CS410 MP4---Topic Models

Due: April 22 2017 at 11:59pm CT.

Topic models are among the most popular techniques for mining and analyzing topics in text data. Since they are unsupervised, they can be applied to any text data in any natural language to discover topics in text and facilitate their coverage in text documents. The basic topic models such as Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA) are often sufficient for many applications, though there are also many extensions of them. In this assignment, you will have an opportunity to experiment with LDA, learn how to apply it to some sample text datasets, and analyze its results to understand its strength and limitation. Specifically, you will use MeTA to run LDA on a few datasets and explore the results

1. **Background: Topic Models with LDA**

Recall that a topic here is a distribution of words with a common theme. LDA will take a text corpus as input, and the output will be the word distribution of topics, and the topic proportions of documents.

As an example of word distribution, a topic representing “working part-time during college” could have the following words (sorted by likelihood):

```
job: 0.156795
part: 0.127387
student: 0.12414
colleg: 0.0905726
time: 0.0708495
money: 0.064298
work: 0.0372687
studi: 0.0274538
learn: 0.0253973
earn: 0.015446
experi: 0.0141118
school: 0.00852317
parent: 0.00835476
societi: 0.00793623
```

Note that “working part-time during college” is never stated here; rather, it’s a label that a human assigns after looking at the distribution. Also, note the words have been stemmed (e.g. running and runner changed to run).

The topic proportions of documents may look like:

```
doc0
----
```

```
topic0: 0.1
topic1: 0.7
topic2: 0.2

doc1
----
topic0: 0.4
topic1: 0.4
topic2: 0.2

doc3
...
```

Naturally, each document topic proportion sums to one. We see that doc0 mainly uses topic1 words, and doc1 seems to be mainly about topic0 and topic1.

2. Setup

We will use metapy to explore LDA. Use the following commands to get started.

```
# Ensure your pip is up to date
pip install --upgrade pip

# install metapy
pip install metapy
```

If you have already installed metapy, use the following command to make sure it's up to date:

```
pip install --upgrade metapy
```

You need to get the datasets we will use in this assignment. Use the following command to create a directory, download and extract the datasets:

```
mkdir MP4
cd MP4
wget http://sifaka.cs.uiuc.edu/ir/textdatabook/cs410-mp4.tar.gz
tar --strip-components=1 -xzvf cs410-mp4.tar.gz
```

You also put the python files `run_lda.py, lda_reader.py` and `config.toml` in the same directory. The `run_lda.py` is the file you run topic model on a dataset, and `config.toml` is the configuration file. For example, if you want to run LDA on abstract dataset, you change the following lines in `config.toml`:

```
prefix = "."
dataset = "abstracts"
corpus = "line.toml"
index = "abstracts"
```

You also put `lemur-stopwords.txt` in the directory to get rid of the stop words. Assume you want 5 topics, then you run the following command

```
python run_lda.py config.toml lda_model 5
```

to run LDA and get two output files, `lda_model.phi` and `lda_model.theta`, which are word distributions of topics and topic proportions of documents, respectively. You can view and contents of these two files to get an idea of the output. To better read the output files, you can also run

```
python lda_reader.py config.toml lda_model
```

to analyze the output.

3. Tasks

1. (40 points) Run LDA on “newegg” dataset with topic numbers 3, 4 and 5. Report the starting and ending log-likelihood values for all three settings, and for the setting where topic number is 3, report the top 10 words of each topic with their probabilities using the format described later.

2. (60 points) Run LDA on “abstracts” dataset with topic number 6.

    2.1 (30 points) Report the top 20 words of each topic with their probabilities, and try to label the 6 topics you’ve got (use a few words as a label to describe what this topic is about). You don’t need to be 100% accurate when labeling the topic.

    2.2 (30 points) Report the topic proportions of the first 5 documents (doc_id from 0 to 4) and look at their contents. Do you think their topic proportions make sense? Why or why not?

3. (**Bonus** 30 points) Run LDA on NIPS 2000, 2006 and 2012 datasets with topic number 10. Give an idea of how you would use LDA to model and analyze the topical trends in NIPS over time.

4. Submission

Please include your answers in a text file called **CS410_mp4_yournetid.txt** and submit it via compass. You must **structure your file** using the example text file we have provided.

For Question 1, assume that your starting and ending likelihood values are -1 and -0.5 for all settings, your format would be like this:

```
Q1:-1 -0.5 -1 -0.5 -1 -0.5
Topic 0
w1 p1
```

```
w2 p2
…
w10 p10
Topic 1
w1 p1
w2 p2
…
w10 p10
Topic 2
…
```

Here w1 is the term and p1 is the corresponding probability.

For Question 2.1, you first report the top words like Q1, and append your labels after the words and probabilities. Your format is like:

```
Q2.1
Topic 0
w1 p1
…
w20 p20
Topic 1
…
Topic 5
…
Your label for Topic 0
Your label for Topic 1
…
Your label for Topic 5
```

For Question 2.2, you first give the proportions and append your comments.
Q2.2 (pi_j is the topic proportion, and brackets are required here)

```
Q2.2
doc_id 0
[pi_1 pi_2 … pi_6]
doc_id 1
[pi_1 pi_2 … pi_6]
…
doc_id 4
[pi_1 pi_2 … pi_6]
##Comment: I think the result does/doesn’t make sense because…
```

There is no specific format for Question 3. Just append your answer after a line “Q3”:

```
Q3
I think we can do it in this way…
```