Report: Generative Processes of Static and Dynamic LDA Models

Arijus Pleska

December 1, 2016

During this week I have familiarised with the generative processes of static and dynamic LDA models. In this report I will present the results by plotting the generated corpora, address the initialisation of the parameters α and β, and raise questions for the next meeting.

I have managed to generate documents using both static and dynamic variations. The static approach in generating corpus is given in Figure 1, and the dynamic approach is given in Figure 2.

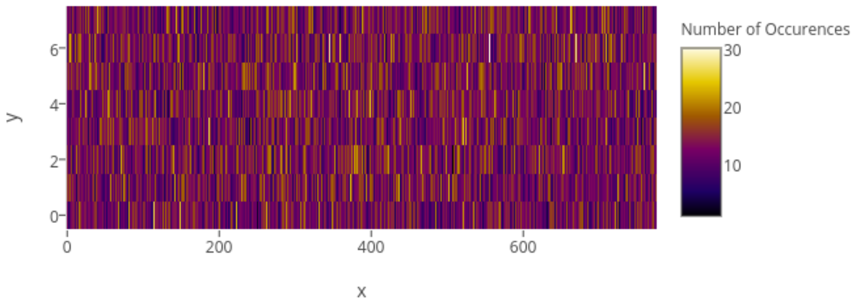

Figure 1: The distribution of a word in a corpus produced by the static generative process.

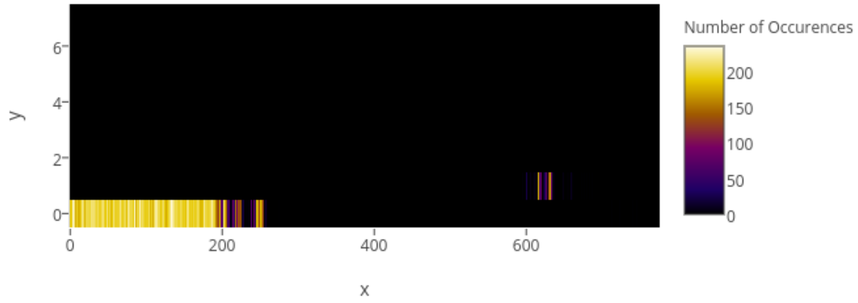

Figure 2: The distribution of a word in a corpus produced by the dynamic generative process.

In Figure 1 we can see that the word is present throughout the corpus, whereas Figure 2 displays a smoother distribution of the word. That is, the dynamic generative process induces a structure of the word's existence in the corpus. Also, note that both corpus contain the same number of documents and the number of words in a document is drawn from the Poisson distribution on the same mean $\xi = 200$.

The implementation of the generative processes follows directly from the original LDA paper [1] and the paper on dynamic topic models [2]. The Python Notebook containing the implementation of the processes can be found on the following link: `https://github.com/perdaug/mlinb/blob/master/notebooks/generating_corpus.ipynb`. Note that the initial values of the parameter α are positive and random, whereas the initial values of the parameter β are equal and normalised for each topic. This formulation of β suggests that each word has an equal probability of getting drawn in the first document generation of the dynamic model and throughout all generated documents of the static model. Also, note that every generated document in the dynamic model is unique with respect to other documents. In other words, each dynamic time-slice contains one document.

During the upcoming meeting I would like to address the following questions:

1. Should we use the variational inference for the implementation of the dynamic topic model? If not, would Gibbs sampling scale with the future datasets?

2. Why Dirichlet distribution is amenable in the dynamic topic modelling? (This is claimed in the last paragraph of the first page of the dynamic topic modelling paper [2])

3. For the dynamic model implementation, should we update α_{t-1} (as discussed in Subsection 5.3 of the original LDA paper [1]) before drawing α_t?

References

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[2] David M. Blei and John D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 113–120, New York, NY, USA, 2006. ACM.