Summary of "Hierarchical Topic Models and the Nested Chinese Restaurant Process"

Matt Dickenson mcd31

STA571/CS590.01 Due: 27 January, 2014

Blei, Griffiths, Jordan and Tenenbaum (2003) show how the Chinese restaurant process can be extended to hierarchical models of document topics (although the method itself is more general). The nested process adds another level of generalization to the model: there are an infinite number of Chinese restaurants in a city, and each table at each restaurant refers to another restaurant. The process results in a large number of trees connecting restaurants to one another. One disadvantage is that all topics must share a topic associated with the root restaurant of a tree, resulting in less flexibility in this one respect. However, in practice most corpora are likely to be selected due to some shared (though perhaps abstract) characteristic, making this restriction less of a convern. In many applications (including the final example in the paper) this root node will be trivial words that are often removed (such as articles and common prepositions).

One major advantage of this model is that it allows a single document to represent multiple topics; each word in a document is assigned to a given topic. Another benefit is the ability to express uncertainty over hierarchies, as mentioned above. The model also does not require an auxiliary list of words of words to ignore, as most language models do. One disadvantage originates from the use of a hierarchical LDA model: this can result in the Gibbs sampler getting stuck in local maxima. To avoid this, the sampler must be run a number of times from different starting points. Bayesian nonparametrics are useful here to address the model selection problem, allowing the nested CRP to indicate uncertainty about possible trees. The model's nonparametric nature also allows it to be used with data sets that are growing over time, meaning that it could be applied in near-online modeling scenarios.