Topic Model Tutorial

A basic introduction on latent Dirichlet allocation and extensions for web scientists

Christoph Carl Kling[1] Lisa Posch[1] Arnim Bleier[1] Laura Dietz[2]
[1]GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany
[2]Data and Web Science Group, University of Mannheim, Germany
ckling@uni-koblenz.de, lisa.posch@gesis.org, arnim.bleier@gmail.com,
dietz@informatik.uni-mannheim.de

CCS Concepts

•Mathematics of computing → Probability and statistics; •Information systems → Data mining;

Keywords

Topic models, Dirichlet distribution, LDA, HDP

1. OUTLINE

In this tutorial, we teach the intuition and the assumptions behind topic models. Topic models explain the co-occurrences of words in documents by extracting sets of semantically related words, called topics. These topics are semantically coherent and can be interpreted by humans. Starting with the most popular topic model, *Latent Dirichlet Allocation* (LDA), we explain the fundamental concepts of probabilistic topic modeling. We organise our tutorial as follows: After a general introduction, we will enable participants to develop an intuition for the underlying concepts of probabilistic topic models. Building on this intuition, we cover the technical foundations of topic models, including graphical models and Gibbs sampling. We conclude the tutorial with an overview on the most relevant adaptions and extensions of LDA.

1.1 Developing an Intuition

In the first part, we provide the participants with an intuition of the ideas and assumptions behind *probabilistic topic models*. First, we present easily understandable metaphors (following the Pólya urn scheme) to introduce the multinomial and the Dirichlet-multinomial distribution and the role of the parameters for the symmetric Dirichlet distribution. Furthermore, we introduce the concept of modelling a corpus of documents as a mixture of Dirichlet-multinomial distributions. We then train LDA on text corpora and demonstrate the effects of different parameter settings on the trained topic models. In order to deepen the intuition, we conclude this part with a game with a purpose (based on the word-intrusion task by [2]), enabling a human evaluation of model parameters.

1.2 Technical Foundations

After developing the intuition, in the second part of the tutorial we show how the assumptions in the metaphors translate to the single parts of Latent Dirichlet Allocation (LDA)[1], the most cited topic model in the scientific community. We provide a translation of the gained intuition to detailed definitions. In particular, we aim to cover concepts such as *closed form inference, approximate inference* with a focus on Gibbs sampling, *generative storyline* and *plate notation*. For each of the introduced concepts, we provide illustrative implementation examples.

1.3 Adaptations and Extensions

LDA has been adapted and extended to a wide range of specific settings. In this part of the tutorial, we present adaptations relevant for the social sciences. Examples include models exploiting context information such as L-LDA, a supervised variant of LDA; PL-TM, a topic model for multilingual settings; the Citation Influence Model, modelling the influence of citations in a collection of publications [3].

1.4 Evaluation, Discussion of Pros and Cons

While a useful tool for exploitative analysis of unfamiliar data collections, topic models were disputed in the recent past. Common error modes are discussed and the critique about topic models is summarized. We emphasize the importance of evaluating any exploratory tool in domain of interest before drawing conclusions. To enable participants to make an informed decision, we discuss several avenues for in-domain evaluation.

2. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[2] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In *NIPS*, pages 288–296, 2009.

[3] L. Dietz, S. Bickel, and T. Scheffer. Unsupervised prediction of citation influences. In *ICML*, pages 233–240. ACM, 2007.

WebSci '16 May 22-25, Hannover, Germany

ACM ISBN 978-1-4503-4208-7/16/05.

DOI: http://dx.doi.org/10.1145/2908131.2908142