
<td>medical</td>
<td>0.0097</td>
<td>embryos</td>
<td>0.0111</td>
<td>human</td>
<td>0.0111</td>
<td>history</td>
<td>0.0148</td>
</tr>
<tr>
<td>bacteria</td>
<td>0.0983</td>
<td>male</td>
<td>0.0558</td>
<td>theory</td>
<td>0.0811</td>
<td>immune</td>
<td>0.0909</td>
<td>stars</td>
<td>0.0524</td>
</tr>
<tr>
<td>bacterial</td>
<td>0.0561</td>
<td>females</td>
<td>0.0541</td>
<td>physics</td>
<td>0.0782</td>
<td>response</td>
<td>0.0375</td>
<td>star</td>
<td>0.0458</td>
</tr>
<tr>
<td>resistance</td>
<td>0.0431</td>
<td>female</td>
<td>0.0529</td>
<td>physicists</td>
<td>0.0146</td>
<td>system</td>
<td>0.0358</td>
<td>astrophys</td>
<td>0.0237</td>
</tr>
<tr>
<td>coli</td>
<td>0.0381</td>
<td>males</td>
<td>0.0477</td>
<td>einstein</td>
<td>0.0142</td>
<td>responses</td>
<td>0.0322</td>
<td>mass</td>
<td>0.021</td>
</tr>
<tr>
<td>strains</td>
<td>0.025</td>
<td>sex</td>
<td>0.0339</td>
<td>university</td>
<td>0.0113</td>
<td>antigen</td>
<td>0.0263</td>
<td>disk</td>
<td>0.0173</td>
</tr>
<tr>
<td>microbiol</td>
<td>0.0214</td>
<td>reproductive</td>
<td>0.0172</td>
<td>gravity</td>
<td>0.013</td>
<td>antigens</td>
<td>0.0184</td>
<td>black</td>
<td>0.0161</td>
</tr>
<tr>
<td>microbial</td>
<td>0.0196</td>
<td>offspring</td>
<td>0.0168</td>
<td>black</td>
<td>0.0127</td>
<td>immunity</td>
<td>0.0176</td>
<td>gas</td>
<td>0.0149</td>
</tr>
<tr>
<td>strain</td>
<td>0.0165</td>
<td>sexual</td>
<td>0.0166</td>
<td>theories</td>
<td>0.01</td>
<td>immunity</td>
<td>0.0145</td>
<td>stellar</td>
<td>0.0127</td>
</tr>
<tr>
<td>salmonella</td>
<td>0.0163</td>
<td>reproduction</td>
<td>0.0143</td>
<td>aps</td>
<td>0.00987</td>
<td>antibody</td>
<td>0.014</td>
<td>astron</td>
<td>0.0125</td>
</tr>
<tr>
<td>resistant</td>
<td>0.0145</td>
<td>eggs</td>
<td>0.0138</td>
<td>matter</td>
<td>0.00954</td>
<td>autoimmune</td>
<td>0.0128</td>
<td>hole</td>
<td>0.00824</td>
</tr>
</tbody>
</table>
Introduction to topic models

- Expectation Maximization
- Probabilistic LSA
- **Latent Dirichlet Allocation**
  - Variational EM
- Gibbs Sampling
Latent Dirichlet Allocation

• Each topic can be represented as a finite mixture of words

• Each document in the corpus can be represented as a mixture of multiple topics

• “Bag of words” assumption
Latent Dirichlet Allocation

- Generative process:
  - For each topic $z$:
    - Derive distribution $\phi \sim \text{Dirichlet}(\beta)$
  - For each document $d$:
    - Derive $\theta \sim \text{Dirichlet}(\alpha)$
  - For each word $w$:
    - Derive topic $z \sim \text{Multinomial}(\theta)$
    - Derive word $w \sim \text{Multinomial}(\phi)$
Approximate inference

- LDA model is intractable to get the exact posterior distribution
- Two main approximate inference methods:
  - variational EM algorithm
  - Gibbs sampling
Gibbs sampling

\[ p(\mathbf{w}, \mathbf{z} | \alpha, \beta) = \left( \frac{\Gamma \left( \sum_{v \in V} \beta_v \right)}{\prod_{v=1}^{V} \Gamma \beta_v} \right)^K \left( \frac{\Gamma \left( \sum_{k \in K} \alpha_k \right)}{\prod_{k=1}^{K} \Gamma \alpha_k} \right)^M \]

\times \prod_{k=1}^{K} \frac{\prod_{v=1}^{V} \Gamma (n_{k,v} + \beta_v)}{\Gamma \left( \sum_{v=1}^{V} (n_{k,v} + \beta_v) \right)} \prod_{m=1}^{M} \frac{\prod_{k=1}^{K} \Gamma (n_{m,k} + \alpha_k)}{\Gamma \left( \sum_{k=1}^{K} (n_{m,k} + \alpha_k) \right)}

\[ p(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^v + \beta_v}{\sum_{v=1}^{V} (n_{-i,j}^v + \beta_v)} \frac{n_{-i,j}^m + \alpha_k}{\sum_{k=1}^{K} (n_{-i,k}^m + \alpha_k)} \]
Outline

- Review of topic models
- Topic models for IR
- Other applications of topic models
- Conclusion & outlook
LDA-Based Document Models for Ad-hoc Retrieval

- Motivation
- Method
- Experimental results
- Conclusion
LDA-Based Document Models for Ad-hoc Retrieval

× Motivation

1. Use topic models for document representation is an area of considerable interest;
2. Document clustering is helpful for the language modeling framework.
3. Assumption in Unigram model is too simple to effectively model a large-scale collection of documents.
Overview of cluster-based retrieval

- the overall likelihood of observing the document \( d \) from the cluster model is

\[
P(w_1, w_2, \ldots, w_{N_d}) = \sum_{z=1}^{K} P(z) \prod_{i=1}^{N_d} P(w_i | z)
\]

- By incorporating the cluster information into language model as smoothing, we have:

\[
P(w | D) = \frac{N_d}{N_d + \mu} P_{ML}(w | D) + (1 - \frac{N_d}{N_d + \mu}) P(w | \text{cluster})
\]
LDA-Based Document Models for Ad-hoc Retrieval

- Method
  - LDA topic model
  - LDA-based Retrieval
  - Complexity
LDA-Based Document Models for Ad-hoc Retrieval

- Method
  - LDA topic model
  - LDA-based Retrieval
  - Complexity
LDA retrieval model

• Query likelihood model

\[ P(Q|D) = \prod_{q \in Q} P(q|D) \]
**LDA retrieval model**

- **Query likelihood model**

\[
P(Q|D) = \prod_{q \in Q} P(q|D)
\]

- **Combine LDA model with cluster retrieval model**

\[
P(w|D) = \lambda \left( \frac{N_d}{N_d + \mu} P_{ML}(w|D) + (1 - \frac{N_d}{N_d + \mu}) P(w|c) \right) + (1 - \lambda) P_{LDA}(w|D)
\]
LDA retrieval model

× For each document \( d \):

\[
P(w|d, \hat{\theta}, \hat{\phi}) = \sum_{z=1}^{\infty} P(w|z, \hat{\phi}) P(z|\hat{\theta}, d)
\]

\[
P_{LDA}(w|D) = \prod_{d \in D} \sum_{z=1} \frac{n_{-i,j}^{(w)} + \beta}{\sum_{v} (n_{-i,j}^{(v)} + \beta)} \frac{n_{-i,j}^{(d)} + \alpha}{\sum_{k} (n_{-i,k}^{(d)} + \alpha)}
\]
LDA-Based Document Models for Ad-hoc Retrieval

- Experimental setup

Table 1. Statistics of data sets

<table>
<thead>
<tr>
<th>Collection</th>
<th>Contents</th>
<th># of docs</th>
<th>Size</th>
<th>Queries</th>
<th># of Queries with Relevant Docs</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>Associated Press newswire 1988-90</td>
<td>242,918</td>
<td>0.73Gb</td>
<td>TREC topics 51-150 (title only)</td>
<td>99</td>
</tr>
<tr>
<td>FT</td>
<td>Financial Times 1991-94</td>
<td>210,158</td>
<td>0.56Gb</td>
<td>TREC topics 301-400 (title only)</td>
<td>95</td>
</tr>
<tr>
<td>SJMN</td>
<td>San Jose Mercury News 1991</td>
<td>90,257</td>
<td>0.29Gb</td>
<td>TREC topics 51-150 (title only)</td>
<td>94</td>
</tr>
<tr>
<td>LA</td>
<td>LA Times</td>
<td>131,896</td>
<td>0.48Gb</td>
<td>TREC topics 301-400 (title only)</td>
<td>98</td>
</tr>
<tr>
<td>WSJ</td>
<td>Wall Street Journal 1987-92</td>
<td>173,252</td>
<td>0.51Gb</td>
<td>TREC topics 51-100 &amp; 151-200 (title only)</td>
<td>100</td>
</tr>
</tbody>
</table>
LDA-Based Document Models for Ad-hoc Retrieval

<table>
<thead>
<tr>
<th>Rel. Retr.</th>
<th>QL</th>
<th>CBDM</th>
<th>LBDM</th>
<th>%chg over QL</th>
<th>%chg over CBDM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>21819</td>
<td>21819</td>
<td>21819</td>
<td>+10.09*</td>
<td>+12.21*</td>
</tr>
<tr>
<td>0.10</td>
<td>0.6422</td>
<td>0.6485</td>
<td>0.6795</td>
<td>+5.8*</td>
<td>+4.8*</td>
</tr>
<tr>
<td>0.20</td>
<td>0.3477</td>
<td>0.3713</td>
<td>0.4131</td>
<td>+18.8*</td>
<td>+11.2*</td>
</tr>
<tr>
<td>0.30</td>
<td>0.2977</td>
<td>0.317</td>
<td>0.3661</td>
<td>+23.0*</td>
<td>+15.5*</td>
</tr>
<tr>
<td>0.40</td>
<td>0.2454</td>
<td>0.2668</td>
<td>0.311</td>
<td>+26.8*</td>
<td>+16.6*</td>
</tr>
<tr>
<td>0.50</td>
<td>0.2081</td>
<td>0.2274</td>
<td>0.2666</td>
<td>+28.1*</td>
<td>+17.2*</td>
</tr>
<tr>
<td>0.60</td>
<td>0.1696</td>
<td>0.1794</td>
<td>0.2245</td>
<td>+32.4*</td>
<td>+25.1*</td>
</tr>
<tr>
<td>0.70</td>
<td>0.1298</td>
<td>0.1444</td>
<td>0.1665</td>
<td>+28.3*</td>
<td>+15.3*</td>
</tr>
<tr>
<td>0.80</td>
<td>0.0865</td>
<td>0.1002</td>
<td>0.118</td>
<td>+36.5*</td>
<td>+17.8*</td>
</tr>
<tr>
<td>0.90</td>
<td>0.0480</td>
<td>0.0571</td>
<td>0.0694</td>
<td>+44.7</td>
<td>+21.6</td>
</tr>
<tr>
<td>1.00</td>
<td>0.0220</td>
<td>0.0201</td>
<td>0.0187</td>
<td>-15.1</td>
<td>-6.8</td>
</tr>
<tr>
<td>Avg</td>
<td>0.2179</td>
<td>0.2326</td>
<td>0.2651</td>
<td>+21.64*</td>
<td>+13.97*</td>
</tr>
</tbody>
</table>
Topic models for query expansion

- Method
- Experiments
- Conclusion
Topic models for query expansion

- Method
- Experiments
- Conclusion
Topic models for query expansion

- Use topic models to calculate multinomial distribution $\theta = p(w|q)$ for a given query $q = \{q_1, q_2 \ldots q_k\}$
- Following RM method, investigate whether topics can be used for query expansion

$$p_{TM}(w|q) = \sum_{t_i} p_{TM}(w|t_i)p_{TM}(t_i|q)$$
Topic models for query expansion

- Combine topics models with RM query expansion method

\[
p(w|q) = \sum_{D_i \in C} (\gamma p_{RM}(w|D_i) + (1 - \gamma)p_{TM}(w|D_i, q) \times p(D_i|q))
\]

\[
p_{TM}(w|D_i, q) = \sum_{t_m} p(w|t_m)p(t_m|D_i, q) \propto \sum_{t_m} p(w|t_m)p(t_m|D_i)p(q|t_m)
\]
Topic models for query expansion

- Method
- Experiments
- Conclusion
Table 2. Retrieval Performance with TREC topics 301-400 (title-only) on one testing corpus (FT) by using different topic models for query expansion and for document smoothing. There are overall 3233 relevant documents. Bold font highlights the best result in each column. Parameters tuned on the training corpus for using typical topic models on the top feedback documents are not well generalized to this FT testing corpus: Q-CBQE, Q-LBQE, MFB perform worse than the QL baseline.

<table>
<thead>
<tr>
<th></th>
<th>Rel. Retr.</th>
<th>Interpolated Recall - Precision</th>
<th>Precision:</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.00</td>
<td>0.10</td>
<td>0.20</td>
</tr>
<tr>
<td>QL</td>
<td>1879</td>
<td>0.6142</td>
<td>0.4615</td>
<td>0.3987</td>
</tr>
<tr>
<td>CBDM</td>
<td>2092</td>
<td>0.6057</td>
<td>0.4766</td>
<td>0.4106</td>
</tr>
<tr>
<td>BT-LBDM</td>
<td>2074</td>
<td>0.6082</td>
<td>0.4821</td>
<td>0.4068</td>
</tr>
<tr>
<td>BT-PBDM</td>
<td>2034</td>
<td>0.6147</td>
<td>0.4747</td>
<td>0.4127</td>
</tr>
<tr>
<td>RMDE-1</td>
<td>1946</td>
<td>0.6067</td>
<td>0.4832</td>
<td>0.4329</td>
</tr>
<tr>
<td>MBDM</td>
<td>2099</td>
<td>0.5983</td>
<td>0.4810</td>
<td>0.4076</td>
</tr>
<tr>
<td>LBDM</td>
<td>2216</td>
<td>0.6338</td>
<td>0.4899</td>
<td>0.4072</td>
</tr>
<tr>
<td>PBDM</td>
<td>2226</td>
<td><strong>0.6341</strong></td>
<td><strong>0.4993</strong></td>
<td>0.4201</td>
</tr>
<tr>
<td>RMDE</td>
<td>2134</td>
<td>0.6320</td>
<td>0.4914</td>
<td>0.4294</td>
</tr>
<tr>
<td>CBQE</td>
<td>2016</td>
<td>0.6067</td>
<td>0.4681</td>
<td>0.4005</td>
</tr>
<tr>
<td>LBQE</td>
<td>2007</td>
<td>0.6198</td>
<td>0.4770</td>
<td>0.4034</td>
</tr>
<tr>
<td>PBQE</td>
<td>1981</td>
<td>0.6203</td>
<td>0.4648</td>
<td>0.3917</td>
</tr>
<tr>
<td>Q-CBQE</td>
<td>2151</td>
<td>0.5638</td>
<td>0.4517</td>
<td>0.3719</td>
</tr>
<tr>
<td>Q-LBQE</td>
<td>2028</td>
<td>0.5856</td>
<td>0.4671</td>
<td>0.3709</td>
</tr>
<tr>
<td>MFB</td>
<td>2283</td>
<td>0.5351</td>
<td>0.4303</td>
<td>0.3658</td>
</tr>
<tr>
<td>LDA-RM</td>
<td>2266</td>
<td>0.6113</td>
<td>0.4874</td>
<td>0.4295</td>
</tr>
<tr>
<td>RM</td>
<td><strong>2313</strong></td>
<td>0.6103</td>
<td>0.4844</td>
<td>0.4326</td>
</tr>
</tbody>
</table>
Topic models for query expansion

- Method
- Experiments
- Conclusion
Topic models for query expansion

- Topics discovered in the query related feedback documents can help retrieval
- Topic-based methods using these query related topics for retrieval are sensitive to parameters
- Method that combines RM successfully improves some queries’ results
Outline

- Review of topic models
- Topic models for IR
- Other applications of topic models
- Conclusion & outlook
Topic models in social text streams

- Dynamic topic modeling on social streams
- User behavior modeling on Twitter
- Topic models for entity linking
Topic modeling in social text streams

- Dynamic topic modeling on social streams
- User behavior modeling on Twitter
- Topic models for entity linking
Dynamic topic modeling

Social text streams: concept drifting phenomenon

- **Input data:**
  - input documents $X_1$
  - input documents $X_2$
  - …
  - input documents $X_i$

- **Output:**
  - topic distribution $p(z|t)$ at each time period $t$
Application: Hierarchical multi-label classification on social text streams

Hierarchical multi-label classification
• learn a hypothesis function \( f : \mathcal{X} \to \{0, 1\}^C \) from training data \( \{(x^{(i)}, y^{(i)})\}_{i=1}^{|D|} \) to predict a \( y \) when given input document \( x \)
• Follow \( T\)-property

some social texts streams belong to to hierarchical multiple labels

There are quite cramped trains
I really feel like Smullers
I think the train will soon stop again because of snow...
200,000 people travel with book as ticket
Hierarchical multi-label classification on social text streams

Hierarchical multi-label classification for short documents in social streams
- Learn from previous time periods, and predict an output when a new document arrives
- Concept drift phenomenon

- Document expansion
- Dynamic topic modeling (DTMs, Blei et al 2006)
- Structural learning based text classification
Experimental setup

- Dataset
  - tweets related to a transportation company
  - from 18th January 2010 to 5th June 2012
  - 145,692 tweets posted by 77,161 Twitter users
  - annotations 493 nodes in 13 subsets
Time-aware topic extraction (1)
Topic modeling in social text streams

- Dynamic topic modeling on social streams
- User behavior modeling on Twitter
- Topic models for entity linking
Application: personalized time-aware tweets summarization

- **Time-aware tweets summarization**
  Select the most representative tweets for each time period as summary

- **Personalized time-aware tweets summarization**
  Summary needs to be relevant to user’s interests

- **Data preprocessing**
- **Tweets propagation model**
- **Document summarization: sentence extraction**
Tweet Propagation Model

Key idea
Tweet Propagation Model

- Probability of user’s interests at each time period
- Probability of topics at each time period
Tweet Propagation Model (1)

A tweet can be represented using probabilistic distributions $p(z \mid p,t)$

$$Z_t = Z_t^u \cup Z_t^{com} \cup Z_t^B$$

- $z \in Z_t^u$ private topics $\theta_{u,t}^u \sim \theta_{u,t-1}^u$
- $z \in Z_t^{com}$ common topics $\theta_{u,t}^{com} \sim \left\{ \theta_{u_i,t-1}^{com} \mid u_i \in C_u \right\}$
- $z \in Z_t^B$ burst topics
Social-aware interests

\[ p(z | u_{t-1}) \]

Time-aware topics tracking

\[ p(w | z_{t-1}) \]
Tweet Propagation Model (2)

Gibbs EM sampling is used in the inference process.

Gibbs sampling + EM parameter estimation: $\alpha_{u,t}$ and $\beta_t$

Derive parameter $\lambda_{u,c,t}$ via Markov random walks

$$\lambda_{u,c,t} = \mu \sum_{i \neq j} \text{sim}(\theta_{c_i,t}, \theta_{c_j,t} | \theta_{u,t}) \cdot \lambda_{u,c_i,t} + \frac{(1-\mu)}{|C_{u,t}|}$$
Approach

Preprocessing & data enrichment

Dynamic topic models based on author topic model

- Detect dynamic interests
- Detect dynamic topics
- Novelty, coverage, diversity
- Relevant to user’s interests

Personalized time-aware tweets summarization
### Overall performance

<table>
<thead>
<tr>
<th>Metrics</th>
<th>TPM-A</th>
<th>TPM-T</th>
<th>TPM-S</th>
<th>UBM</th>
<th>TLDA</th>
<th>AT</th>
<th>TF-IDF</th>
<th>Centroid</th>
<th>Lex-R</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-1</td>
<td>0.428</td>
<td>0.403</td>
<td>0.395</td>
<td>0.387</td>
<td>0.374</td>
<td>0.355</td>
<td>0.341</td>
<td>0.302</td>
<td>0.291</td>
</tr>
<tr>
<td>R-2</td>
<td>0.125</td>
<td>0.119</td>
<td>0.116</td>
<td>0.114</td>
<td>0.112</td>
<td>0.102</td>
<td>0.095</td>
<td>0.081</td>
<td>0.077</td>
</tr>
<tr>
<td>R-W</td>
<td>0.159</td>
<td>0.153</td>
<td>0.149</td>
<td>0.146</td>
<td>0.142</td>
<td>0.144</td>
<td>0.137</td>
<td>0.118</td>
<td>0.115</td>
</tr>
</tbody>
</table>

**40 tweets per period**

<table>
<thead>
<tr>
<th>Metrics</th>
<th>TPM-A</th>
<th>TPM-T</th>
<th>TPM-S</th>
<th>UBM</th>
<th>TLDA</th>
<th>AT</th>
<th>TF-IDF</th>
<th>Centroid</th>
<th>Lex-R</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-1</td>
<td>0.513</td>
<td>0.497</td>
<td>0.482</td>
<td>0.461</td>
<td>0.374</td>
<td>0.423</td>
<td>0.411</td>
<td>0.362</td>
<td>0.369</td>
</tr>
<tr>
<td>R-2</td>
<td>0.149</td>
<td>0.142</td>
<td>0.139</td>
<td>0.134</td>
<td>0.127</td>
<td>0.122</td>
<td>0.116</td>
<td>0.097</td>
<td>0.095</td>
</tr>
<tr>
<td>R-W</td>
<td>0.197</td>
<td>0.191</td>
<td>0.189</td>
<td>0.178</td>
<td>0.176</td>
<td>0.166</td>
<td>0.161</td>
<td>0.131</td>
<td>0.119</td>
</tr>
</tbody>
</table>

**60 tweets per period**
Topic modeling in social text streams

- Dynamic topic modeling on social streams
- User behavior modeling on Twitter
- Topic models for entity linking
Author topic model

distribution of topics over words

distribution of authors over topics

uniform distribution of documents over authors
Author topic model

× For each topic $k$
  × Draw a multinomial $\phi$ from Dirichlet prior $\beta$
× For each author $a$
  × Draw a multinomial $\theta_a$ from Dirichlet prior $\alpha$
× For each document $m$
  × For each word $n$
    × Draw an author $x$ from $a$
    × Draw a topic $z_{m,n}$ from $\theta_x$
    × Draw a word $w_{m,n}$ from multinomial $\phi_{z_{m,n}}$
### Author topic model

<table>
<thead>
<tr>
<th>TOPIC 19</th>
<th>WORD</th>
<th>PROB.</th>
<th>TOPIC 24</th>
<th>WORD</th>
<th>PROB.</th>
<th>TOPIC 29</th>
<th>WORD</th>
<th>PROB.</th>
<th>TOPIC 87</th>
<th>WORD</th>
<th>PROB.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LIKELIHOOD</td>
<td>0.0539</td>
<td></td>
<td>RECOGNITION</td>
<td>0.0400</td>
<td></td>
<td>REINFORCEMENT</td>
<td>0.0411</td>
<td></td>
<td>KERNEL</td>
<td>0.0683</td>
</tr>
<tr>
<td></td>
<td>MIXTURE</td>
<td>0.0509</td>
<td></td>
<td>CHARACTER</td>
<td>0.0336</td>
<td></td>
<td>POLICY</td>
<td>0.0371</td>
<td></td>
<td>SUPPORT</td>
<td>0.0377</td>
</tr>
<tr>
<td></td>
<td>EM</td>
<td>0.0470</td>
<td></td>
<td>CHARACTERS</td>
<td>0.0250</td>
<td></td>
<td>ACTION</td>
<td>0.0332</td>
<td></td>
<td>VECTOR</td>
<td>0.0257</td>
</tr>
<tr>
<td></td>
<td>DENSITY</td>
<td>0.0398</td>
<td></td>
<td>TANGENT</td>
<td>0.0241</td>
<td></td>
<td>OPTIMAL</td>
<td>0.0208</td>
<td></td>
<td>KERNELS</td>
<td>0.0217</td>
</tr>
<tr>
<td></td>
<td>GAUSSIAN</td>
<td>0.0349</td>
<td></td>
<td>HANDWRITTEN</td>
<td>0.0169</td>
<td></td>
<td>ACTIONS</td>
<td>0.0208</td>
<td></td>
<td>SET</td>
<td>0.0205</td>
</tr>
<tr>
<td></td>
<td>ESTIMATION</td>
<td>0.0314</td>
<td></td>
<td>DIGITS</td>
<td>0.0159</td>
<td></td>
<td>FUNCTION</td>
<td>0.0178</td>
<td></td>
<td>SVM</td>
<td>0.0204</td>
</tr>
<tr>
<td></td>
<td>LOG</td>
<td>0.0263</td>
<td></td>
<td>IMAGE</td>
<td>0.0157</td>
<td></td>
<td>REWARD</td>
<td>0.0165</td>
<td></td>
<td>SPACE</td>
<td>0.0188</td>
</tr>
<tr>
<td></td>
<td>MAXIMUM</td>
<td>0.0254</td>
<td></td>
<td>DISTANCE</td>
<td>0.0153</td>
<td></td>
<td>SUTTON</td>
<td>0.0164</td>
<td></td>
<td>MACHINES</td>
<td>0.0168</td>
</tr>
<tr>
<td></td>
<td>PARAMETERS</td>
<td>0.0209</td>
<td></td>
<td>DIGIT</td>
<td>0.0149</td>
<td></td>
<td>AGENT</td>
<td>0.0136</td>
<td></td>
<td>REGRESSION</td>
<td>0.0155</td>
</tr>
<tr>
<td></td>
<td>ESTIMATE</td>
<td>0.0204</td>
<td></td>
<td>HAND</td>
<td>0.0126</td>
<td></td>
<td>DECISION</td>
<td>0.0118</td>
<td></td>
<td>MARGIN</td>
<td>0.0151</td>
</tr>
<tr>
<td></td>
<td>Tresp_V</td>
<td>0.0333</td>
<td></td>
<td>Simard_P</td>
<td>0.0694</td>
<td></td>
<td>Singh_S</td>
<td>0.1412</td>
<td></td>
<td>Smola_A</td>
<td>0.1033</td>
</tr>
<tr>
<td></td>
<td>Singer_Y</td>
<td>0.0281</td>
<td></td>
<td>Martin_G</td>
<td>0.0394</td>
<td></td>
<td>Barto_A</td>
<td>0.0471</td>
<td></td>
<td>Scholkopf_B</td>
<td>0.0730</td>
</tr>
<tr>
<td></td>
<td>Jebara_T</td>
<td>0.0207</td>
<td></td>
<td>LeCun_Y</td>
<td>0.0359</td>
<td></td>
<td>Sutton_R</td>
<td>0.0430</td>
<td></td>
<td>Burges_C</td>
<td>0.0489</td>
</tr>
<tr>
<td></td>
<td>Ghahramani_Z</td>
<td>0.0196</td>
<td></td>
<td>Denker_J</td>
<td>0.0278</td>
<td></td>
<td>Dayan_P</td>
<td>0.0324</td>
<td></td>
<td>Vapnik_V</td>
<td>0.0431</td>
</tr>
<tr>
<td></td>
<td>Ueda_N</td>
<td>0.0170</td>
<td></td>
<td>Henderson_D</td>
<td>0.0256</td>
<td></td>
<td>Parr_R</td>
<td>0.0314</td>
<td></td>
<td>Chapelle_O</td>
<td>0.0210</td>
</tr>
<tr>
<td></td>
<td>Jordan_M</td>
<td>0.0150</td>
<td></td>
<td>Revow_M</td>
<td>0.0229</td>
<td></td>
<td>Dietterich_T</td>
<td>0.0231</td>
<td></td>
<td>Cristianini_N</td>
<td>0.0185</td>
</tr>
<tr>
<td></td>
<td>Roweis_S</td>
<td>0.0123</td>
<td></td>
<td>Platt_J</td>
<td>0.0226</td>
<td></td>
<td>Tsitsiklis_J</td>
<td>0.0194</td>
<td></td>
<td>Ratsch_G</td>
<td>0.0172</td>
</tr>
<tr>
<td></td>
<td>Schuster_M</td>
<td>0.0104</td>
<td></td>
<td>Keeler_J</td>
<td>0.0192</td>
<td></td>
<td>Randlov_J</td>
<td>0.0167</td>
<td></td>
<td>Laskov_P</td>
<td>0.0169</td>
</tr>
<tr>
<td></td>
<td>Xu_L</td>
<td>0.0098</td>
<td></td>
<td>Rashid_M</td>
<td>0.0182</td>
<td></td>
<td>Bradtke_S</td>
<td>0.0161</td>
<td></td>
<td>Tipping_M</td>
<td>0.0153</td>
</tr>
<tr>
<td></td>
<td>Saul_L</td>
<td>0.0094</td>
<td></td>
<td>Sackinger_E</td>
<td>0.0132</td>
<td></td>
<td>Schwartz_A</td>
<td>0.0142</td>
<td></td>
<td>Sollich_P</td>
<td>0.0141</td>
</tr>
</tbody>
</table>
Extension: entity-topic model for entity linking

- Extended from the Author-topic model
- Entity linking task
Given the topic knowledge $\phi$, the entity name knowledge $\psi$ and the entity context knowledge $\xi$:

1. For each doc $d$ in $\textbf{D}$, sample its topic distribution $\theta_d \sim \text{Dir}(\alpha)$;
2. For each of the $M_d$ mentions $m_i$ in doc $d$:
   a) Sample a topic assignment $z_i \sim \text{Mult}(\theta_d)$;
   b) Sample an entity assignment $e_i \sim \text{Mult}(\phi_{z_i})$;
   c) Sample a mention $m_i \sim \text{Mult}(\psi_{e_i})$;
3. For each of the $N_d$ words $w_i$ in doc $d$:
   a) Sample a target entity it describes from $d$’s referent entities $a_i \sim \text{Unif}(e_{m_1}, e_{m_2}, \cdots, e_{m_d})$;
   b) Sample a describing word using $a_i$’s context word distribution $w_i \sim \text{Mult}(\xi_{a_i})$. 

Entity-topic model
Take home notes

- Overview of topic models
- LDA for ad-hoc retrieval
- LDA for query expansion
- Topic models of social text streams
Future work of topic models

- Non-parametric topic models
- Large-scale topic modeling
- Parallel processing to enhance the efficiency
- Active learning to enhance performance
- Continuous time periods for time-aware topic modeling
- Topic models and its applications
- Zhaochun Ren
- ISLA, University of Amsterdam
- z.ren@uva.nl