Improving Quality of LDA Models

Henri Trenquier

Supervisor:
Dr. Carlos Ortiz Martinez
MSc Security and Network Engineering

University of Amsterdam

July 5, 2018
Context
Forensics

- Accelerate forensic investigations
- Large document collections

_A Forensic Analysis Solution of the Email Network Based on Email Contents_

- L Xie, Y Liu, G Chen (2015)
- Email network analysis
Latent Dirichlet allocation

- David Blei, Andrew Ng, and Michael I. Jordan (2003)
- Cited over 23K times
- Machine learning

Statistical model

- Bayesian
- generative & probabilistic
- for a collection of discrete data
- Topic discovery
To: Jack@company.com

Cc:

Subject: Random question

From: Henri Trenquier – henri.trenquier@os3.nl

Hi Jack,

What is a human computer interface?

Best,

Henri

- Preprocessing
- Bag of word: (‘human’, ‘interface’, ‘computer’)
Example

Corpus

1. ‘human’, ‘interface’, ‘computer’
3. ‘eps’, ‘user’, ‘interface’, ‘system’
5. ‘user’, ‘response’, ‘time’
6. ‘trees’
7. ‘graph’, ‘trees’
8. ‘graph’, ‘minors’, ‘trees’
9. ‘graph’, ‘minors’, ‘survey’
Example
Corpus

1. ‘human’, ‘interface’, ‘computer’
3. ‘eps’, ‘user’, ‘interface’, ‘system’
5. ‘user’, ‘response’, ‘time’
6. ‘trees’
7. ‘graph’, ‘trees’
8. ‘graph’, ‘minors’, ‘trees’
9. ‘graph’, ‘minors’, ‘survey’

Expected topics
1. Human machine interface
2. Graph theory
Topic modeling LDA

Example

Topic representation

- system
- user
- eps
- interface
- human
- trees
- response
- time
- computer
- minors
- graph
- survey

(topic_1) (topic_2)
**Example**

**Topic modeling LDA**

---

**Document representation**

- ['graph', 'minors', 'survey']
- ['graph', 'minors', 'trees']
- ['graph', 'trees']
- ['trees']
- ['user', 'response', 'time']
- ['system', 'human', 'system', 'eps']
- ['eps', 'user', 'interface', 'system']
- ['survey', 'user', 'computer', 'system', 'response', 'time']
- ['human', 'interface', 'computer']

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

- **topic_1**
- **topic_2**
### Example

**Topic modeling LDA**

Expected topics:

1. Human machine interface
2. Graph theory

<table>
<thead>
<tr>
<th>Model</th>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good_Model</td>
<td>('system', 'user', 'eps', 'human', 'interface')</td>
</tr>
<tr>
<td></td>
<td>('graph', 'trees', 'minors', 'survey', 'time')</td>
</tr>
<tr>
<td>Bad_Model</td>
<td>('computer', 'system', 'user', 'trees', 'graph')</td>
</tr>
<tr>
<td></td>
<td>('system', 'graph', 'trees', 'user', 'eps')</td>
</tr>
</tbody>
</table>

**Table:** Good and Bad models
Expected topics:

1. Human machine interface
2. Graph theory

<table>
<thead>
<tr>
<th>Model</th>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bad Model</td>
<td>(‘computer’, ‘system’, ‘user’, ‘trees’, ‘graph’)</td>
</tr>
<tr>
<td></td>
<td>(‘system’, ‘graph’, ‘trees’, ‘user’, ‘eps’)</td>
</tr>
</tbody>
</table>

Table: Good and Bad models

- More words over all topics
- **More similar** words **within** a topic
- **Less similar** words **across** topics
Context

Enron

- Accounting fraud
- ~500K e-mails database
- Topic modeling dataset
- quickly target incriminating e-mails
Research Question

How to improve the quality of LDA models?

- What is the optimal number of topics for a LDA model
- How does the number of iterations influence the quality of models?
- Can we improve semantic quality evaluation?

Scope: Enron e-mail dataset
State of the art
Coherence

- Evaluation metric for topic modeling

*Optimizing Semantic Coherence in Topic Models*

- D Mimno et al. (2011)
- 542 citations

\[
C(t; V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})} \tag{1}
\]
State of the art

Coherence

- Evaluation metric for topic modeling

*Optimizing Semantic Coherence in Topic Models*

- D Mimno et al. (2011)
- 542 citations

\[ C(t; V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)})}{D(v_l^{(t)})} + 1 \]  

Measure evaluated by a survey:

- "good", "intermediate" or "bad"
- no literal definition of coherence
- lack of "inter-topic" evaluation
- \( C_v \) and \( U_{MASS} \)
State of the art

Coherence

A Practical Algorithm for Topic Modeling with Provable Guarantees

- S Arora et al. (2013)
- 229 citations
- introduces "inter-topic similarity"
Evaluation metric

Topic Coherence

\[ C_{\text{word2vec}} \] coherence measure

- Semantic space
- word2vec model trained on Google News
Evaluation metric

Topic Coherence

$C_{\text{word2vec}}$ coherence measure

- Semantic space
- word2vec model trained on Google News

1. intra_topic_similarity
2. inter_topic_similarity

$$C_{\text{word2vec}} = \frac{\text{avg}(\text{intra_topic_similarity})}{\text{avg}(\text{inter_topic_similarity})}$$ (2)
### Evaluation metric

**Topic Coherence**

$c_{word2vec}$ coherence measure

- Semantic space
- word2vec model trained on Google News

1. **intra_topic_similarity**
2. **inter_topic_similarity**

$$c_{word2vec} = \frac{\text{avg}(\text{intra_topic_similarity})}{\text{avg}(\text{inter_topic_similarity})}$$

<table>
<thead>
<tr>
<th>Model</th>
<th>Topics</th>
<th>$c_{word2vec}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good_Model</td>
<td>(<code>system</code>, <code>user</code>, <code>eps</code>, <code>human</code>, <code>interface</code>) (<code>graph</code>, <code>trees</code>, <code>minors</code>, <code>survey</code>, <code>time</code>)</td>
<td>0.887</td>
</tr>
<tr>
<td>Bad_Model</td>
<td>(<code>computer</code>, <code>system</code>, <code>user</code>, <code>trees</code>, <code>graph</code>) (<code>system</code>, <code>graph</code>, <code>trees</code>, <code>user</code>, <code>eps</code>)</td>
<td>0.604</td>
</tr>
</tbody>
</table>
Experiment
Pipeline

Figure: Similarity measures

- Modeling: I, K
- Coherence analysis: $C_v$, $u_{mass}$, $C_{word2vec}$
Results

CV

Figure: Influence of the number of topics on the $CV$ coherence
Figure: Influence of the number of topics on the $U_{MASS}$ coherence
Figure: Influence of the number of topics on the $C_{word2vec}$ coherence
Results

Low & High number of iterations

Figure: Influence of the number of iterations on the $C_V$ coherence
Results
Low & High number of iterations

Figure: Influence of the number of iterations on the $U_{MASS}$ coherence
Results

Low & High number of iterations

Figure: Influence of the number of iterations on the $C_{\text{word2vec coherence}}$
E-mail information density

Preprocessing phase

word2vec semantic representation is not perfect

\[
sim(['th', 'de', 'er', 'ed', 'ng', 'enron', 'nd', 'es', 'al', 'ing']) = 1.28669572453
\]

\[C_{\text{word2vec}}\] coherence still too simplistic
How to improve the quality of LDA models?

- Impression of model coherence
- New semantic coherence
- Results do not reveal an optimum number of topic
- Number of iterations has no visible impact
Future Work

- Better preprocessing: stemming
- Refine $C_{\text{word2vec}}$ coherence
  - weight the words of a topic
  - word2vec training dataset
  - compare similar models
- Hierarchical topics
Thank you for your attention