Programming assignment 3: Topic modelling

8 Points Possible

2025-11-24

Attempt 1

In Progress

NEXT UP: Submit assignment

Add comment

Unlimited Attempts Allowed

2025-11-16 to 2025-11-24

Details

This is the Canvas submission page for Assignment 3 on topic modeling.

Your tasks here are to

1. Write your own code for doing Gibbs sampling for LDA. Here you can choose if you want to do Alternative 1 or collapsed Gibbs, i.e. Alternative 2, in the lecture slides. However, Alternative 1 may run slower than collapsed Gibbs, so I recommend collapsed Gibbs.
2. Evaluate your results visually and via coherence score.

Before implementing Gibbs sampling, stop word removal and tokenization is necessary. There are Python functions for doing that. It is also a good idea to remove words that do not appear in the corpus more than a certain number of times, e.g. 10. For text corpus you are free to choose one that you like, but a tip is that news article corpora work well for topic models. If you do that, you can try the AP Corpus, the Reuter Corpus or 20 News Groups. You will only be able to run the algorithm on a small subset of the chosen corpus. The larger the better of course, but a corpus of at least 200 000 words should be manageable.

Run this for different hyperparameters. You can try $\alpha = \beta = 0.1$ and $\alpha = \beta = 0.01$. (A cross-validation search for optimal values will probably be too slow.) Run also for different numbers of topics, e.g. $K = 10$ and $K = 50$.

From experience, it seems that something between 100 and 200 iterations over the corpus is sufficient for the Gibbs sampler to mix, so that what you get is something that is sufficiently close to the posterior distribution of topics given the text. Collect only one sample. (Since the topic number assigned to a topic in each run is completely random, using multiple samples will cause the extra challenge of identifying topics between the runs.)

Assess model performance in tables over the twenty top words for the topics that seem to make the most sense. Classifying top words can be interpreted in two ways: 1) the most common words by raw count for each topic, or 2) the most common words by relative frequency, i.e. the number of times the word appears in the topic divided by the total number of times that the word appears in the corpus. You are free to choose which interpretation to use or if you want to try both (after having done one, doing the other will probably take very little effort).

Assess also the topic qualities with the Umass coherence score. Observe if a good coherence score really corresponds with your own judgment of if a topic makes sense or not. Use the top twenty words for the Umass score.

Umass coherence score:

https://dl.acm.org/doi/pdf/10.5555/2145432.2145462.

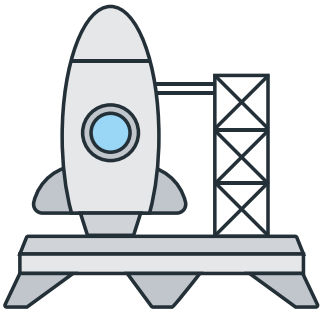
You should submit a report and code (source files, notebooks, or Colab links). The report must be a pdf document and it should be in IMRAD format. It can be short, but the introduction must at least tell the reader:

- how the LDA model works, including a presentation of the algorithm with its mathematical formality,
- that you do Gibbs sampling and how that works for sampling from LDA (e.g. by pseudo-code or a bullet list with the same information),
- a definition of the Umass score.

The analysis and discussion can be just one part, where you state your thoughts about why the results were what they were and what could maybe improved if your were allowed to come back to spend some more time on the problem. The report must be reader friendly, so that the reader can easily read what you do and why. It should be written so that your text is understandable also for the reader who is not already aquainted with them.

Remark: One may run training in batches, processing a subset of the documents at a time. However, this will lead to the same problems of identifying topics between different batches (i.e. the same problem as with running the whole algorithm several times, as mentioned above). So I do not recommend batch training.

Keep in mind, this submission will count for everyone in your Assignment 3 group.



Choose a file to upload

File permitted: PDF, IPYNB, PY

or

Canvas Files

Submit assignment