

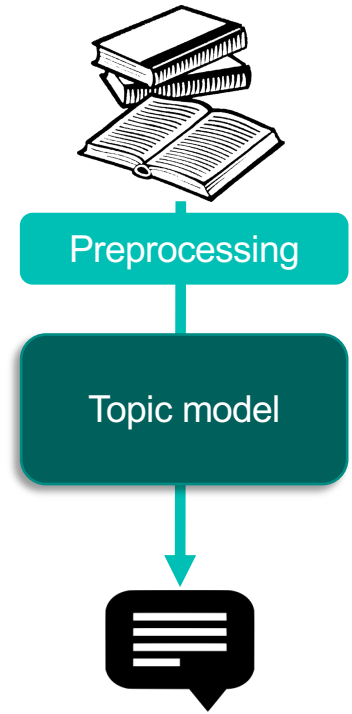
## 3.2 Topic Modelling

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)

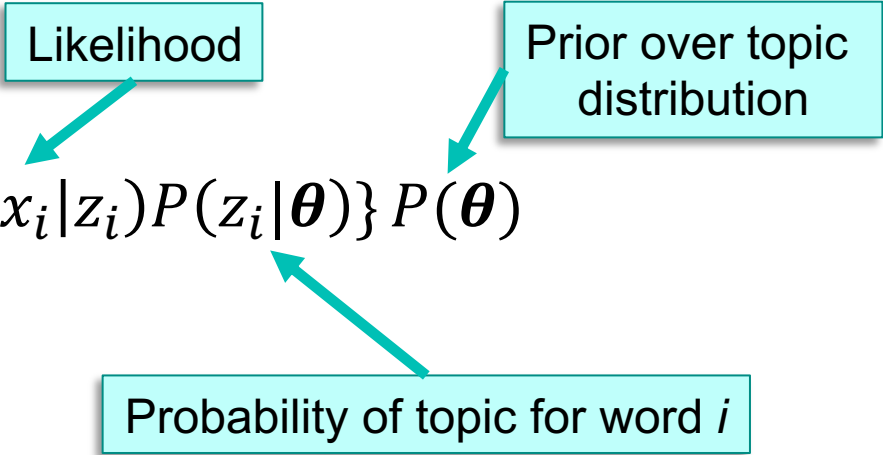
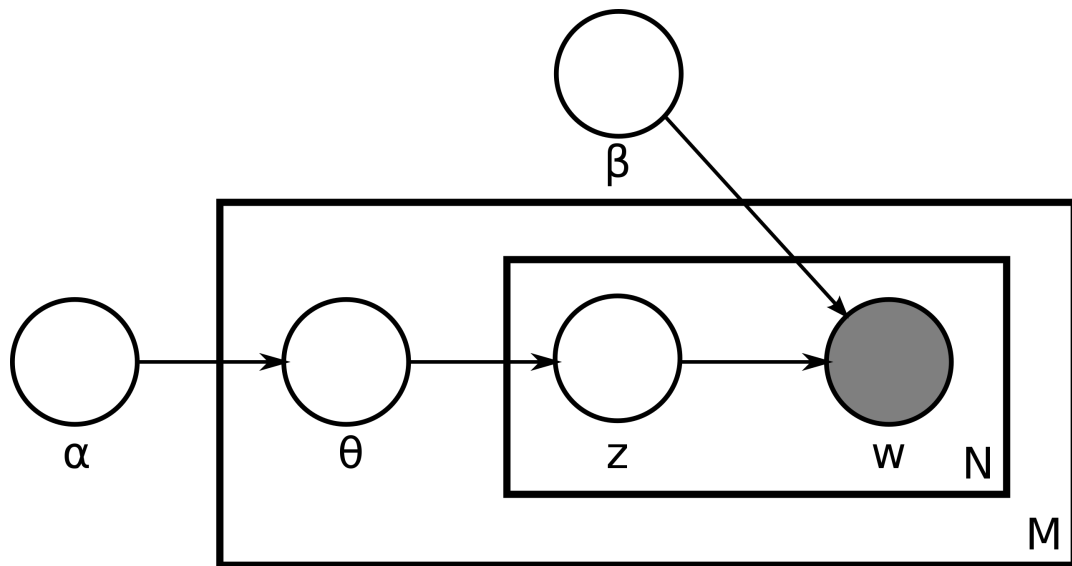
$$P(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta}) = \prod_{i=1}^N \{P(x_i|z_i)P(z_i|\boldsymbol{\theta})\} P(\boldsymbol{\theta})$$


Diagram illustrating the components of the LDA joint probability equation:

- Likelihood**: Points to  $P(x_i|z_i)$
- Prior over topic distribution**: Points to  $P(\boldsymbol{\theta})$
- Probability of topic for word  $i$** : Points to  $P(z_i|\boldsymbol{\theta})$

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

# Latent Dirichlet Allocation (LDA)



# Latent Dirichlet Allocation (LDA)

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

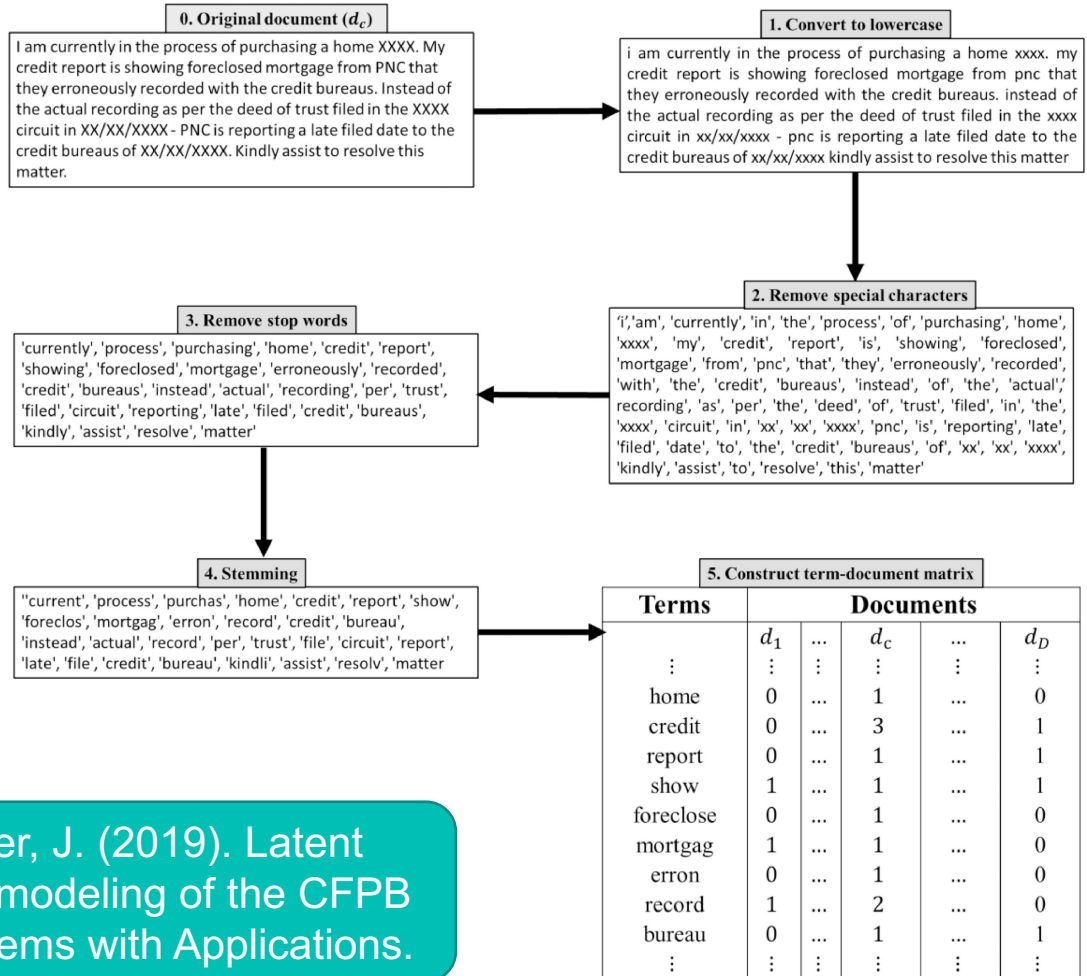
# Latent Dirichlet Allocation (LDA)

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

# 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  $l$
- 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  $P(\boldsymbol{\theta})$ , 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

## Documents

XXXX XXXX I purchased a vehicle from XXXX  
XXXX XXXX XXXX which I traded in my  
XX/XX/XXXX Volvo. I then signed contract and  
release of liability to the dealer. I still have the  
contract. Three years later I received a letter  
from a collection agency that I owe them XXXX  
dollars for the car I traded in, that was towed  
from XXXX XXXX XXXX XXXX said at the time  
the car was still in my name. So I went back to  
the dealer and the dealer before was sold to  
another company. I spoke with XXXX XXXX and  
did what they told me and it is still on my credit  
report. I am really frustrated on what I am going  
through. The collectors will not listen to me.  
What can I do. The agency is XXXX Collections  
in XXXX XXXX California.

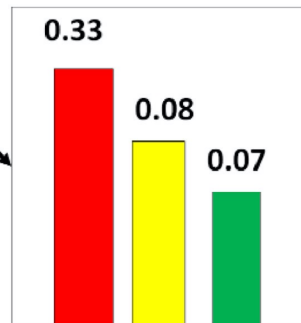
## Topics $\beta_k$

car	0.23
vehicle	0.18
finance	0.09
..	

collect	0.25
agenc	0.13
recover	0.05
..	

receiv	0.23
letter	0.17
send	0.1
..	

## Topic proportions $\theta_d$



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