Capstone Project Proposal

By Harold Neal

Project Definition

This project will use deep-learning NLP to extract topics and measure sentiment from a large set of news articles. The system will identify when sentiment on a topic has changed by a certain threshold amount over a specified period of time, and will graph sentiment for a given topic.

This tool could identify potential threats to a business brand or a political candidate by finding when opinion shifts against a topic. This tool may also be used to research how news sentiment maps to other measures. For example, the sentiment measures for the topic "Economy" could be plotted with a polling measure such as "consumer confidence". This may give insight into whether news sentiment leads or follows such measures. This could be used to determine if news sentiment could be used in prediction.

Although outside the scope of this project, the data generated could be used to group websites based on topic sentiment similarity.

While the project will capture data on many topics, I plan on focusing on:

1. Economy
2. Individual presidential candidates in the US
3. Climate Change

Project Data

The data for this project will come from the Common Crawl data set. I will filter the data to focus on several dozen recognized news, opinion, and analysis websites, such as the New York Times, CNN, The Economist, etc. The selection of sites for this project will have a focus on US news, but a few sources will come from non-US-based sites.

The project will start with the latest data available and move backwards. The goal would be to process one years' worth of data, and more if time and cost considerations allow.

Project Plan

This project will begin with determining efficient ways to process the very large crawl files that exist at the Common Crawl archive. This step will also include determining relevant content--I

will want to be able to extract title, article content, and date for each article, and I will determine if there are other relevant data that should be collected.

I will next need to identify the appropriate NLP libraries to use. Selection criteria for libraries to use will include processing costs and differentiation--the project needs for the sentiment measures to have a meaningful range. I may want to use more than one library to attempt to get a better overall measure of sentiment.

Cost Containment

Although it is difficult to estimate computing resources needed without some prototyping, we have several options in how the project could be resized to manage costs. These include:

- The number of sites that will be processed
- How many months of data will be used
- The number of sentiment libraries to use
- The computational costs of the libraries picked

It is expected that the size of data will be the most significant factor in determining costs.