jLDADMM: A Java package for the LDA and DMM topic models

Dat Quoc Nguyen

Department of Computing
Macquarie University, Australia
dat.nguyen@students.mq.edu.au

Abstract. The Java package jLDADMM is released to provide alternatives for topic modeling on normal or short texts. It provides implementations of the Latent Dirichlet Allocation topic model and the one-topic-per-document Dirichlet Multinomial Mixture model (i.e. mixture of unigrams), using collapsed Gibbs sampling. In addition, jLDADMM supplies a document clustering evaluation to compare topic models.

1. Introduction

jLDADMM is released to provide alternative choices for topic modeling on normal or short texts. Probabilistic topic models, such as Latent Dirichlet Allocation (LDA) [2] and related models [1, 8], are widely used to discover latent topics in document collections. However, applying topic models for short texts (e.g. Tweets) is more challenging because of data sparsity and the limited contexts in such texts. One approach is to combine short texts into long pseudo-documents before training LDA. Another approach is to assume that there is only one topic per document.

jLDADMM provides implementations of the LDA topic model [2] and the one-topic-per-document Dirichlet Multinomial Mixture (DMM) model (i.e. mixture of unigrams) [9]. These implementations of LDA and DMM use Gibbs sampling for inference as described in [3] and [10], respectively. Furthermore, jLDADMM supplies a document clustering evaluation to compare topic models, using two common metrics of Purity and normalized mutual information (NMI) [5].

jLDADMM is available to download at http://sourceforge.net/projects/jldadmm/

Bug reports, comments and suggestions about jLDADMM are highly appreciated. As a free open-source package, jLDADMM is distributed on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.

2. Using jLDADMM for topic modeling

This section is to describe the usage of jLDADMM in command line or terminal, employing the pre-compiled `jLDADMM.jar` file. Here, we supposed that Java is set to run in command line or terminal (e.g. adding Java to the environment variable `path` in Windows OS).

When unzipping the downloaded file `jLDADMM_v1.0.zip`, users can find the pre-compiled `jLDADMM.jar` file and source codes in the `jar` and `src` folders, respectively. The users can recompile the source codes, simply running `ant` (also supposed that `ant` is already installed). In addition, uses can find input examples in the `test` folder.

Input corpus format: Similar to the `corpus.txt` file in the `test` folder, jLDADMM assumes that *each line in the input corpus file represents a document* (i.e. a sequence words/tokens separated by white space characters). The users should preprocess the input corpus before training the LDA or DMM models, for example: down-casing, removing non-alphabetic characters and stop-words, removing words shorter than 3 characters and words appearing less than a certain times in the input corpus.

Now, we can train LDA or DMM by executing:

```
$ java [-Xmx1G] -jar jar/jLDADMM.jar -model <LDA_or_DMM> -corpus <Input_corpus_file_path>
[-ntopics <int>] [-alpha <double>] [-beta <double>] [-niters <int>] [-twords <int>]
[-name <String>] [-sstep <int>]
```

where parameters in [ ] are optional.

- `-model`: Specify the topic model LDA or DMM
- `-corpus`: Specify the path to the input training corpus file.
- `-ntopics <int>`: Specify the number of topics. The default value is 20.
- `-alpha <double>`: Specify the hyper-parameter alpha. The default value is 0.1. See experimental details in [10, 4].
- `-beta <double>`: Specify the hyper-parameter beta. The default value is 0.01 which is a common setting in the literature [3]. Following [10], the users may consider to the beta value of 0.1 for short texts.
- `-niters <int>`: Specify the number of Gibbs sampling iterations. The default value is 2000.
- `-twords <int>`: Specify the number of the most probable topical words. The default value is 20 (people normally use top-10 or top-15 or top-20 topical words to evaluate topic coherence [7, 6]).
- `-name <String>`: Specify the name to the topic modeling experiment. The default value is *"model"*.
- `-sstep <int>`: Specify the step to save the sampling outputs. The default value is 0 (i.e. only saving the output from the last sample).

Examples:

```
$ java -jar jar/jLDADMM.jar -model LDA -corpus test/corpus.txt -name testLDA
```

The output files are saved in the same folder as the input training corpus file, in this case in the `test` folder. We have output files of `testLDA.theta`, `testLDA.phi`, `testLDA.topWords`, `testLDA.topicAssignments` and `testLDA.paras`, referring to the document-to-topic distributions, topic-to-word distributions, top topical words, topic assignments and model parameters, respectively. Similarly, we perform:

```
$ java -jar jar/jLDADMM.jar -model DMM -corpus test/corpus.txt -beta 0.1 -name
testDMM
```

We have output files of `testDMM.theta`, `testDMM.phi`, `testDMM.topWords`, `testDMM.topicAssignments` and `testDMM.paras`.

3. Using jLDADMM for evaluating topic models in document clustering task

This section is to evaluate topic models in document clustering task. Here, we treat each topic as a cluster, and we assign every document the topic with the highest probability given the document [4].

To get the Purity and NMI clustering scores, we perform:

```
$ java -jar jar/jLDADMM.jar -model Eval -label <Golden_label_file_path> -dir
<Directory_path> -prob <Document-topic-prob/Suffix>
```

- `-label`: Specify the path to the ground truth label file. Each line in this label file contains the golden label of the corresponding document in the input training corpus. See the `corpus.LABEL` and `corpus.txt` files in the `test` folder.
- `-dir`: Specify the path to the directory containing document-to-topic distribution files.
- `-prob`: Specify a document-to-topic distribution file or a group of document-to-topic distribution files in the specified directory.

Examples:

```
$ java -jar jar/jLDADMM.jar -model Eval -label test/corpus.LABEL -dir test -prob testLDA.theta
```

```
$ java -jar jar/jLDADMM.jar -model Eval -label test/corpus.LABEL -dir test -prob testDMM.theta
```

The above commands will produce the clustering scores for the `testLDA.theta` and `testDMM.theta` files in the `test` directory, separately. Given the following command:

```
$ java -jar jar/jLDADMM.jar -model Eval -label test/corpus.LABEL -dir test -prob theta
```

This command will produce the clustering scores for all the document-to-topic distribution files with their names ending by `theta`. In this case, the distribution files are `testLDA.theta` and `testDMM.theta`. It also provides the *mean* and *standard deviation* of the clustering scores.

To improve evaluation scores (for example, 5% absolute improvements in clustering and classification tasks), the users may consider to the latent feature topic models [8] with the source codes at https://github.com/datquocnguyen/LFTM

4. Citation

Please cite jLDADMM in all publications reporting on results obtained with the help of jLDADMM:

```
Dat Quoc Nguyen. 2015. jLDADMM: A Java package for the LDA and DMM topic models. http://jldadmm.sourceforge.net/. [.bib]
```

References

[1] Blei, D. M., 2012. Probabilistic Topic Models. Communications of the ACM 55 (4), 77–84.
[2] Blei, D. M., Ng, A. Y., Jordan, M. I., 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022.
[3] Griffiths, T. L., Steyvers, M., 2004. Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America 101 (Suppl 1), 5228–5235.
[4] Lu, Y., Mei, Q., Zhai, C., 2011. Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. Information Retrieval 14, 178–203.
[5] Manning, C. D., Raghavan, P., Schütze, H., 2008. Introduction to Information Retrieval. Cambridge University Press.
[6] Mimno, D., Wallach, H. M., Talley, E., Leenders, M., McCallum, A., 2011. Optimizing Semantic Coherence in Topic Models. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. pp. 262–272.
[7] Newman, D., Lau, J. H., Grieser, K., Baldwin, T., 2010. Automatic Evaluation of Topic Coherence. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 100–108.
[8] Nguyen, D. Q., Billingsley, R., Du, L., Johnson, M., 2015. Improving Topic Models with Latent Feature Word Representations. Transactions of the Association for Computational Linguistics 3, 299–313.
[9] Nigam, K., McCallum, A., Thrun, S., Mitchell, T., 2000. Text Classification from Labeled and Unlabeled Documents Using EM. Machine learning 39, 103–134.
[10] Yin, J., Wang, J., 2014. A Dirichlet Multinomial Mixture Model-based Approach for Short Text Clustering. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 233–242.