3.2 Topic Modelling

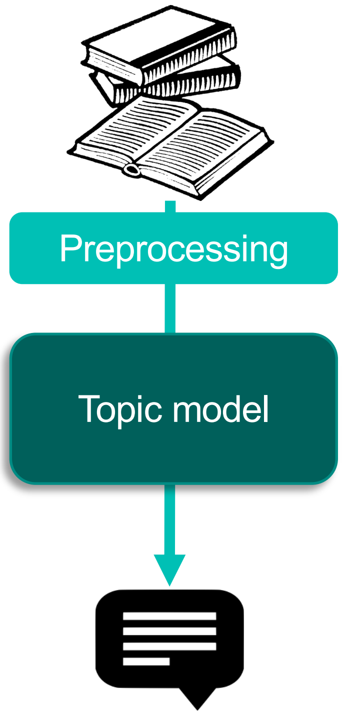
Edwin Simpson

Department of Computer Science,

University of Bristol, UK.

Topic Model

- Document is composed of multiple topics;
- Each word belongs to a single topic.



Latent Dirichlet Allocation (LDA)

Likelihood

Prior over topic distribution

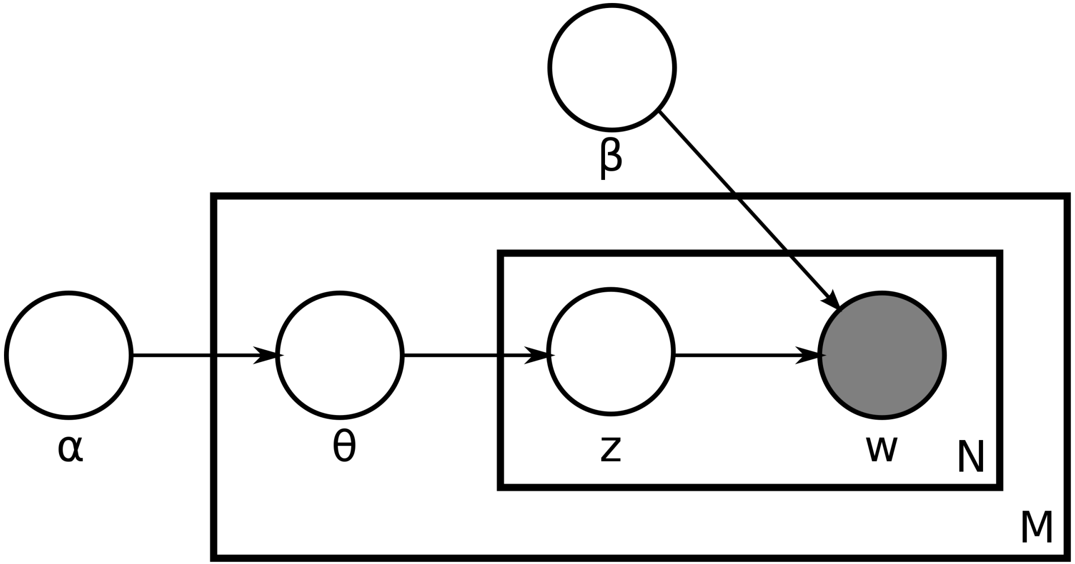
$$P(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\theta}) = \prod_{i=1}^{N} \{P(x_i | z_i) P(z_i | \boldsymbol{\theta})\} P(\boldsymbol{\theta})$$

Probability of topic for word *i*

- x_i: token at position *i*
- z_i: cluster label for word *i*
- $\boldsymbol{\theta}$: distribution over topics in this document

David Blei, Andrew Ng and Michael I. Jordan, Latent Dirichlet Allocation, JMLR (2003)

Latent Dirichlet Allocation (LDA)



David Blei, Andrew Ng and Michael I. Jordan, Latent Dirichlet Allocation, JMLR (2003)

Latent Dirichlet Allocation (LDA)

- The distributions of the LDA model have the following parametric forms:
  - $P(\boldsymbol{\theta}|\boldsymbol{\alpha}) = Dirichlet(\boldsymbol{\alpha})$: prior over topic distributions;
  - $P(x_i = w|z_i = c, \boldsymbol{\beta}) = Categorical(\boldsymbol{\beta}_c)$: likelihood of word *w* given topic *c*

David Blei, Andrew Ng and Michael I. Jordan, Latent Dirichlet Allocation, JMLR (2003)

Latent Dirichlet Allocation (LDA)

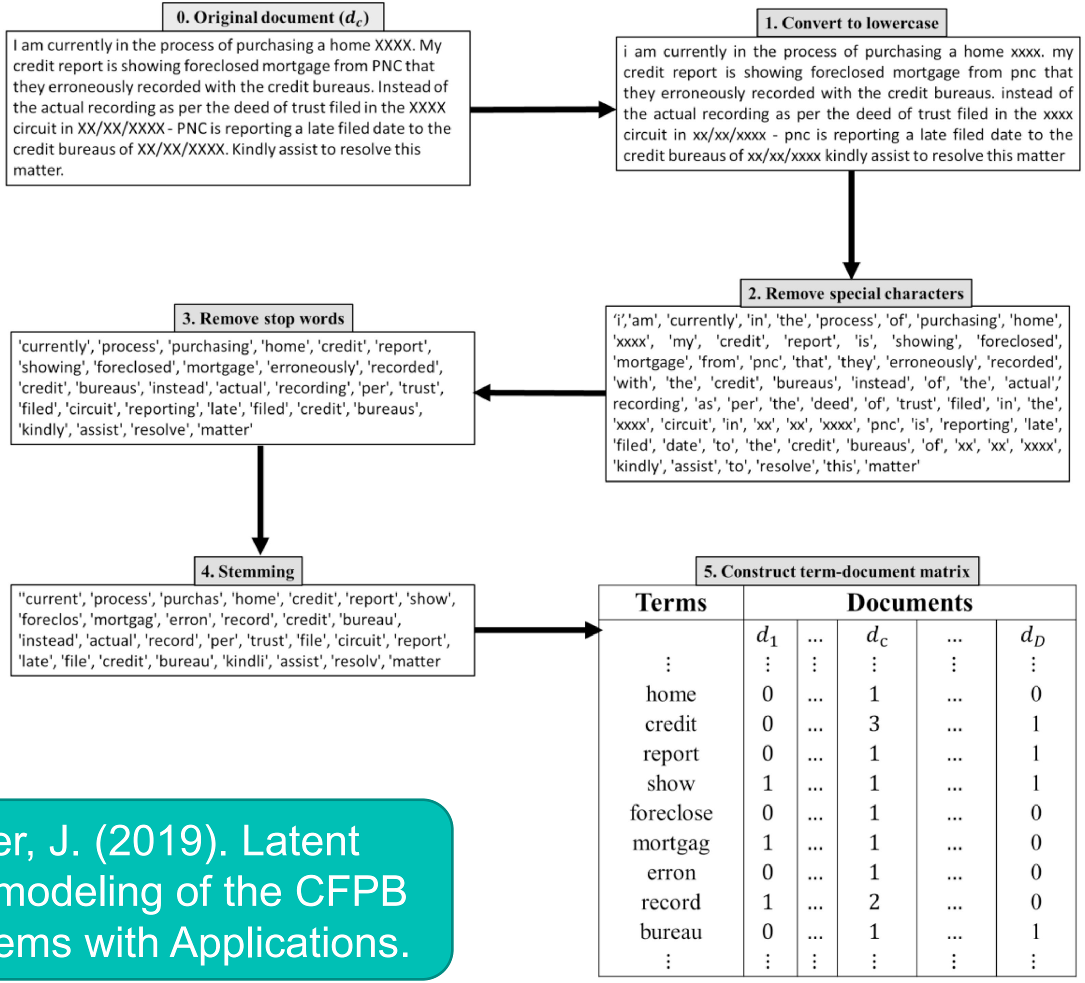
- When we perform topic modelling on a dataset, our aim is to obtain a posterior distribution over $\boldsymbol{\theta}$ for all documents.
- To do this, we need to deal with the unknown values of $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and $\boldsymbol{z}$.
- We cannot compute the posterior in closed form, so we obtain an approximation using an approach called **variational inference.**

David Blei, Andrew Ng and Michael I. Jordan, Latent Dirichlet Allocation, JMLR (2003)

Unsupervised Learning with Variational Inference

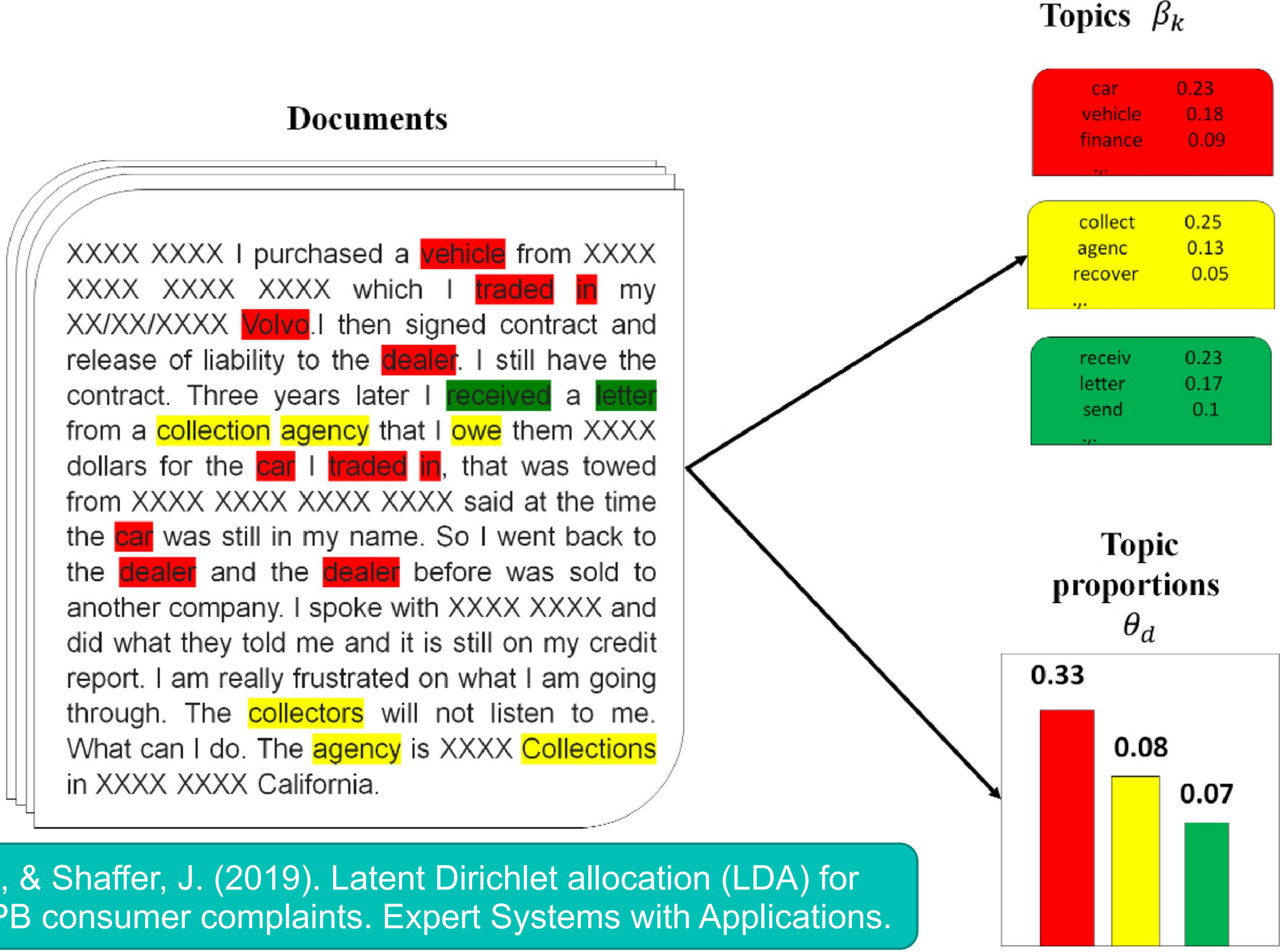
- Randomly initialise the distribution of words in each topic, $P(x_i|z_i, \boldsymbol{\beta})$
- Randomly initialise the prior over topic distributions $P(\boldsymbol{\theta}|\boldsymbol{\alpha})$
- E-step: loop over documents d:
  - Compute the **expectations** of z_i given current distributions of words for all topics
  - Compute the counts of topics in document d given expectations of z_i for all words *I*
- M-step:
  - Compute maximum likelihood estimates of the per-topic word distributions, $\boldsymbol{\beta}_{z_i} = P(x_i|z_i)$, using current expectations of z_i
  - Compute maximum likelihood estimates of the distribution $P(\theta)$, using the current counts of topics in each document

Topic Modelling: Preprocessing Pipeline



Bastani, K., Namavari, H., & Shaffer, J. (2019). Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. Expert Systems with Applications.

Example of LDA Complaint Topics



Bastani, K., Namavari, H., & Shaffer, J. (2019). Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. Expert Systems with Applications.

Summary

- Topic models identify multiple clusters – topics – within individual documents
- Latent Dirichlet allocation (LDA) is one of the most popular topic models
- Standard LDA uses bag-of-words representations
- Uses variational inference to learn the topics and topic distribution for each document