Is my LDA on topic?

Using external labels for model selection and validation

Introduction

Two questions surface in the literature on topic model validation: what constitutes a good topic model and do we (still) need humans? In the context of specific practical applications, generating synthetic documents or referring to known labeled sets may not suffice. Often, we use the wordlists associated with topic models as the basis for our evaluation of models. (Mimno, Wallach, Talley, Leenders, & McCallum, 2011; Rosen-Zvi, Griffiths, Steyvers, & Smyth, 2004). Like reading tealeaves, this approach may leave something to be desired. Humans tend to evaluate topic models based on the consistency of words within a given topic (Chang, Gerrish, Wang, Boyd-graber, & Blei, 2009). In practice, what models may aim to provide topics which would behave in parallel to those based in qualitative coding. To assess correspondence between automatically generated topics and those conceptualized by humans, validation may require human coders which interpret the sense of words-in-topics or topic-in-documents to validate and possibly compare topic models. However, this may not be feasible in larger contexts, as human coders are relatively expensive and the amount of topics may be extensive. With varying starting parameters and preprocessing decisions, the number of potential topic models may increase exponentially. In contexts without labeled training sets, a quick sanity-check approach to picking the most promising models may facilitate better research and provide grounding for parameter selection decisions. In this abstract, I explore the potential of exogenous labels as a way

to glean which model in a set of topic-models may best serve the purpose of document classification for a given task.

Topics are often used in the domain of written texts belonging to labeled categories. Some modeling approached – such as structural LDA models (Griffiths & Tenenbaum, 2004) - incorporate this information in the modeling process[1]. External labels may denote categories on other levels of abstraction than the topics of interest, e.g.: the economy section (label) of a newspaper may discuss the European crisis (event). Such external categories may serve as a reference ordering of topics.

Assumptions

In this preliminary setup, I assume external labels can serve as categories – or fields- of semantic meaning. Specifically, I assume that a good topic model would provide topics that 1) are positively associated with each category, their predictive value and 2) can discriminate between categories, their discriminative validity. Neither are assumed to be strong, so that simply true/false guesses or single/multiple relations cannot be used as quality indicators.

The proposed process

Ideal topic models should have a set of topics that can together predict each of the exogenous categories. Although multiple topics may associate with the same category in the same way multiple topics may be associated with the same document (Blei, Ng, & Jordan, 2003). I propose looking at this with

1 In our study, the use of this information in topic modeling would confound the effect of shared topics on association with joint membership of categories.

simple negative binomial regression models (NB), which should be able to correctly model their predictive power while correcting for the inflated zero rate associated with uniquely coded labels. Each topic model can be tested in predictive power related to each possible exogenous label, using a three-step process for each of the labels of interest:

1. Explore predictive power of topics by running a full NB model, remove values which cannot be estimated.

Depending on label-topic incidence rates, some topics may not occur for a given label and can therefore not be estimated. These topics have no predictive power and can be dropped.

2. Find strongest associated topics, retain significant predictors.

Initial models will include the full range of topics, leading to increasing odds of multicollinearity problems. To remove noise-based influence of non-significant topics on the predictive power of the model these parameters are dropped.

3. Depending on specific application: Drop non-positive predictors.

In some cases, having a negative predictor of labels is not of interest. Such topics may have spurious low occurrence within the context of labeled documents, or there may be a substantive reason why some topics do not occur in the set of documents under this label.

After these steps, the topic model under consideration can be described using the vector of BIC values for each of labels' model, the number of predictors (topics) associated with each label and the ratio of unique predictors.

Interpretation

The BIC and predictor count vectors can be used as summaries of the predictive validity of topic models. Compared to a baseline model (BIC of the

prediction of the intercept only), the improvement in BIC of each model indicates the extent to which the supposedly best fitting subset of topics are associated with this label. BIC is preferred over AIC or simple log-likelihood because of the focus on parsimony. The vector of predictor counts allows topic model comparison based indicated complexity within a label. High predictor counts would indicate the number of topics associated with a specific label rises, which may indicate some models are better at subdividing some categories.

The ratio of unique to total predictors provides some indication about each models ability to distinguish topics discussed in documents under different labels. Although the interpretation of shared predictors depends on the nature of labels specific to the application, they may be indicative of super-categorical word sets, such as words associated with the Euro-crises that positively predict both international and financial categories, but are not able to distinguish the focus of documents.

Our test case

Our use-case revolves around organizational websites in the non-profit sector which provide or receive grants in the United States. These organizations aim to fulfill a societal aims of various kinds, and are classified by the IRS in 26 distinct categories ranging from A: Arts to Y: mutual benefit organizations (Z is reserved for Unknown/unclassified). We gathered the text content of 344.000 websites between 2008-2011 from the Wayback Archive[2] . Our analysis aims to

2 http://archive.org/web/

include discussed topics on their websites as shared interests. The assumption of partial correspondence in this context is based on the insight that the kinds of missions should relate to topics discussed; Art organizations can discuss museums, which mutual benefits (such as fraternities) are unlikely to discuss.

Of particular interest is the use of noun versus adjective based topic models in our dataset. Secondary is the thresholds applied to filter out too common or infrequent words. The combination of both aspects quickly multiplies the number of possible topic models to use in classifying these websites. Using either adjectives or nouns, with a upper-bound fraction of occurrences for words between 0.85, 0.9 and 0.95. we have 6 candidate models to use to classify documents for further analyses.

Expected Results

We aim to test the outlined procedure to look for the best model in our context. The analysis should provide us with an initial comparison base of the models, which we will use as a basis for qualitative assessment. By looking at the word-topic loadings, topic-document loadings and document-category relation of selected items, we aim to assess the validity indicators as proposed. If usefull, we will further develop the pipeline, simple graphical summaries and interpretation steps for future use. In summary, by providing a model comparison framework for associated topic models based on distinct wordlists, parameters or other characteristics but sharing the same underlying set of documents together with external labels, we hope to ease the selection of topic-models in applications where human coding is not feasible.

Literature

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems* (pp. 288–296). Retrieved from http://machinelearning.wustl.edu/mlpapers/paper_files/NIPS2009_0125.pdf

Griffiths, D., & Tenenbaum, M. (2004). Hierarchical topic models and the nested Chinese restaurant process. *Advances in Neural Information Processing Systems*, *16*, 17.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 262–272). Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=2145462

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence* (pp. 487–494). AUAI Press. Retrieved from http://dl.acm.org/citation.cfm?id=1036902