Is Your Anchor Going Up or Down? Fast and Accurate Supervised Topic Models

Thang Nguyen, Jordan Boyd-Graber, Jeff Lund, Kevin Seppi, and Eric Ringger
daihang@umiacs.umd.edu, Jordan.Boyd.Graber@colorado.edu, {jefflund, kseppi}@byu.edu, ringger@cs.byu.edu
University of Maryland, College Park / University of Colorado, Boulder / Brigham Young University

Motivation
• Supervised topic models leverage latent document-level themes to capture nuanced sentiment, create sentiment-specific topics and improve sentiment prediction.
• Examples include Supervised LDA (Blei et al., 2007), Labelled LDA (Ramage et al., 2009), Med LDA (Zhu et al., 2009), etc.
• The downside for Supervised LDA is that it is slow, which this work addresses.

Contribution
• We create a supervised version of the anchor word algorithm (ANCHOR) (Arora et al., 2013).
• This supervised anchor word algorithm (SUP ANCHOR) is very fast because it inherits fast inference from the anchor algorithm.
• Experiments on three sentiment datasets show that SUP ANCHOR is comparable to Supervised LDA (SLDA) in terms of prediction accuracy.
• Anchor words learned by SUP ANCHOR provide great insight.

Supervised Anchor Word (Cont.)
• Anchor words form a convex hull that encloses all other words in the vocabulary.
• Adding sentiment related dimensions moves words UP or DOWN; forming sentiment-specific anchor words.

Experiments
• Goal: Evaluate the new topics generated by the proposed model in a prediction task. We focus on binary classification in sentiment analysis datasets.
• Sentiment datasets.
  
<table>
<thead>
<tr>
<th>Corpus</th>
<th>Train Docs</th>
<th>Test Docs</th>
<th>Tokens</th>
<th>Types</th>
<th>Positive Sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMAZON</td>
<td>13,300</td>
<td>3,314</td>
<td>1,031,659</td>
<td>2,662</td>
<td>52.2%</td>
</tr>
<tr>
<td>TRIPADVISOR</td>
<td>115,394</td>
<td>28,828</td>
<td>12,752,444</td>
<td>4,867</td>
<td>41.5%</td>
</tr>
<tr>
<td>YELP</td>
<td>13,955</td>
<td>3,482</td>
<td>1,142,555</td>
<td>2,585</td>
<td>27.7%</td>
</tr>
</tbody>
</table>

Table: Statistics for the datasets employed in the experiments.

Runtime Analysis

<table>
<thead>
<tr>
<th>Method</th>
<th>Runtime in seconds</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>891s</td>
</tr>
<tr>
<td>SLDA</td>
<td>4,762s</td>
</tr>
<tr>
<td>SUP ANCHOR</td>
<td>118s</td>
</tr>
</tbody>
</table>

Future Directions
• Explore other richer word representations (e.g. word2vec, more smoothing on added dimensions).
• Work on inferring document-topic distributions directly.
• Incorporate into Interactive Topic Modeling and Active Learning.