Topic Model Tutorial
Part 1 – The Intuition

Christoph Carl Kling
ckling@uni-koblenz.de
Conference Dinner
Conference dinner

- I sit at a table with a probability proportional to the number of people already sitting there
Conference dinner

- I sit at a table with a probability proportional to the number of people already sitting there

- If everybody does the same and there are more and more people entering, the probabilities for choosing the tables converge
Conference dinner

- I sit at a table with a probability proportional to the number of people already sitting there

- If everybody does the same and there are more and more people entering, the probabilities for choosing the tables converge

- The scheme yields a sample of a *Dirichlet distribution*

  Parameters: initial number of participants at each table
Dirichlet Distribution

- The scheme yields a sample of a *Dirichlet distribution*

  Parameters: initial number of participants at each table

- “rich get richer”, preferential attachment

- Initial settings of < 1 participant at each table produce *sparse* distributions
In reality, I choose tables based on the number of people AND the topic they talk about!
In reality, I choose tables based on the number of people AND the topic they talk about!
Topics
Articles are labelled with tags (e.g. politics, economy, sports, ...)
Articles are labelled with tags (e.g. politics, economy, sports, ...)

Politics: election, party, vote, candidate, ...
Economy: dollar, crisis, financial, market, ...
Sports: soccer, basketball, match, score, ...
Articles are labelled with tags (e.g. politics, economy, sports, ...)

Politics: election, party, vote, candidate, ...
Economy: dollar, crisis, financial, market, ...
Sports: soccer, basketball, match, score, ...

Topics
Topic Modelling
Topic Modelling

Automatically extract topics from text documents!
Latent Semantic Analysis
Term-document matrix
Term-document matrix

term frequencies document 4

- high occurrence
- low occurrence
Term-document matrix

|   | 8  | 9  | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 |
|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| air|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| aliens|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| amnesty|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| arms|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| blood|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| booster|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| defense|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| drug|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| environmental|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| health|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| illegal|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| immigration|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| launch|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| nuclear|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| pollution|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| power|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| shuttle|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| space|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| study|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| treaty|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |

how often does document 4 contain the word “blood”? 

- high occurrence
- low occurrence
Latent Semantic Analysis (LSA)

- Topic model based on “matrix decomposition”
Latent Semantic Analysis (LSA)

- Topic model based on “matrix decomposition”

- Topics are described by “loadings” over the terms
The Test Dataset
Test dataset

document 0: probabilistic topic model
document 1: probabilistic topic model
document 2: probabilistic topic model
document 3: probabilistic topic model
document 4: probabilistic topic model
document 5: probabilistic topic model
document 6: probabilistic topic model
document 7: famous fashion model
document 8: famous fashion model
document 9: famous fashion model
document 10: famous fashion model
document 11: famous fashion model
document 12: famous fashion model
document 13: famous fashion model
document 14: famous fashion model
document 15: famous fashion model
document 16: famous fashion model
document 17: famous fashion model
document 18: famous fashion model
document 19: famous fashion model
Test dataset

Topic 1: famous, fashion, model
Topic 2: model, probabilistic, topic

Expected topics
Test dataset

Term-document matrix
Test dataset

Term-document matrix
Test dataset

Term-document matrix
LSA

Topic 1

Topic 2
LSA

Topic 1

Topic 2
LSA – Weaknesses

- Topic loadings can be negative → hard to interpret!

- LSA has problems to cope with word ambiguities
Probabilistic LSA
Probabilistic LSA (PLSA)
- Based on *categorical* distributions
Probabilistic LSA (PLSA)

- Based on *categorical* distributions

- *Probabilistic model* that explains the creation of documents
Probabilistic LSA (PLSA)

The PLSA model for the creation of words in documents:

1) Documents have each a categorical distribution $t$ over the topics
Probabilistic LSA (PLSA)

The PLSA model for the creation of words in documents:

1) Documents have each a categorical distribution $t$ over the topics

2) Topics have each a categorical distribution $f$ over all words
Probabilistic LSA (PLSA)

The PLSA model for the creation of words in documents:

1) Documents have each a categorical distribution \( t \) over the topics

2) Topics have each a categorical distribution \( f \) over all words

3) Creation of a word in document \( i \):
   1) Draw a topic \( z \) from \( t_i \)
   2) Draw a word from \( f_z \)
Probabilistic LSA (PLSA)

Topic 1

Topic 2
Probabilistic LSA (PLSA)

Topic 1

Topic 2
Probabilistic LSA (PLSA)

Topic 1

Topic 2
Probabilistic LSA (PLSA)

Document 0 (probabilistic topic model)
Probabilistic LSA (PLSA)

Document 0 (probabilistic topic model)  Document 7 (famous fashion model)
PLSA – Strengths & Weaknesses

- Topics are probability distributions and easy to interpret!

- PLSA still has problems to cope with ambiguous words
Latent Dirichlet Allocation
Latent Dirichlet Allocation (LDA)

- A word in a document is likely to belong to the same topic as the other words of that document
Latent Dirichlet Allocation (LDA)

- A word in a document is likely to belong to the same topic as the other words of that document

*document 7: famous fashion model*
Latent Dirichlet Allocation (LDA)

- A word in a document is likely to belong to the same topic as the other words of that document

**document 7:** famous fashion model

Topic 1  Topic 1  ?
Latent Dirichlet Allocation (LDA)

- A word in a document is likely to belong to the same topic as the other words of that document

**document 7:** famous fashion model

\[ \text{Topic 1} \quad \text{Topic 1} \quad \rightarrow \quad \text{Topic 1} \]
Latent Dirichlet Allocation (LDA)

- A word in a document is likely to belong to the same topic as the other words of that document

**document 0:** probabilistic topic model
Latent Dirichlet Allocation (LDA)

- A word in a document is likely to belong to the same topic as the other words of that document

document 0: probabilistic topic model

Topic 2  Topic 2  ?
Latent Dirichlet Allocation (LDA)

- A word in a document is likely to belong to the same topic as the other words of that document.

*document 0: probabilistic topic model*

- Topic 2
- Topic 2

→ Topic 2
Latent Dirichlet Allocation (LDA)

- A word in a document is likely to belong to the same topic as the other words of that document

- We would need some preference for already assigned topics in a document
Latent Dirichlet Allocation (LDA)

- A word in a document is likely to belong to the same topic as the other words of that document

- We would need some preference for already assigned topics in a document

→ Dirichlet distribution!
Dirichlet distribution

Total customers: 15
Function: Dir(3.2, 4.3, 2.1)

Politics

Topic Models

Politics

Topic Models

Politics

Politics

Politics

Customers: 3
Probability: 0.2

Customers: 2
Probability: 0.133

Customers: 4
Probability: 0.266

Customers: 3
Probability: 0.2

Customers: 2
Probability: 0.133

Customers: 1
Probability: 0.066
Dirichlet distribution
Latent Dirichlet Allocation (LDA)

Topic 1

Topic 2
Probabilistic topic model (with sparse Dirichlet)

Document 0 (probabilistic topic model)  Document 7 (famous fashion model)
LDA – Strengths

- LDA can cope with ambiguous words!

- Most popular topic model
(Human) Evaluation
<table>
<thead>
<tr>
<th>PLSA</th>
<th>LDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1</td>
<td>family, registered, like, hard, members, ...</td>
</tr>
<tr>
<td>Topic 2</td>
<td>high, left, planned, organization, story, ...</td>
</tr>
<tr>
<td>Topic 3</td>
<td>normal, predicted, first, chief, health, ...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>first, network, time, won, week, third, ...</td>
</tr>
<tr>
<td>...</td>
<td>two, house, found, police, car, home, ...</td>
</tr>
<tr>
<td>...</td>
<td>cents, futures, cent, lower, higher, ...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Topic Model Game

- Tests the semantic coherence of topics

- Given the top-5 words of a topic and an intruder word from a different topic – find the intruder word!
Topic Model Game

Given the top-5 words of a topic and an intruder word from a different topic – find the intruder word!

air pollution power blood environmental nuclear
**Topic Model Game**

Given the top-5 words of a topic and an intruder word from a different topic – find the intruder word!

- air pollution
- power
- blood
- environmental
- nuclear
Topic Model Game

https://tinyurl.com/tmt16
Summary
Summary

- Dirichlet distribution (Polya urn scheme)

- Latent Semantic Analysis (LSA)

- Probabilistic Latent Semantic Analysis (PLSA)

- Latent Dirichlet Allocation (LDA)

- Human evaluation of topic models