Incorporating Prior Lexical Information in Topic Models

Jagadeesh Jagarlamudi
jags@umiacs.umd.edu
University of Maryland

Raghavendra Udupa
raghavu@microsoft.com
Microsoft Research India

Hal Daumé III
hal@umiacs.umd.edu
University of Maryland

Recently topic models have emerged as a powerful tool to analyze document collections in an unsupervised fashion. The seminal work by Blei. et al. [1], starts by assuming that each document is a mixture of a set of topics where each topic is in turn a combination of words from the vocabulary. When fit to a document collection, the model inherently uses the co-occurrence information to group semantically related words into a single topic. Since then, many extensions have been proposed to improve the semantic coherence of the words in each topic. Because of the unsupervised nature of these approaches, a user has no way to specify the intended topics and moreover he/she is left with the job of making sense of the topics learnt by the model.

In this paper, we address this problem by providing a simple yet effective mechanism to guide the model to learn desired topics by providing seed words in each topic. For example, the user can gather seed words for each of the dmoz categories and provide them as input. This enables the model to analyze a document collection in terms of these well known categories.

Like in LDA, each document is assumed to be a mixture over topics but each topic is a convex combination of a seed topic and a traditional LDA style topic. Here we assume that there is a one-to-one correspondence between seed topics and LDA style topics. But this can be easily modified to handle the case where a topic is associated with multiple seed topics and vice versa. To understand the intuition, consider a seed topic (say 4) with words {grain, wheat, corn} now by assigning all the related words such as 'tonnes', 'agriculture', 'production' etc. to the same topic (i.e. topic 4) the model can potentially put high probability mass on topic 4 for agriculture related documents. Otherwise the model has to distribute the probability mass on the topic 4 and also the other topic which contains the new agriculture related words and as a result it will pay more penalty. Thus the model starts from seed topics and groups related words into the same topic and as a consequence we hope the document topic distributions become more focussed.

We assume that the corpus is generated based on the following generative process:

1. For each topic k=1...T, choose $\phi_k \sim \mathrm{Dir}(\beta_1)$.
2. For each *seed* topic k=1...T, choose $\phi_k^s \sim \mathrm{Dir}(\beta_2)$[1].
3. For each document d, choose $\theta_d \sim \mathrm{Dir}(\alpha)$.
    - For each token $i = 1 \cdots N_d$:
        (a) Select a topic $z_i \sim \mathrm{Multinomial}(\theta_d)$.
        (b) Select an indicator variable $x_i \sim \mathrm{Binomial}(\mu_{z_i})$.
        (c) if x_i is 0
            – Select a word from $p(w_i|\phi_{z_i})$.   // choose from LDA style topic
        (d) if x_i is 1
            – Select a word from $p(w_i|\phi^s_{z_i})$.   // choose from seed topic

[1] Note that for a seed topic only user input words will have non-zero probability.

Experiments

We use the document clustering task to evaluate our model in comparison with the LDA model. We consider only the top-5 categories of the Reuters-21578 [3] corpus. The number of documents in each of these categories is shown in Table 1. As it can be seen that the number of documents in each cluster varies considerably. Under such circumstances, since topic models are trained to maximize the likelihood of the data, in the absence of any guidance the model focusses on explaining the documents of the majority class. The Table also shows the seed words used for each category. There are many possible ways to obtain the seed words. For the experiments reported here, seeds are obtained manually using the related searches information provided by a popular search engine. Also note that the number of seed words in each class are different. We use collapsed Gibbs sampling [2] to learn the parameter values with five topics. We use the learnt document-topic probability distributions to cluster documents in the following way: for each document, we take the most likely topic and declare the document as belonging to that cluster.

Class	# of documents	Seed words
Earn	2709	company, share, billion, pct, note, oper quarter, record, avg, shrs, earnings
Acquisition	1488	acquisition, procurement, transfer, development consolidation, integration, merge
Foreign Exchange	461	forex, fixed, income, currency, trading, euro, dollar, rate interbank, commodity, foreign, exchange, rates, converter
Grain	395	grain, wheat, corn, forage, oilseed, silage
Crude	351	natural, gas, oil, fuel, products, petrol, energy, arabia

Table 1: Top-5 categories of Reuters corpus and the seed words used for each of the class. The user is required to input only seed words and the probabilities are learnt automatically

The results are shown in Table 2. First we report the average entropy of the document-topic probability distributions (second column). The decrease in the entropy values with our model suggests that our model successfully learns focussed document-topic distributions. We also report the clustering accuracy measured in terms of F-measure, Rand Index (RI) and Variational Information (VI) in the same Table. Note that, for VI a lower value is indicative of a better performance. From the results it is clear that the document-topic distributions learnt by our model better capture the underlying topicality of the document collection.

Method	Avg. Entropy	F-measure	RI	VI
LDA	0.91	0.6578	0.8051	1.127
Our model	0.89	**0.685**	**0.8132**	**1.116**

Table 2: Clustering accuracy of LDA and our model on the top-5 categories of Reuters-21578 corpus.

References

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Maching Learning Research*, 3:993–1022, 2003.

[2] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of National Academy of Sciences USA*, 101 Suppl 1:5228–5235, April 2004.

[3] D. D. Lewis, Rose Yang, Y., T., and F Li. Rcv1: A new benchmark collection for text categorization research. In *Journal of Machine Learning Research*, volume 5, pages 361–397, 2004.