ACM Word Template for SIG Site

Egezon Berisha
RIT
exb3825@rit.edu

Thad Billig
RIT
tabdar@rit.edu

Eitan Romanoff
RIT
ear7631@rit.edu

ABSTRACT
Expression of ideas and sentiment over social networks has been increasingly popular over the last few years, leading to both a time where such information is easy to broadcast, as well as easy to acquire, notably through services like Facebook and Twitter. Likewise, there is an obvious desire for people to utilize this information within a variety of different domains, usually for the purpose of trend and sentiment analysis for marketing purposes. However, just as the information is freely available, it is also abundant, and performing analysis on datasets of exceedingly large sizes is an increasingly common issue. In this paper, we will discuss the methodology utilized to analyze large samples of twitter data on a chosen subject for sentiment and trend analysis.

General Terms
Algorithms, Measurement, Standardization, Languages.

Keywords
Amazon Web Services, Twitter, Mahout, Hadoop.

1. INTRODUCTION
Many businesses need to evaluate the outcome of social media efforts. To accomplish this there are big data acquisition and storage problems that need to be solved in addition to the actual natural language processing algorithms that need to be used in order to evaluate sentiment on a given topic. Distributed systems and distributed analytic tools are necessary to accomplish these tasks on acquired data.

This paper will demonstrate a solution built on top of several technologies that target the storage and handling of large datasets. Our project is an exploratory one that seeks to answer several key questions in this application of data mining – how can large sets of short fragmented natural language texts be stored and analyzed, and how can one meaningfully analyze these data sets in relevant ways to this domain of marketing?

Our processes in mining the data must be scalable over large datasets, and will use current technologies marketed towards this use case. Thus, Amazon Web Services (AWS), and its various services will be leveraged, including the AWS Simple Storage Service (AWS S3) and the AWS Elastic Compute Cloud (AWS EC2) service. The AWS EC2 platform has out-of-the-box functionality built on top of the Hadoop map-reduce framework. Furthermore, the Mahout data mining platform will be leveraged for all our natural language processing and mining. Because Mahout is based on the Hadoop system, it can also run on the EC2 cloud.

For data analysis itself, with regards to Twitter data, one can propose two core questions with regards to any Twitter dataset. Firstly, one may wish to know what the prevalent topics contained within the tweets are, and secondly, what the overall sentiment is towards those topics. The Mahout system contains several data mining and machine learning algorithms that can tackle these problems, including Latent Dirichlet Allocation – an algorithm that serves to identify key topics in text corpora, as well as a variety of standard classification techniques which can be applied to the classification of positive or negative sentiment.

The overall workflow process is as follows.

1. A topic of interest will first be targeted prior to data acquisition.

2. Twitter data will then be acquired by taking advantage of the Twython Python library, using a Python program, and pulling data from twitter VIA the streaming API. The streaming API that twitter publicly provides allows for a limit of 1% of the entire Twitter stream, and also allows for filters over that stream.

3. Streamed data will be stored as JSON in textual files, and uploaded to an AWS S3 bucket in fixed size chunks.

4. Data stored in an AWS S3 bucket can then be used by an AWS EC2 instance seamlessly. Various classification, and natural language processes will be executed at this stage.

The rest of the paper will be organized as follows: We will discuss the chosen natural language processing algorithms applied to the data, as well as their various purposes in section 2. In section 3, we will discuss the architecture of our system, and the dynamics of communication between its parts as well as the problems our method of storage solves. In section 4 we will present the results of our experimentation against identified twitter data sets. In section 5, we will present recommendations for additional work and unresolved issues.

RIT 13 Month 3, 2013, Rochester, NY, USA.