Project Title:

Tweets Analysis with Spark

Team:

Code Monkey

Team members:

Harshil Patel (8), Matthew Boerner (1), Stephanie Retzke (10) and Yong Zheng (14)

Goal and objectives:

The goal and objectives for this project are for us to use the knowledge that we learned from class and explore what we can do with these knowledge. For this particular project, we will use Spark to analysis tweets from Twitter with Spark SQL, Spark ML, and Spark GraphFrames. We will begin with 100K tweets as our dataset.

Motivation:

Twitter has a lot of data from various people and topics. I want to use Twitter as my data source then use Spark to perform several different analysis with the tweets from Twitter.

Significance:

Spark is a popular unified analytics engine for big data processing. It consists of streaming, SQL, machine learning, and graph processing. I will use a subset of these 4 major modules for the analysis tasks. We will try to find some meaningful insights from the tweets we collect.

Features:

1. Use REST API to stream data from Twitter and persist data for future analysis
2. Use Spark SQL to perform additional data analysis tasks with the persisted data
3. Use Spark ML to apply some machine learning algorithms with the persisted data
4. Use Spark GraphFrames to build a network graph

References:

http://spark.apache.org/