Data Representation Notes

Hang Dong

May 4, 2019

Abstract

This document contains the technical steps of using Latent Dirichlet Allocation (LDA) with MALLET[1] for Data Representation of social tags, as a supplementary material of the paper *Knowledge Base Enrichment by Relation Learning from Social Tagging Data*.

1 LDA training using MALLET

The input is a set of documents, each as a "bag of tags" originally in the Academic Social Bookmarking System Bibsonomy[2]. We cleaned the tags by grouping different tag variants and filtering low frequent or non-English tags. The final cleaned "bag of tags", with the compatible MALLET format, are in "bibsonomy_tags_cleaned_tags.txt".

MALLET command line commands to run LDA:

import file

```
bin\mallet import-file –input bibsonomy_tags_cleaned_tags.txt –output res-3-new.mallet –keep-sequence TRUE –token-regex "[\p{L}\p{N}\p{P}]+"
```

split file to training and testing/held-out data

```
bin\mallet split –input res-3-new.mallet –training-file train-new.mallet –testing-file test.mallet –training-portion 0.9
```

train topics

```
bin\mallet train-topics –input train-new.mallet –num-topics 50 –use-symmetric-alpha true –alpha 50 –beta 0.01 –evaluator-filename evaluator.mallet
```

evaluate topics

```
bin\mallet evaluate-topics –input test-new.mallet –evaluator evaluator.mallet –output-doc-probs doc-probs.txt –output-prob prob.txt
```

get length of documents

```
bin\mallet run cc.mallet.util.DocumentLengths –input test-new.mallet >test-new-doc-length.txt
```

train topics on the whole data

```
bin\mallet train-topics –input res-3-new.mallet –num-topics 1000 –num-threads 4 –use-symmetric-alpha true –alpha 50 –beta 0.01 –evaluator-filename evaluator.mallet –inferencer-filename inferencer.mallet –output-state state.gz –output-topic-keys twords.txt –output-doc-topics pzd.txt –topic-word-weights-file ptz.txt –word-topic-counts-file assign.txt
```

The input "bibsonomy_tags_cleaned_tags.txt" and one output "twords.txt" are in the same folder.

2 Data representation from the LDA outputs

Then we derive the $p(z)$, $p(z|C_a)$ and $p(C_a|z)$ from the outputs as in the Section 3.2 of the paper *Knowledge Base Enrichment by Relation Learning from Social Tagging Data.*

The $p(z)$ is calculated as

$$p(z) = \frac{N_z}{N} \tag{1}$$

,where N_z is the number of cleaned tags assigned using the topic z.

The $p(z|C_a)$, where C_a is a cleaned tag, is calculated as

$$p(z|C_a) \propto p(C_a|z) * p(z) \tag{2}$$

The tag vector representation $v(C_a)$ is thus

$$v(C_a) = \{p(\mathbf{z}_i|C_a)\}_{i=1}^{|\mathbf{z}|} \tag{3}$$

[1] http://mallet.cs.umass.edu/index.php

[2] Data version 2015-07 from https://www.kde.cs.uni-kassel.de/wp-content/uploads/bibsonomy/

We used Matlab to process the LDA outputs from MALLET. The final $p(z)$, $p(z|C_a)$ and $p(C_a|z)$ are in "pz.mat", "pzt.mat" and "ptz.mat" under the folder "Feature Generation, Hierarchy Generation, Relation-level evaluation" in the GitHub repository[3].

[3] `https://github.com/acadTags/tag-relation-learning`