COMP-SCI 5540 Principles of Big Data Management

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

Project 2

- Due Date: April 8, 2017 7:00:00 PM*
- Points:
  - Main Requirements: 100
  - Extra Requirement: 10
- Overview:
  - Develop a system to analyze Twitter's data using MapReduce.
  - Teams can have 1-5 members.
    - Teams with five members are required to implement the Extra Requirement and will not receive the additional points from this requirement.
  - Goal: become familiar with Hadoop MapReduce Framework.
- Submit:
  - Your properly documented code and a report that includes sample output for the requirement.
- Tasks†:
  - Note: storing data on HDFS is not required.
  - Main Requirements:
    - Using the collection of tweets from Project 1 (or collect a new set), implement MapReduce programs to determine the vocabulary uniqueness of your dataset:
      - M/R: Find the list of words that have duplicates in the tweets' text.
      - M/R: Find the list of words that are unique in the tweets' text.
      - Store the lists in two text files: dups.txt and uniqs.txt
      - Print the ratio of the number of unique words to the number of words with duplicates.
  - Extra Requirement:
    - Implement a MapReduce program to determine the best time to post a tweet.
      - Propose the metric/criterion of your choice based on the tweet JSON format.
      - Run your program and return the top ten best times to post a tweet on twitter.
- Prerequisite Skills:
  - Create, open, read, and write files using a local file system.
  - Write a basic word count function in MapReduce.
  - 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 April 8, 2017 7:00:01 PM

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