COMP-SCI 5540 Principles of Big Data Management

University of Missouri-Kansas City
Department of Computer Science and Electrical Engineering

Project 1

- Due Date: February 26, 2017 7:00:00 PM*
- Points:
    - Main Requirements: 100
    - Extra Requirement: 10
- Overview:
    - Develop a system to archive a social network's (e.g. Twitter's) data.
    - Each team should have exactly four members and only one submission per team.
    - Goal: become familiar with application program interfaces (APIs) and the Hadoop Distributed File System (HDFS).
- Submit:
    - Your properly documented code and a report that includes screenshots of your file system directory.
- Tasks†:
    - Main Requirements:
        - Collect tweets in JavaScript Object Notation (JSON) format (at least 100K record).
            - Find the list of top ten used hashtags in your collection.
        - Create a directory in HDFS for each hashtag from the top ten hashtag list.
            - Create additional two directories: "Others" and "None"
        - Store the tweets on files in HDFS
            - If a tweet contains a hashtag from the top ten list, store the tweet in that hashtag's directory.
            - If a tweet contains one or more hashtags, but none of the hashtags are in the top ten list, store the tweet in the "Others" directory.
            - If a tweet does not contain a hashtag, store it in the "None" directory.
    - Extra Requirement:
        - Implement a function that counts the number of times a `keyword` appears in one of two tweet JSON attributes (*text* and *hashtags*) in all of 12 directories that were created on HDFS:`int countWord(String keyword, String attr)`
- Prerequisite Skills:
    - Create, open, read, and write files using a local file system.
    - Write a basic word count function.
    - Read and parse a JSON file. Perform a word count on one attribute on a list of JSON objects.

* 10-point late submission penalty for every day starting on February 26, 2017 7:00:01 PM

† All requirements must be implemented in a program(s) and NOT via shell commands or scripts.

- Tips and Steps:
    - Create a developer account (on Twitter) and acquire the access tokens.
    - You will need to figure out how to add the necessary libraries (or JARs) to use some of the APIs (e.g. HDFS APIs).
    - Write a program to retrieve the data using Twitter’s public APIs and store the tweets in a local file.
    - Perform a word count on the hashtags, sort the counts in a non-increasing order, and create the 12 directories.
    - Read the local tweets file and store the tweets on HDSF according to the requirements.