Introduction to Topic Modelling – workshop 2

Bertie Vidgen & Taha Yasseri

Oxford Internet Institute (University of Oxford) & Alan Turing Institute

Tuesday 24th April 2018

Overview

Session 1	• Introduction to topic modelling • Practical – implement a topic model • Brief discussion
Session 2	• Advanced topic modelling • Practical – fit topic model parameters • Brief discussion

Aims

- How a topic model works
- Introduction to model evaluation
- Fitting of a topic model
- Analyzing topic model output

David Blei – he's kind of a big deal in topic modelling

Recap from yesterday

- LDA topic models are great…

  … *But only if you clean and prepare your data properly*

- Topics are the ‘missing link’ between documents and words
- Basics of topic modelling
  - Each document is a multinomial distribution over topics
  - Each topic is a multinomial distribution over the entire vocabulary

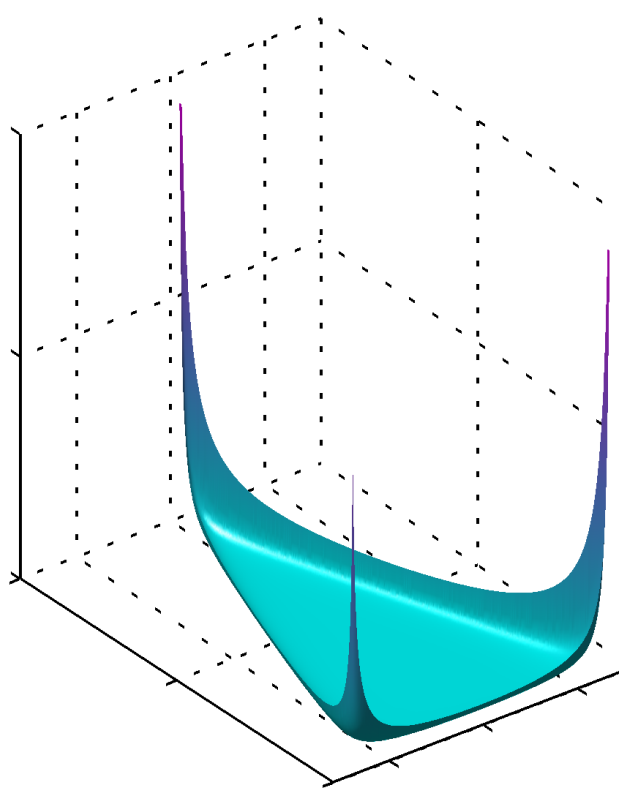
- **Any questions?**

How a topic model works – the generative process

- For each document the distribution over topics is called *theta* θ
- The Dirichlet is a distribution over all possible multinomial top distributions (all the different values of theta)
- To generate a document:
  - Sample a value of theta from the Dirichlet
    - Sample a topic from theta
    - Sample a word from the topic

… Continue until you have selected all the words in your document

Then, move onto the next document



How a topic model works – the algorithm

1

Words in documents are assigned to topics randomly

LDA usually uses an equal prior – but there are methods for introducing domain knowledge

2

Iterate through every word in every document

Update the word to topic assignments at each iteration

After very very many iterations… the model reaches convergence (i.e. there are few changes at each iteration)

3

Calculate posterior values – these are the document-level topic proportions (theta)

The algorithm tries to reverse engineer the generative process

Evaluating topic models (is not always easy)

	Qualitative	Quantitative
Intrinsic	Are the topics coherent? Many methods for qualitative assessment, such as 'intrusion'	Use just the data itself and a statistical test – semantic coherence, log likelihood and perplexity are good metrics
Extrinsic	Manually compare topic model with another dataset (typically using experts) – time consuming and subjective but insightful	Statistically compare your topic model's output with another dataset or with relevant metadata

Evaluation is difficult as topic modelling is an *unsupervised* machine learning method – there is no 'right' answer

Assessing model fit using perplexity (1)

- Perplexity
  - We have a trained topic model and want to know how well it performs at assigning topics to previously unseen documents
  - Perplexity is defined mathematically

$$b^{-\frac{1}{N}\sum_{i=1}^{N}\log_b q(x_i)}$$

- Implementation
  - We decide how many ‘folds’ to use (usually 5 or 10) and split data into that many chunks
  - We train the model, leaving aside one of the chunks (the ‘testing data’)
  - We then test the perplexity of model on the heldout testing data

Assessing model fit using perplexity (2)

- Interpretation
  - Lower perplexity means the model fits better. It finds the unseen data ‘less perplexing’
  - Perplexity should be used to *compare* different models, it is hard to interpret on its own
  - Caution – this is just a statistical test! It doesn’t ensure the model is useful or that topics intuitively make sense. You often need to play around with parameters to get useful output

Hyperparameters alpha and beta

Alpha	Beta
• Prior for the per-document topic distributions in the Dirichlet (the *shape* of documents' multinomial distribution over topics) • Lower values of alpha means documents are likely to have just a few topics • Typical values for alpha are either 0.1 or 50 / Number of topics	• Prior for the per-topic word distributions (the *shape* of topics' multinomial distributions over words) • Lower values of beta mean each topic is dominated by just a few top words • Typical values for beta are either 0.1 or 200 / Size of vocabulary

Hyperparameters are parameters which determine the shape of the *prior* distributions

First practical – fitting parameters for topic models

- Script
  - topic-modelling-workshop 2 – practical 1.R
- Data
  - workshop-2-practical-1.Rdata

Some advice on model fitting

- Only adjust one parameter at a time
- Estimate how long the model fitting will take
  - If you have 10 folds you are fitting 10 models for each level of the parameter. Then, if you have 10 values for the parameter you are fitting 10 * 10 = 100 models. *And* higher values of k usually increase the run time. So a model with 100 topics will take longer to fit than a model with 10 topics
- Don’t bore readers with your model fitting
  - Model fitting details is just the perfect thing to slot into an appendix
- Don’t be constrained by the results of fitting
  - Particularly when picking the number of topics, go with values which work rather than exactly what your model fitting indicates
- Save your results!

Analyze the output of topic models

- Topic proportions in the entire corpus
  - What are the most prevalent topics?
- Topic prevalence over time
  - How does topic prevalence fluctuate? Are there discernible patterns?
  - Can we link changes in topic prevalence to external events?
- Topic connections
  - Which topics co-occur with each other in documents?
  - Which topics are similar in terms of the distribution over words?
- Topic prevalence for different groups

Second practical – analyzing the output

- Script
  - topic-modelling-workshop 2 – practical 2.R
- Data
  - workshop-2-practical-2.Rdata

Thank you!

- Email me if you have any issues with the code or want advice on implementing a topic model on your own data

bertie.vidgen@oii.ox.ac.uk

References

- Mark Steyvers (2004) ‘Probabilistic topic models’ in Landauer et al. (eds.) *Latent semantic analysis: a road to meaning* New York: Laurence Erlbaum
  - In-depth overview of how topic models work, with some (fairly) accessible explanation of the maths and the infamous ‘plate notation’
- Mimno (2011) ‘Optimizing semantic coherence in topic models’, *Proceedings of the 2011 Conference on Empirical Methods in NLP*, pp. 262-272
  - Good overview of the problem of topics evaluation, and also puts forward a nice measure for evaluating topics (the UMass measure of topic coherence)
- Baumer et al. (2017) ‘Comparing grounded theory and topic modelling: extreme divergence or unlikely convergence’, *Journal of the Association for Information Science and Technology*, 54: 1
  - Interesting discussion of the comparisons between topic modelling and grounded theory
- Discussion on topic evaluation (implementation in Python but it is still relevant)
  - https://www.youtube.com/watch?v=UkmIljRIG_M&t=2073s
- Introductory material on perplexity
  - https://jamesmccaffrey.wordpress.com/2016/08/16/what-is-machine-learning-perplexity/
  - https://planspace.org/2013/09/23/perplexity-what-it-is-and-what-yours-is/
  - https://www.youtube.com/watch?v=NlmKb0X-nkA

Some extensions to the classic LDA model

- Correlated topic models (CTM)
  - Explicitly models dependencies between topic co-occurrence in documents
  - Blei and Lafferty (2007) 'A correlated topic model of Science'
- Hierarchical topic models (hLDA)
  - Models nested topic hierarchies; can be used to automatically pick the right number of topics
  - Blei et al. (2004) 'Hierarchical topic models and the nested Chinese restaurant process'
- Supervised topic models (sLDA)
  - Uses document-level topic distribution to predict a fixed category
  - Blei and McAuliffe (2010). 'supervised topic models'
- Structured topic models (STM)
  - Introduces arbitrary metadata to the topic model; you can see how topics' content and prevalence varies
  - Roberts et al. (2013) 'The structural topic model and applied social science'
- Topics over Time (ToT)
  - Explicitly models how topic prevalence changes over time
  - Wang and McCallum (2006) 'Topics over time: a non-Markov continuous-time model of topical tends'
- And a plethora of other topic models incorporating various metadata and content data, such as author, sentiment, geography…