Monitoring Sentiments on Twitter Using Parallelized NLP

Project Proposal

Frazier N. Baker

November 6, 2017

Introduction.

There is an immense amount of data available from online social media in the form of personal statements. Because these statements are often made publicly and given a timestamp, I can analyze the correlation between these statements and concurrent events[2]. In addition, I can filter this data down using other available data, such as location and hashtags. However, with the massive amount of data available, it would be advantageous to parallelize the filtering and analysis of data.

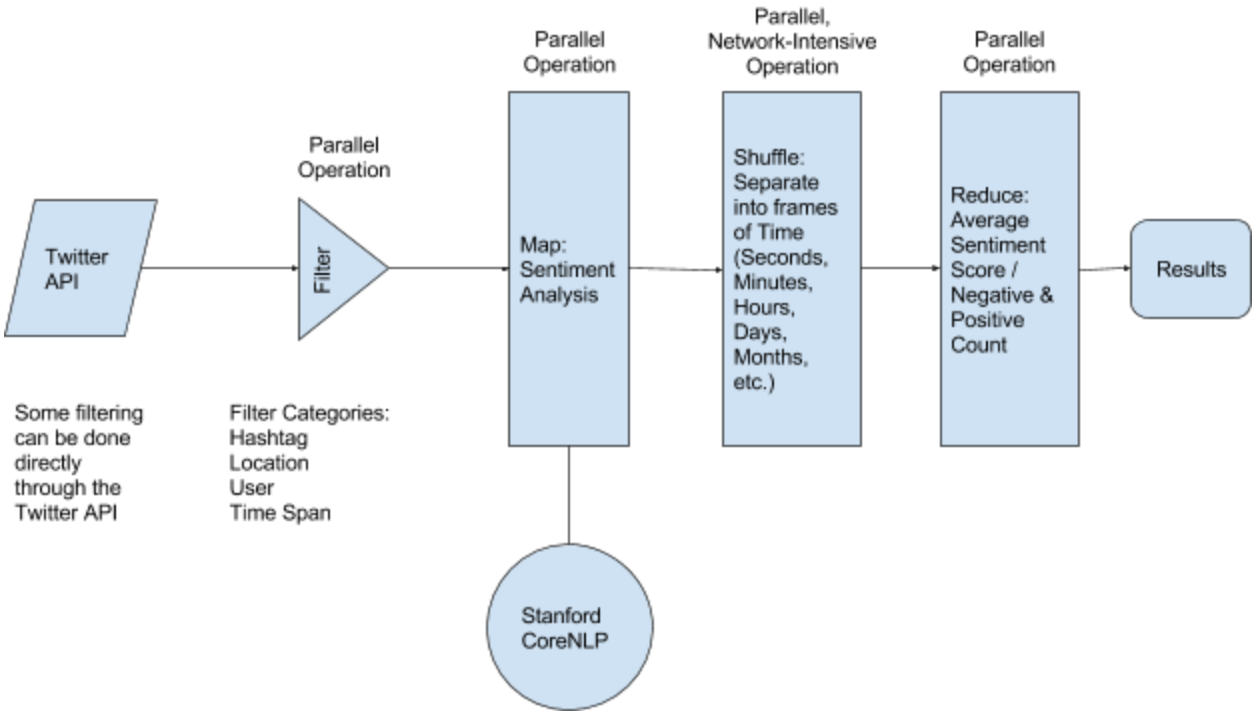
To mine information from these posts, I will utilize sentiment analysis, which has been called both a subfield of natural language processing and a special case of general natural language processing. Sentiment analysis often focuses on a simple positive/negative score or indication for the sentiment of the supplied input text, it does not delve into deeper emotions[3].

Objectives.

In this project, I propose a tool to run parallelized sentiment analysis on Twitter data over time filtered to specific regions and hashtags using a Spark[7] cluster running on Docker[5] containers spread across multiple hosts. I will be using the Stanford NLP Toolkit[4] for sentiment analysis[6]. I will deliver a command-line interface tool to run parallelized twitter sentiment analyses on my local cluster, along with a docker configuration and instructions to reproduce a similar cluster. I will also deliver the results of such an analysis on Twitter data from geographic areas surrounding colleges leading up to exam weeks of three different schools.

In addition, I may choose other events to examine with my analysis. I may temporarily make my cluster and software accessible over the internet for the purpose of demonstration for the class. In addition, I will optionally implement a simple graphical visualization of the results of this analysis over time to show how the overall sentiment of posts changes around dates, events, locations, and trending ideas. I may use D3.js[1] to help me create visualizations of this data.

Design.



Performance Goal.

Considering that I have 2 hosts with four cores on one host and two cores on the other, I would like to see at least a 4x speed up when running all nodes compared to running this analysis on one node. I take into consideration that there is some overhead with network communication, and I have tried to keep my connection distance short by using a local area network.

Schedule.

Task	**Due Date**
Set Up Cluster and Install Lightweight Linux OS	11/6/2017
Implement Docker Compose Spark Cluster Across Hosts	11/8/2017

Create Script to Download Twitter Data to HDFS and Download an Example Set	11/10/2017
Create Script to Run Sentiment Analysis	11/12/2017
Run Sentiment Analysis on Twitter Data For Exam Weeks and Analyze Results	11/15/2017
Benchmark Analysis On 1 Host versus Multiple Hosts	11/17/2017
Construct Visualizations & Presentation	11/19/2017

Bibliography.

[1] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D^3: Data-Driven Documents. *IEEE Trans. Vis. Comput. Graph.* 17, 12 (December 2011), 2301–2309.

[2] E. Gilbert, K. Karahalios - ICWSM, and 2010. 2010. Widespread Worry and the Stock Market. *aaai.org* (2010). Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1513/1833

[3] Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

[4] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. DOI:https://doi.org/10.3115/v1/p14-5010

[5] Dirk Merkel. 2014. Docker: Lightweight Linux Containers for Consistent Development and Deployment. *Linux J.* 2014, 239 (March 2014). Retrieved from http://dl.acm.org/citation.cfm?id=2600239.2600241

[6] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, 1631–1642.

[7] Matei Zaharia, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez, Scott Shenker, Ion Stoica, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, and Shivaram Venkataraman. 2016. Apache Spark: a unified engine for big data processing. *Commun. ACM* 59, 11 (October 2016), 56–65.