LDA Based Topic Modeling

- Background
Information Retrieval (IR) Methodology

- **tf-idf scheme** (Salton and McGill, 1983)
  - Documents are reduced to a vector of real numbers, each number represents ratios of counts.
  - Counts terms in documents, reduces arbitrary length of documents to fixed-length lists of numbers.
  - Fails to reveal statistical structure inside/between documents.

- **aspect model** (Hofmann, 1999)
  - A generative probabilistic latent semantic indexing (pLSI).
  - Models each word in a document as a sample from mixture of ‘topics’.
  - Each document is represented as a probability distribution on a fixed number of ‘topics’.
Fundamental Probabilistic Assumption of pLSI

- ‘bag-of-words’
  - Order of words in a document can be neglected
  - Order of documents in a corpus can be neglected
  - Assumption of ‘exchangeable random variables’

- Not providing model at the level of documents
  - A document is just a list of numbers (mixing proportions for topics)
  - Number of parameters grows linearly, leads to over-fitting
  - Cannot assign probability to a document outside of training set
Any collection of exchangeable random variables has a representation as a mixture distribution – in general an infinite mixture

- ‘exchangeability’ means *conditionally* independent and identically distributed, where conditioning is with respect to an underlying *latent parameter* of a probability distribution
- Need mixture models to capture exchangeability for both words and documents
Latent Dirichlet Allocation (LDA)

- A generative probabilistic model for collections of discrete data such as text corpora
- A three-level hierarchical Bayesian model
- Each document is a random mixture over latent topics
- Each topic is a distribution over words
LDA Model

LDA assumes the following generative process for each document \( w \) in a corpus \( \mathcal{D} \):

1. Choose \( N \sim \text{Poisson}(\xi) \).
2. Choose \( \theta \sim \text{Dir}(\alpha) \).
3. For each of the \( N \) words \( w_n \):
   
   (a) Choose a topic \( z_n \sim \text{Multinomial}(\theta) \).
   
   (b) Choose a word \( w_n \) from \( p(w_n \mid z_n, \beta) \), a multinomial probability conditioned on the topic \( z_n \).

Before proceeding, a couple of assumptions are made in this basic model, some of which we will return to in subsequent sections. First, the dimensionality \( k \) of the Dirichlet distribution (and thus the dimensionality of the topic variable \( z \)) is assumed known and fixed. Second, the word probabilities are parameterized by a \( k \times V \) matrix \( \beta \) where \( \beta_{ij} = p(w^j = 1 \mid z^i = 1) \), which for now we treat as a fixed quantity that is to be estimated. Finally, the Poisson assumption is not critical to anything that follows and the parameter \( \alpha \) is a \( k \)-vector with components \( \alpha_i > 0 \),
Comparison of pLSI and LDA Model

\[ \alpha \text{ and } \beta \text{ are corpus-level parameters, sampled once in corpus generating process} \]
\[ \text{Variables } \theta \text{ are document-level variables, sampled once per document} \]
\[ \text{Variables } z \text{ and } w \text{ are word-level variables, sampled once for each word in each document} \]
Some Related Modification of LDA

- **DF-LDA** (Andrzejewski, 2009)
  - User set must-link or cannot-link constraints, to reduce non-trivial grounding numbers
  - Only for aspect extraction
- **MaxEnt-LDA** (Zhao, 2010)
  - Jointly modeling aspect and sentiment in review text analysis
  - Maximum-Entropy to train a switch variable to separate aspect and sentiment words
  - Unsupervised
- **Seeded Aspect and Sentiment model** (SAS)
  - Semi-supervised by seeds provided by user, although in a different
Problems in Max-Ent LDA

- Without seeds (unsupervised), many discovered aspects are not meaningful to users
- Manually label data in training
SAS and ME-SAS

- Assumption: one review sentence usually talks about one aspect
- $D$: all documents,
- $S_d$: all sentences in document $d_{1...D}$
- $N_{d,s}$: all words in $s \in S_d$, also use $\text{Sent}^d_s$ to denote $s$ in $d$
- Indicator (switch) variable for distinguish between aspect and sentiment terms, $r_{d,s,j} \in \{\hat{a},\hat{o}\}$ for $w_{d,s,j}$ which is the j-th term in $\text{Sent}^d_s$
- $\Psi_{d,s}$: distribution of aspects and sentiments in $\text{Sent}^d_s$; it is also the success probability of emitting an
SAS and ME-SAS

- $V$: all unique non-seed terms in corpus
- $C$: all seed sets, each seed set, $Q_l \ (l=1...C)$, is a group of semantically related terms
- $T$: all aspects, $\varphi_t$ is the t-th aspect, $\varphi_0$ is the model for $\varphi_t$
- $\Omega$: distribution of seeds, $\Omega_{t,l}$ is the distribution of seeds for the t-th aspect, in seed set $Q_l$
SAS Model

1. For each aspect $t \in \{1, \ldots, T\}$:
   i. Draw $\varphi^o_t \sim \text{Dir} (\beta^o)$
   ii. Draw a distribution over terms and seed sets $\varphi^A_t \sim \text{Dir} (\beta^A)$
      a) For each seed set $l \in \{Q_1, \ldots, Q_C\}$
         Draw a distribution over seeds $\Omega_{t,l} \sim \text{Dir} (\gamma)$

2. For each (review) document $d \in \{1, \ldots, D\}$:
   i. Draw $\theta_d \sim \text{Dir} (\alpha)$
   ii. For each sentence $s \in \{1, \ldots, S_d\}$:
      a) Draw $z_{d,s} \sim \text{Mult} (\theta_d)$
      b) Draw $\psi_{d,s} \sim \text{Beta} (\delta)$
      c) For each term $w_{d,s,j}$ where $j \in \{1, \ldots, N_{d,s}\}$:
         I. Draw $r_{d,s,j} \sim \text{Bernoulli} (\psi_{d,s})$, $r_{d,s,j} \in \{\hat{a}, \hat{o}\}$
         II. if $r_{d,s,j} = \hat{a}$ // $w_{d,s,j}$ is a sentiment
             Emit $w_{d,s,j} \sim \text{Mult} (\varphi^A_{t,d})$
             else // $r_{d,s,j} = \hat{o}$, $w_{d,s,j}$ is an aspect
                A. Draw $u_{d,s,j} \sim \text{Mult} (\varphi^A_{t,d})$
                B. if $u_{d,s,j} \in V$ // non-seed term
                   Emit $w_{d,s,j} = u_{d,s,j}$
                   else // $u_{d,s,j}$ is some seed set index say $l_{d,s,j}$
                   Emit $w_{d,s,j} \sim \Omega_{z_{d,s,l_{d,s,j}}}$
Move $\Psi_{d,s}$ to term plate and draw $\Psi_{d,s}$ from Max-Ent($x_{d,s,j}; \lambda$) model
Experiment

- Corpus consisted of 101,234 reviews and 692,783 sentences
- Compare ME-LDA, DF-LDA with ME-SAS and SAS
- The posterior inference was drawn after 5000 Gibbs iterations, with an initial burn-in 1000 iterations
- Max-Ent parameter $\lambda$ was learned from 1000 terms which are automatically generated, and at least appeared 20 times
Experiment

- 9 major aspects (T=9): Dining, Staff, Maintenance, Check in, Cleanliness, Comfort, Amenities, Location, Value for Money (VFM)
Table 1: Top ranked aspect and sentiment words in three aspects (please see the explanation in Section 4.1).
## Experiment

<table>
<thead>
<tr>
<th>No. of Seeds</th>
<th>DF-LDA</th>
<th>DF-LDA-Relaxed</th>
<th>SAS</th>
<th>ME-SAS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P@10</td>
<td>P@20</td>
<td>P@30</td>
<td>P@10</td>
</tr>
<tr>
<td>2</td>
<td>0.51</td>
<td>0.53</td>
<td>0.49</td>
<td>0.67</td>
</tr>
<tr>
<td>3</td>
<td>0.53</td>
<td>0.54</td>
<td>0.50</td>
<td>0.71</td>
</tr>
<tr>
<td>4</td>
<td>0.57</td>
<td>0.56</td>
<td>0.53</td>
<td>0.73</td>
</tr>
<tr>
<td>5</td>
<td>0.59</td>
<td>0.57</td>
<td>0.54</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Table 2: Average $p@n$ of the seeded aspects with the no. of seeds.

<table>
<thead>
<tr>
<th>Aspect</th>
<th>ME-LDA</th>
<th>DF-LDA</th>
<th>DF-LDA-Relaxed</th>
<th>SAS</th>
<th>ME-SAS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P@10</td>
<td>P@20</td>
<td>P@30</td>
<td>P@10</td>
<td>P@20</td>
</tr>
<tr>
<td>Dining</td>
<td>0.70</td>
<td>0.65</td>
<td>0.67</td>
<td>0.50</td>
<td>0.60</td>
</tr>
<tr>
<td>Staff</td>
<td>0.60</td>
<td>0.70</td>
<td>0.67</td>
<td>0.40</td>
<td>0.65</td>
</tr>
<tr>
<td>Maintenance</td>
<td>0.80</td>
<td>0.75</td>
<td>0.73</td>
<td>0.40</td>
<td>0.55</td>
</tr>
<tr>
<td>Check In</td>
<td>0.70</td>
<td>0.70</td>
<td>0.67</td>
<td>0.50</td>
<td>0.65</td>
</tr>
<tr>
<td>Cleanliness</td>
<td>0.70</td>
<td>0.75</td>
<td>0.67</td>
<td>0.70</td>
<td>0.70</td>
</tr>
<tr>
<td>Comfort</td>
<td>0.60</td>
<td>0.70</td>
<td>0.63</td>
<td>0.60</td>
<td>0.65</td>
</tr>
<tr>
<td>Amenities</td>
<td>0.80</td>
<td>0.80</td>
<td>0.67</td>
<td>0.70</td>
<td>0.65</td>
</tr>
<tr>
<td>Location</td>
<td>0.60</td>
<td>0.70</td>
<td>0.63</td>
<td>0.50</td>
<td>0.60</td>
</tr>
<tr>
<td>VFM</td>
<td>0.50</td>
<td>0.55</td>
<td>0.50</td>
<td>0.40</td>
<td>0.50</td>
</tr>
<tr>
<td>Avg.</td>
<td>0.67</td>
<td>0.70</td>
<td>0.65</td>
<td>0.52</td>
<td>0.62</td>
</tr>
</tbody>
</table>

Table 3: Effect of performance on seeded and non-seeded aspects (5 seeds were used for the 6 seeded aspects).