TOM: A library for topic modeling and browsing

Adrien Guille*, Edmundo-Pavel Soriano-Morales*

*Laboratoire ERIC, Université Lumière Lyon 2
adrien.guille@univ-lyon2.fr, edmundo.soriano-morales@univ-lyon2.fr

Abstract. In this paper, we present TOM (TOpic Modeling), a Python library for topic modeling and browsing. Its objective is to allow for an efficient analysis of a text corpus from start to finish, via the discovery of latent topics. To this end, TOM features advanced functions for preparing and vectorizing a text corpus. It also offers a unified interface for two topic models (namely LDA using either variational inference or Gibbs sampling, and NMF using alternating least-square with a projected gradient method), and implements three state-of-the-art methods for estimating the optimal number of topics to model a corpus. What is more, TOM constructs an interactive Web-based browser that makes exploring a topic model and the related corpus easy.

1 Introduction

Topic models are useful tools for unveiling the latent topical structure of text corpora. They can make searching, browsing and summarizing these corpora easier. Several models and algorithms to approximate them have been proposed in the recent years. The quality of the discovered topics depends on the model, the approximation algorithm, the nature of the corpus being studied, as well as the number of topics (Stevens et al., 2012). Therefore, in order to perform an efficient topic-based analysis of a text corpus, it is important to compare several approaches to identify the most relevant topics. However, this is a difficult task because the existing implementations of the approximation algorithms are independent, which means that one has to learn how data are structured in each implementation; what the functions to manipulate each topic model are; how to fit these topic models on the exact same set of features, *etc*. On the other hand, several methods have been proposed to estimate the optimal number of topics to model a corpus, but – to the best of our knowledge – their implementations are not publicly available.

In this short paper we present TOM (TOpic Modeling), an open source library written in Python for analyzing a text corpus from start to finish, via the discovery of latent topics. Apart from advanced corpus preparation functions, TOM offers a unified interface for existing robust implementations of approximation algorithms, that makes fitting and manipulating topic models easy. It also implements several functions to estimate the optimal number of topics to model a corpus. What is more, TOM can automatically build a Web based interface for exploring a topic model and a corpus in an interactive manner.