Latent Dirichlet allocation in web spam filtering

István Bíró\textsuperscript{1} Jácint Szabó\textsuperscript{1} András A. Benczúr\textsuperscript{1}

\textsuperscript{1}Data Mining and Web Search Group
Computer and Automation Institute
Hungarian Academy of Sciences

AIRWeb Workshop, April 22, 2008, Beijing, China.
Latent Dirichlet allocation

- Blei, Ng, Jordan, 2003
- fully generative statistical natural language model
- extension of latent semantic indexing (LSI)
- has better perplexity than LSI
- a lot of extensions and variations of LDA were developed and successfully applied
Latent Dirichlet allocation

Model

- topic: distribution over the words
- document: distribution over the topics
- for every word-position of the corpus, draw a topic for that document, and then draw a word for that topic

Inference by Gibbs sampling

- Markov chain Monte Carlo method
- running time $O((\#\text{topics}) \cdot (\#\text{word pos's}) \cdot (\#\text{iterations}))$
Latent Dirichlet allocation

Model

- topic: distribution over the words
- document: distribution over the topics
- for every word-position of the corpus, draw a topic for that document, and then draw a word for that topic

Inference by Gibbs sampling

- Markov chain Monte Carlo method
- running time $O((\#\text{topics}) \cdot (\#\text{word pos’s}) \cdot (\#\text{iterations}))$
Latent Dirichlet allocation

In practice

- given a collection of documents
- keep only semantic words, delete stopwords, stem
- create vocabulary
- choose an appropriate topic-number (about 100)
- make model inference to create the model
- for a topic, the word distribution gives a semantic theme
- for a document, the topic distribution describes to which themes it belongs
- for a new document, make unseen inference to get its topic distribution
In practice
- given a collection of documents
- keep only semantic words, delete stopwords, stem
- create vocabulary
- choose an appropriate topic-number (about 100)
- make model inference to create the model
- for a topic, the word distribution gives a semantic theme
- for a document, the topic distribution describes to which themes it belongs
- for a new document, make unseen inference to get its topic distribution
In practice

- given a collection of documents
- keep only semantic words, delete stopwords, stem
- create vocabulary
- choose an appropriate topic-number (about 100)
- make model inference to create the model
- for a topic, the word distribution gives a semantic theme
- for a document, the topic distribution describes to which themes it belongs
- for a new document, make unseen inference to get its topic distribution
Multi-corpus LDA

- two corpora: labeled spam and nonspam sites
- build two separate LDA models on them, with $k_s$ and $k_n$ topics
- aggregate these models to get $k_s + k_n$ topics
- make unseen inference for every unlabeled site, and get its topic-distribution: $\sum_{1 \leq i \leq k_s} p_i^s + \sum_{1 \leq i \leq k_n} p_i^n = 1$
- the total probability of spam topics, $\sum_{1 \leq i \leq k_s} p_i^s$, gives a spamness feature

Similar to the compression based spam filter of Cormack, applied with success at Web Spam Challenge 2007.
Multi-corpus LDA on UK2007-WEBSpAM, apparently primarily content spammed

- $\sim 115000$ sites
- $\sim 200$ labeled as spam
- $\sim 3800$ labeled as nonspam

- document: concatenation of all pages of a site
- topic numbers: $k_s = 10$ and $k_n = 50$
Tests

Most frequent words in some topics

<table>
<thead>
<tr>
<th>Spam topic 7</th>
<th>Nonspam topic 4</th>
<th>Nonspam topic 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>loan (0.080)</td>
<td>club (0.035)</td>
<td>music (0.022)</td>
</tr>
<tr>
<td>uk (0.042)</td>
<td>team (0.012)</td>
<td>band (0.012)</td>
</tr>
<tr>
<td>unsecured (0.026)</td>
<td>league (0.009)</td>
<td>film (0.011)</td>
</tr>
<tr>
<td>credit (0.024)</td>
<td>win (0.009)</td>
<td>festival (0.009)</td>
</tr>
<tr>
<td>home (0.022)</td>
<td>home (0.009)</td>
<td>dance (0.008)</td>
</tr>
</tbody>
</table>
Tests

- public: Web Spam Challenge 2008 public features
- text: pivoted tf.idf (Singhal et al.)
- graph: site and page level stacked graphical (see Dávid Siklósi’s Challenge talk)

<table>
<thead>
<tr>
<th>feature set</th>
<th>F</th>
<th>ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>text (SVM)</td>
<td>0.554</td>
<td>0.864</td>
</tr>
<tr>
<td>public &amp; text &amp; graph (log)</td>
<td>0.601</td>
<td>0.954</td>
</tr>
<tr>
<td>public &amp; text &amp; graph &amp; lda (log)</td>
<td>0.667</td>
<td>0.969</td>
</tr>
</tbody>
</table>