Patrick Verga
P3 Progress Report

My project is using clustering to improve pseudo-relevance feedback. This project is based on the paper “A cluster-based resampling method for pseudo-relevance feedback” by Lee, Kyung Soon, Croft, W. Bruce, and Allan, James. I didn’t realize it at the time, but it turns out that robust was the one dataset where the normal relevance model beat the clustering technique.

The paper takes the top 100 documents and uses k-nn clustering with tfidf vectors. The clusters are then scored and documents from the top clusters are used to extract terms for query expansion.

Results for the robust queries 600-700 were included in the paper and I was able to replicate very similar results (exact parameters for robust were not included in the paper). MAP from the paper and what I’ve gotten so far are shown below.

Original Paper Relevance Model	Original Paper cluster resampling	My Relevance Model	My Cluster Resampling
.359	.351	.355	.353

Based on this, I’d say things look pretty good. I don’t know if anything will come of it but I’m experimenting with LDA and a similar cluster scheme. After that, I should be able to just run this on the different data sets we have.