Topic Modelling
(and Natural Language Processing)
workshop

PyCon UK 2019

@MarcoBonzanini

github.com/bonzanini/topic-modelling
Nice to meet you

- Data Science consultant: NLP, Machine Learning, Data Engineering
- Corporate training: Python + Data Science
- PyData London chairperson

github.com/bonzanini/topic-modelling
This tutorial

• Introduction to Topic Modelling

• Depending on time/interest:
  Happy to discuss broader applications of NLP

• The audience (tell me about you):
  - new-ish to NLP?
  - new-ish to Python tools for NLP?

github.com/bonzanini/topic-modelling
Motivation

Suppose you:

• have a huge number of (text) documents
• want to know what they’re talking about
• can’t read them all

github.com/bonzanini/topic-modelling
Topic Modelling

• Bird’s-eye view on the whole corpus (dataset of docs)
• Unsupervised learning
  pros: no need for labelled data
  cons: how to evaluate the model?

github.com/bonzanini/topic-modelling
Topic Modelling

Input:
- a collection of documents
- a number of topics K

github.com/bonzanini/topic-modelling
Topic Modelling

Output:
- K topics
- their word distributions

movie, actor, soundtrack, director, …
goal, match, referee, champions, …
price, invest, market, stock, …

github.com/bonzanini/topic-modelling
Distributional Hypothesis

• “You shall know a word by the company it keeps”  
  — J. R. Firth, 1957

• “Words that occur in similar context, tend to have similar meaning”  
  — Z. Harris, 1954

• Context approximates Meaning
# Term-document matrix

<table>
<thead>
<tr>
<th></th>
<th>Word 1</th>
<th>Word 2</th>
<th>Word N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc 1</td>
<td>1</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Doc 2</td>
<td>3</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Doc N</td>
<td>0</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

[github.com/bonzanini/topic-modelling](https://github.com/bonzanini/topic-modelling)
Latent Dirichlet Allocation

- Commonly used topic modelling approach
- Key idea:
  each document is a distribution of topics
  each topic is a distribution of words

github.com/bonzanini/topic-modelling
Latent Dirichlet Allocation

• “Latent” as in hidden: only words are visible, other variables are hidden

• “Dirichlet Allocation”: topics are assumed to be distributed with a specific probability (Dirichlet prior)

github.com/bonzanini/topic-modelling
Topic Model Evaluation

- How good is my topic model?
  “Unsupervised learning”… is there a correct answer?

- Extrinsic metrics: what’s the task?

- Intrinsic metrics: e.g. topic coherence

- More interesting:
  - how useful is my topic model?
  - data visualisation can help to get some insights

github.com/bonzanini/topic-modelling
Topic Coherence

- It gives a score of the topic quality
- Relationship with Information Theory (Pointwise Mutual Information)
- Used to find the best number of topics for a corpus

github.com/bonzanini/topic-modelling
Demo