CSCI 702 PRE-REGISTRATION SIGNUP

NAME: Will Potter

ID: 00472428

ADVISOR: David Kauchak

ADVISOR SIGNATURE:

**PROPOSED TOPIC**

TITLE: Geolocation through Language Recognition

ONE PARAGRAPH SUMMARY/DESCRIPTION:

For my senior thesis, I’d like to create a system that studies large datasets of content from social media sites (i.e. Facebook statuses and Tweets) and uses a combination of supervised and unsupervised machine learning techniques to determine the physical location of statuses that don’t contain an associated geographic location (“geotag”). Initial experiments would use a dataset of geotagged-posts within a constrained geographic area (i.e. Boston, New York).

Specific objectives are to:

* Predict geographic location without the use of the geotag
* Predict the geographic origin of the writer/speaker
* Track regular travel patterns off of text statuses

I’ve created a 220,000 tweet dataset (still growing as well) that contains only geotagged tweets. I’m also planning on downloading the complete set of Wikipedia articles that are tagged with GPS-coordinates through dbpedia.com. While these will be my initial two datasets as they are easy to download, I’d like to use any bodies of text that are tagged with a location.

Previous Work

Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. 2012. Supervised text-based geolocation using language models on an adaptive grid. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '12). Association for Computational Linguistics, Stroudsburg, PA, USA, 1500-1510.