Project Proposal

Dan Crankshaw, Dan Deutsch, Ryan Cotterell

February 22, 2013

The goal of our project is to create an infinite mixture model using a Direchlet process and Gibbs sampling to cluster Arabic social media data by dialect. This process will have several steps.

- We will begin with a literature review to understand the theoretical underpinnings of mixture models, infinite mixture models, Dirichlet processes, and Gibbs sampling. At the end of this we will create tutorial including sample code that will attempt to explain these concepts to someone with only an undergraduate computer science background.
- We will then implement an Infinite Mixture Model (IMM) that uses a Dirichlet process for the priors on the clusters and Gibbs sampling to perform inference. We will check the implementation against a known data set to verify its correctness.
- We will also implement a finite mixture model and derive a Gibbs sampler for this model as well. We will also test this implementation against a known data set.
- Finally, we will apply both the IMM and the FMM to a data set of short Arabic social media messages (SMS's, tweets, etc) to cluster these messages by dialect.