Big Data Project
Collection of Twitter Data

In this as in all your data mining projects, you need to explain all steps and results clearly and cogently, so that a reasonably intelligent though statistically naïve manager could understand it.

In this project, you must apply all data preparation steps you deem necessary.

Project Description

You are being tasked of collecting a vast amount of information from Twitter.

The objective of this project is to collect Tweeter data through the Tweeter API and store it on the Hadoop platform. These data will be used for analytical purposes by a different team and they have not provided any information on how the data should be stored, the only constraint is they need the data stored in HDFS. However, they need to know the format in which the data has been stored so they can read it using Spark.

Tasks

1. Retrieve live data from Twitter. Twitter provides a developer API that allows the capturing of real-time Twitter conversations by querying the Twitter API using specific keywords. The API returns the tweets matching your keywords.
    a. Pick a very general keyword in order to capture a very diverse set of tweets.
    b. Query the Twitter API to retrieve the related tweets. Execute the same query every 30 minutes over a period of time in order to collect enough data to test your approach.
    c. You will need to create a Twitter developer account in order to activate your account to retrieve Twitter data.
    d. The Twitter API has limitation on the amount of twitters you can retrieve and how often you can query the API. However, these limitations are well within the limits needed for this project
    e. You can write this utility using the programming language of your choice.

2. The data collected in (1) should be:
    a. Stored on HDFS
    b. Available for querying using Hive or Impala

3. The data retrieved from Twitter will be in JSON format.
    a. Select the most appropriate format for storing in HDFS
    b. Try to define a well specific tabular structure
    c. Make the most appropriate decisions on the format type to use

4. Write MapReduce jobs that generate the following statistics:

a. Most frequent words
b. Most frequent bi-grams
c. Most active user

Your submission is due on May 1st and must include:

1. A brief report describing your approach and reasons to support your decisions. Moreover, report the insight on the question asked at (4).
2. All code used in this project (MapReduce, Twitter retrieval, etc.)
3. A sample of the data stored in HDFS.

Good Luck!