Extensions and Adaptations of LDA

Lisa Posch

Topic Model Tutorial
Hannover, 2016
Why do we want to extend LDA?
We want to extend LDA so that we can

• include different characteristics
• explore different aspects

of our dataset.
Outline

• Quick Recap of LDA

• Extensions and Adaptations
Outline

• Quick Recap of LDA

• Extensions and Adaptations:
  – Labeled LDA
  – Polylingual Topic Model
  – Author-Topic Model
  – Topics over Time
  – Citation Influence Model
Goal of this Session

- You know that there are different topic models that are based on LDA.
- You have seen some specific adaptations of LDA
- and you know what they are used for.
LDA (recap)
LDA

Collection of Documents
LDA

Collection of Documents
LDA

Collection of Documents

Topic 1
websci
conference
germany
hannover
social
computer

Topic 2
food
restaurant
eat
steak
cafe

Topic 3
sociology
social
behavior
relationships
quantitative
LDA

Collection of Documents

LDA

Topic 1
websci conference
germany hannover
social computer

Topic 2
food restaurant
pizza eat
steak cafe

Topic 3
sociology social
society behavior
relationships quantitative

<table>
<thead>
<tr>
<th></th>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc 1</td>
<td>40%</td>
<td>40%</td>
<td>20%</td>
</tr>
<tr>
<td>Doc 2</td>
<td>50%</td>
<td>10%</td>
<td>40%</td>
</tr>
<tr>
<td>Doc 3</td>
<td>5%</td>
<td>45%</td>
<td>50%</td>
</tr>
</tbody>
</table>
LDA – Generative Storyline

1. For each document, draw a distribution over topics
2. For each topic, draw a distribution over the vocabulary
3. For each word in each document:
   – Draw a topic
   – Draw a word from this topic
LDA – Plate Notation

\[ \theta_d \rightarrow z_{di} \rightarrow w_{di} \rightarrow \phi_k \]

\[ \forall i \in [1, N_d] \]
\[ \forall d \in [1, D] \]

\[ \forall k \in [1, K] \]

**document-topic distribution**

**topic**

**word**

**topic-term distribution**
Labeled LDA
Labeled LDA

L-LDA is supervised variant of LDA which takes labeled documents as input and creates a topic for each label.

Labeled LDA

L-LDA is supervised variant of LDA which takes labeled documents as input and creates a topic for each label.

Dataset: text documents with multiple labels
Labeled LDA

Collection of labeled Documents
Collection of labeled Documents

Labeled LDA
Labeled LDA

Collection of labeled Documents
Labeled LDA

Collection of labeled Documents

<table>
<thead>
<tr>
<th></th>
<th>WebSci</th>
<th>Food</th>
<th>Sociology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc 1</td>
<td>70%</td>
<td>30%</td>
<td>0%</td>
</tr>
<tr>
<td>Doc 2</td>
<td>45%</td>
<td>0%</td>
<td>55%</td>
</tr>
<tr>
<td>Doc 3</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
</tr>
</tbody>
</table>

WebSci:
- websci
- conference
- germany
- hannover
- social
- computer

Food:
- food
- restaurant
- pizza
- eat
- steak
- cafe

Sociology:
- sociology
- social
- society
- behavior
- relationships
- quantitative
L-LDA – Generative Storyline

1. For each document, draw a distribution over topics, restricted to the document’s labels
2. For each topic, draw a distribution over the vocabulary
3. For each word in each document:
   - Draw a topic, from the permitted topics
   - Draw a word from this topic
L-LDA – Plate Notation

\[ \alpha \overset{\theta_d}{\rightarrow} \overset{z_{di}}{\rightarrow} \overset{w_{di}}{\rightarrow} \overset{\phi_k}{\rightarrow} \beta \]

\[ \forall d \in [1,D] \]

\[ \forall i \in [1,N_d] \]

\[ \forall k \in [1,K] \]
L-LDA – Plate Notation

\[ \forall d \in [1,D] \]

\[ \forall i \in [1,N_d] \]

\[ \forall k \in [1,K] \]
L-LDA – Examples

• Publications, labeled with a classification system
  – Create a topic for each class in the classification system

• Tagged Blog entries
  – Create a topic for each tag
Polylingual Topic Model
Polylingual Topic Model

PLTM is a topic model for corpora where the documents are available in several languages.

The sets of documents should be loosely equivalent to each other.

Polylingual Topic Model

PLTM uses a separate vocabulary for each language, and each topic has a word distribution for each language.
Polylingual Topic Model

Collection of multilingual Documents
Polylingual Topic Model

Collection of multilingual Documents
Polylingual Topic Model

Collection of multilingual Documents

- **Topic 1**
  - websci
  - conference
  - germany
  - hannover
  - social
  - computer

- **Topic 2**
  - food
  - restaurant
  - pizza
  - eat
  - steak
  - cafe

- **Topic 3**
  - sociology
  - social
  - society
  - behavior
  - relationships
  - quantitative
Polylingual Topic Model

Collection of multilingual Documents
PLTM – Generative Storyline

1. For each document, draw a distribution over topics

2. For each topic, in each language, draw a distribution over the vocabulary of this language

3. For each word in each language in each document:
   – Draw a topic
   – Draw a word from this language-specific topic
PLTM – Plate Notation

\[ \alpha \rightarrow \theta_d \rightarrow z^l_{di} \rightarrow w^l_{di} \rightarrow \phi^l_k \rightarrow \beta^l \]

\[ \forall d \in [1, D] \]
\[ \forall i \in [1, N^l_d] \]
\[ \forall k \in [1, K] \]
\[ \forall l \in [1, L] \]
PLTM – Plate Notation

- $\alpha \rightarrow \theta_d 

- \forall d \in [1,D]

- $\forall i \in [1,N_d^l]$  

- $z_{di}^l \rightarrow w_{di}^l 

- \forall k \in [1,K]$  

- $\phi_k^l 

- \forall l \in [1,L]$  

- $\beta^l 

- \text{multiple languages}$
PLTM - Examples

• Wikipedia articles in several languages
  – Create topics for each language

• Documents that are annotated with a controlled vocabulary
  – Create topics for both the natural language and the controlled vocabulary
Author-Topic Model
Author-Topic Model

- The Author-Topic model extends LDA to include authorship information.

- Each author has a distribution over topics.

Author-Topic Model

• For each word in a document
  – choose an author,
  – then choose a topic from that author‘s topic distribution and
  – generate a word from that topic.

Bleier, Arnim, and Andreas Strotmann. "Towards an Author-Topic-Term-Model Visualization of 100 Years of German Sociological Society Proceedings."
Topics over Time
Topics over Time

• Topics generate both words and observed timestamps.

• Jointly models word co-occurrences and localization in time.
Topics over Time

<table>
<thead>
<tr>
<th>Topic</th>
<th>Value</th>
<th>Topic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>government</td>
<td>0.02928</td>
<td>world</td>
<td>0.01875</td>
</tr>
<tr>
<td>united</td>
<td>0.02132</td>
<td>states</td>
<td>0.01717</td>
</tr>
<tr>
<td>states</td>
<td>0.02067</td>
<td>security</td>
<td>0.01710</td>
</tr>
<tr>
<td>islands</td>
<td>0.01167</td>
<td>soviet</td>
<td>0.01664</td>
</tr>
<tr>
<td>canal</td>
<td>0.01014</td>
<td>united</td>
<td>0.01491</td>
</tr>
<tr>
<td>american</td>
<td>0.00872</td>
<td>nuclear</td>
<td>0.01454</td>
</tr>
<tr>
<td>cuba</td>
<td>0.00834</td>
<td>peace</td>
<td>0.01408</td>
</tr>
</tbody>
</table>

Citation Influence Model
Citation Influence Model

- Estimates the weight of edges in a citation graph, i.e. the strength of influence one publication has on another.

- Incorporates the aspects of topical innovation and topical inheritance via citations.

Citation Influence Model

original LDA Paper
Summary

- Labeled LDA: for labeled documents
- Polylingual Topic Model: multilingual documents
- Author-Topic Model: authors’ interests
- Topics over Time: topics’ localization in time
- Citation Influence Model: strength of influence
Which topic model you want to use depends on your data and on which questions you want to answer.
Questions,

but first….
How could topic models be useful for your research?

Which aspects of datasets would you want to explore with topic models?

... discuss with your neighbor!

What questions do you have?
Which things are still unclear to you?

What would you like to know more about?

e etc. etc. etc
Thank you!