Final Project Proposal *

Chad Brubaker & Alex Clemmer

February 6, 2012

1 Overview

Machine learning in the domain of NLP has traditionally been an *offline* affair, with scientists injesting batches of data, analyzing them, and reporting the results.

Resistance to a fully-online methodology has so far been thanks mostly to the fact that NLP researchers have historically been ill-equipped to deal with the issues of massive data. For example, as [Talbot and Osborne, 2007] point out, simply storing (let alone learning from) the n-grams of a large corpus requires a precipitously large amount of space, both slowing down processing time, and making a large portion of streaming NLP work infeasible on common desktop machines.

A combination of recent advances in many critical areas, like feature representation [Van Durme & Lall, 2011] and locality sensitive hashing [Van Durme & Lall, 2009] makes interesting streaming work on commodity machines viable.

The goal of our project is to demonstrate that the requirements of moving to a fully online setting is not does not require a radical change in methodology, which we will show by adapting an offline classifer to a streaming setting. Specifically, we will train a model offline and extend it to predict well over time in a streaming context. We expect to use spectral Bloom filters to efficiently represent the n-gram features and their counts (see above citations). This will be the beating heart of the algorithm, although there may be other ancillary problems to solve, like feature hashing.

Our initial plan is to use the GeoDoc [GeoDoc, 2012] dataset to classify geospatial locations in text.

References

[Talbot and Osborne, 2007] David Talbot and Miles Osborne, "Randomised Language Modelling for Statistical Machine Translation," *Proceedings of ACL*, 2007.

[GeoDoc, 2012] Geospatial Information and Documents, http://www2.lirmm.fr/ mroche/GeoDoc2012/, *PAKDD Workshop*, 2012

[Van Durme & Lall, 2011] Benjamin Van Durme and Ashwin Lall, "Efficient Online Locality Sensitive Hashing via Reservoir Counting", *ACL Short*, 2011.

[Van Durme & Lall, 2009] Benjamin Van Durme and Ashwin Lall, "Probabilistic Counting with Randomized Storage", *IJCAI*, 2009.

*CS 6955 Data Mining; Spring 2012 Instructor: Jeff M. Phillips, University of Utah