Department of Computer Science
The University of Melbourne
COMP90042 WEB SEARCH AND TEXT ANALYSIS (Semester 1, 2018)

Workshop exercises: Week 9

Discussion

1. In modelling terms, what is the difference between **topics** and **classes**?

   A document can have multiple topics but a single class only. Also, both have semantic intepretations but classes have labels while topics have not.

2. What is a **topic model**? What is the difference between topic modelling and text classification?

   From Blei (2012): topic models are algorithms for discovering the main themes in a large and unstructured collection of documents. Standard topic modelling is unsupervised and can model multiple topics per document. Text classification is supervised and assume a single class per document.

3. Give 3 example applications for topic models. Explain why it is not feasible to use text classification for these applications.

   - Organising historical documents
   - Finding trending topics on Twitter
   - Make sense of scientific publications
   - Stance detection on social media
   - Mining parallel data for translation
   - Query expansion in IR

   These applications usually do not have annotated data and most do not have a specific class taxonomy to apply. Therefore it is not practical to perform classification.

4. It is possible to train a topic model using unsupervised HMMs but this is not ideal. Why? How it can be improved?

   Because standard HMMs assume that the topic of a word is independent of the document where that word is. A simple improvement is to allow per-document HMMs.

5. How can you evaluate topic models automatically?

   Using perplexity on held-out test corpora.

6. Cite 2 example visualisations for evaluating topic models manually.

   - Word lists
   - Word clouds
   - Labelling using article names
   - Labelling using pictures

7. Cite 3 extensions of LDA and what kind of problems they address.
   - LDA-HMM: remove "bag-of-words" assumption
   - Hierarchical LDA: models topic hierarchy ("sports" -¿ "football")
   - Correlated LDA: assume similarity between topics ("football" / "rugby" vs. "football" / "genetics")
   - Dynamic LDA: assume that topics change over time (words that form a topic in 1920 are different from the words from the same topic in 2000) (check Blei (2012), Figure 5 for an example)
   - Non-parametric LDA: does not need to fix the number of topics

Programming

1. Go through the `WSTA_N15_topic_models` notebook. What kind of topics do you get in your final output? Can you label all of them? How would you improve the interpretability of these topics?

Catch-up

- What is a language model?
- What is the difference between n-gram LMs and neural LMs?
- How do you evaluate language models?

Get ahead

- Try some of the extensions proposed in the `WSTA_N15_topic_models` notebook.
- Try the Gensim tutorial on finding topics on Wikipedia (`https://radimrehurek.com/gens`
  Beware though: training on Wikipedia can take quite a long time.