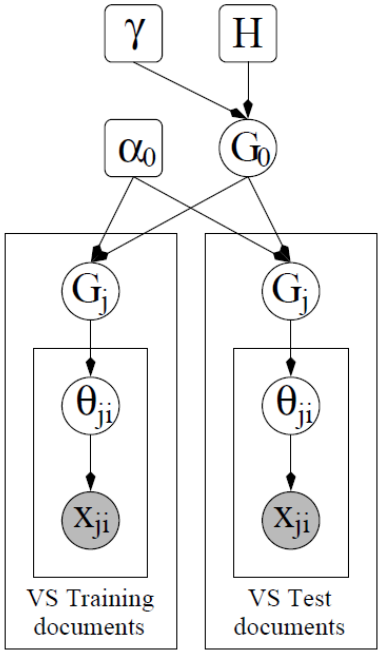
Small Problem 4: HDP Topic Model

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 H is the Dirichlet prior over topic-multinomials; γ is the concentration parameter of the top-level Dirichlet Process G_0; α_0 is the concentration parameter of the Dirichlet Process G_j for each document j. θ_{ji} is the topic that generated word token i of document j, x_{ji}. We will fix the parameters as follows:

Parameter (from TJBB)	Value	Explanation
γ	1	Concentration parameter of the top-level DP
α_0	1	Concentration parameter of the per-document DPs
η	0.01	Concentration parameter of the H Dirichlet distribution over words
W	10473	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 $i{:}\,n$ pairs where i indexes the word (in vocab.txt) and n 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 d_i denote the ith test document; ℓ_i denote the number of words in the complete (ground truth) document, and τ_i 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 $P(w|D, \ell_i)$, where D consists of all of the training documents, all of the incomplete test documents, and the length ℓ_i of the target test document i.

Metric 1: $\sum_w |\mathbb{I}[w \in d_i] - P(w|D, \ell_i)|$, which is the absolute difference between the indicator variable for whether word w appears in test document d_i and the predicted probability that it appears in the document.

Query 2: For each test document d_i, compute the most likely completion of the document of length τ_i, where τ_i 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.