Lucene Requirments for the Structural Topic Model

Molly Roberts and Brandon Stewart

August 10, 2012

Note: a lot of the below is already written in Lowe's JFreq software (LDA format, preprocessing stuff). It doesn't do all the cool subsetting, data organization that we do, though, so putting the two together is really important. It might be a good reference even if the code can't be used directly. http://www.williamlowe.net/software/jfreq/

1 Preprocessing

In general, in the software we probably want to include a preprocessing step that can be used after the subsetting step but before the term document matrix is outputted. We would have to figure out the details for this proprocessing, but typically such preprocessing would include options for:

1. Minimum word length
2. Removal of stopwords
3. Removing numbers
4. Removing XML/html tags
5. Lower threshold of word appearance (must appear in x% of documents)
6. Upper threshold of word appearance (must appear in less than y% of documents)
7. Remove currency

We might also be able to choose (more complicated to code, maybe longer term):

1. Unigrams, bigrams, trigrams
2. Stem words (with the ability to map them back into their most frequently used full form)

2 Term Document Matrix Format

For the Structural Topic Model, the term document matrix looks a little different than typical. It will be identical to the CTM in C, which is where the data is a file where each line is of the form:

$[\text{M}][\text{term}_1] : [\text{count}_1][\text{term}_2] : [\text{count}_2] \ldots [\text{term}_N] : [\text{count}_3]$

- $[M]$ is the number of unique terms in the document
- $[\text{term}_i]$ is an integer associated with the i-th term in the vocabulary.
- $[\text{count}_i]$ is how many times the i-th term appeared in the document.

3 Additional Output

3.1 Metadata

We already have an option to export metadata, and we would use this. Metadata would be in a csv file where each line corresponds in the same order to the documents in the TDM.

3.2 Vocabulary

We need an option that outputs the vocabulary associated with the numbers in the above term document matrix. This file should be a .dat file and should have one word per line ordered according to the numbering of words in the data.

3.3 Unique Covariate Profiles

Not sure if this fits into the Lucene software 100%, but it would be great to have the software output the number of unique covariate profiles and which documents are associated with which covariate profiles.