

ENGR-E517 – INTRODUCTION TO HIGH PERFORMANCE COMPUTING

PROJECT PROPOSAL – HIGH PERFORMANCE TWITTER HASHTAG ANALYSIS WITH ANIMATED MAPPING

Peter Annable, Spring 2018

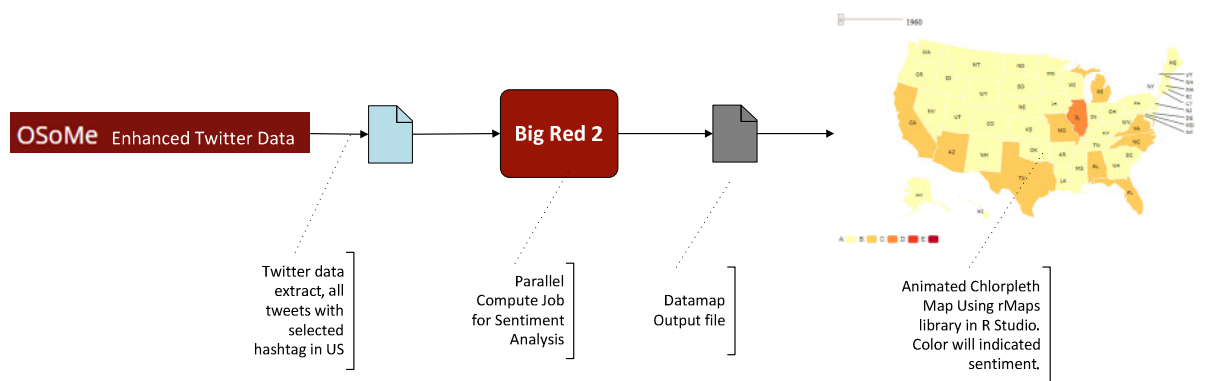
Overview

This project is inspired by the Pulse of the Nation, an analysis of tweets in the United States that was used to infer the mood of the people and use an animated map to show how the mood shifts over a period of 24 hours. In a similar manner, marketers are interested to understand how a hashtag and the sentiment about that hashtag moves over time. To understand this, I will use a set of tweets that contain a specified hashtag over a period of several days or weeks. Then I will determine the location, overall sentiment by state, and volume of tweets for each daily period. The measurements will be used to create an animated choropleth map of the United States

Solution Architecture

The solution envisioned will have three major components: Extracts from the IU OSoMe project will flow to a C code job in Big Red 2 for sentiment analysis over the input time period. The output will be a file that can be consumed by a visualization program. The C code will perform a simplified sentiment analysis based on keywords found with the tweet and averaged over the set for each time period and each state. As a simplification, only English language words will be analyzed.

Solution Architecture for Twitter Hashtag Analysis



Scaling Techniques

This problem will leverage strong scaling on a fixed set of tweets. The analysis steps should allow for a highly parallel operation, with one thread per time period, per state, or both. Additionally, I may be able to process some analysis steps in parallel.

Additional ideas to explored if time

If time allows, the following additional options will be explored:

- The option for hourly or by-minute analysis
- Animated cartograms to indicate the volume of tweets per each region.
- More sophisticated sentiment analysis.