# Small Problem 4: HDP-LDA Topic Model

## Summary

**Given:**

A corpus of documents.

The hyper-parameters for the HDP-LDA model

A set of partial documents in which exactly half of the words have been removed.

**Find:**

Query 1: For each word not appearing in the partial document, compute the marginal probability that it appears in each of the partial documents.

Query 2: For each partial document, compute the MAP completion of the document.

**Metrics:**

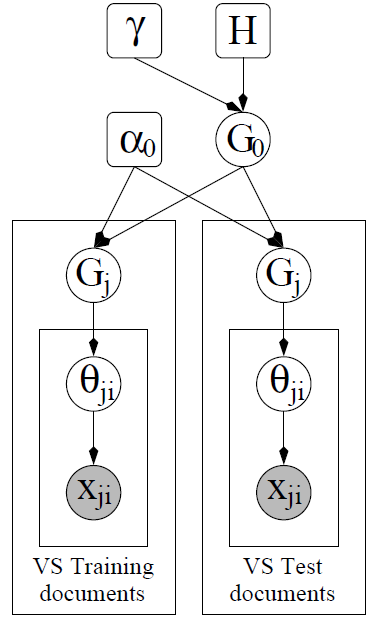
Metric 1: , which is the difference between the indicator variable for whether word appears in test document and the predicted probability that it appears in the document.

Metric 2: Hamming distance (computed at the token level) between the true document and the predicted document. For example, if the word “pickle” appears 3 times in the ground truth document and it is predicted to appear only once, then the Hamming distance is 2.

## Details

You are given a corpus of documents, a lexicon, and five incomplete documents in which half of the words have been deleted. You are to fit a Hierarchical Dirichlet Process topic model (as defined in Teh, Jordan, Beal & Blei, 2006; referred to as TJBB in this document) to the corpus and then, for each incomplete document and each word in the lexicon, you are to compute the probability that the word appears in the document.

The HDP model has the following form (known as model M1 in TJBB):



Here is the Dirichlet prior over topic-multinomials; is the concentration parameter of the top-level Dirichlet Process ; is the concentration parameter of the Dirichlet Process for each document . is the topic that generated word token of document , . We will fix the parameters as follows:

|  |  |  |
| --- | --- | --- |
| Parameter  (from TJBB) | Value | Explanation |
|  |  | Concentration parameter of the top-level DP |
|  |  | Concentration parameter of the per-document DPs |
|  |  | Concentration parameter of the Dirichlet distribution over words |
|  |  | The vocabulary size = number of words |

The following files are provided:

|  |  |
| --- | --- |
| *Name* | *Description* |
| problem-4-training-corpus.dat | This is a subset of the AP corpus containing 2241 documents. Each line consists of the number of words in the document followed by a blank-separated list of pairs where indexes the word (in vocab.txt) and is the number of occurrences of this word in the document |
| problem-4-test-corpus.dat | The five incomplete test documents |
| problem-4-test-ground-truth.dat | The complete test documents |
| problem-4-vocab.txt | The lexicon of 10473 words (in case you want to see the actual words) |

In these queries, we make the word vs. token distinction. A word (e.g., “pickle”) may appear 0 or more times in a document. Each occurrence is called a token. For query 1, we measure the error at the word level. In Query 2, we measure it at the token level. We will let denote the th test document; denote the number of words in the complete (ground truth) document, and be the number of tokens in the complete document.

Query 1: For each test document, for each word in the lexicon that has not already been observed in the document, compute the probability that that word appears in the document at least once. You may use the true lengths of the test documents (from problem-4-test-ground-truth.dat) for this purpose. Denote this probability as , where consists of all of the training documents, all of the incomplete test documents, and the length of the target test document .

Metric 1: , which is the absolute difference between the indicator variable for whether word appears in test document and the predicted probability that it appears in the document.

Query 2: For each test document , compute the most likely completion of the document of length , where is the true number of tokens in the document.

Metric 2: Hamming distance (computed at the token level) between the true document and the predicted document. For example, if the word “pickle” appears 3 times in the ground truth document and it is predicted to appear only once, then the Hamming distance is 2.

This query requires solving the very difficult optimization problem of searching for the MAP document completion.