twitteRStorm: Prototyping a Streaming Framework for Analyzing Tweets with Storm

Douglas Raffle

1. Department of Statistics, West Virginia University

⋆ Contact author: dcraffle@mix.wvu.edu

Keywords: Twitter, Storm, Natural Language Processing, Streaming, R

Storm is a distributed, highly scalable, fault-tolerant framework for realtime analysis of streaming data (Apache Software Foundation, 2015a) used by dozens of organizations, like The Weather Channel, Spotify, and Twitter (Apache Software Foundation, 2015b). Storm works by specifying a streaming topology of “spouts” (data sources), tuples (data objects) and “bolts” (methods for processing tuples). While technologies like Hadoop and Spark process batches of static data, Storm is designed to continuously analyze incoming streams of data. While Spark does have a streaming API, Storm is purpose-built for streaming analysis.

The RStorm package (Kaptein, 2013) allows for rapid prototyping of Storm streams from within R, letting statisticians and data scientists quickly prototype streaming frameworks from the comfort of a familiar environment, which can then be passed to a data engineering team for production. This presentation will use the problem of analyzing tweets in real time as a case study. Tweets will be pulled from Twitter’s REST APIs (Twitter, 2015a) using the twitteR package (Gentry, 2015), a framework prototyped in RStorm, and a sample reporting dashboard created using shiny (Chang et al., 2015). We will also discuss implementing the stream in a production environment using Twitter’s Streaming APIs (Twitter, 2015b) and using the package Storm to write bolts in R using Storm’s Multi-Language Protocol (Day, 2015).

References


