COMP-SCI 5540 Principles of Big Data Management

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

Project 3

- Due Date: May 1, 2017 7:00:00 PM
- Points:
  - Main Requirements: 100
  - Extra Requirement: 20
- Overview:
  - Develop a system to analyze Twitter's data using Apache Spark.
  - 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 Apache Spark Resilient Distributed Datasets (RDDs), Spark SQL and DataFrames (DFs).
- Submit:
  - Your properly documented code and project presentation slides.
- Tasks[*]:
  - Main Requirements:
    - Using a collection of tweets, implement three analytical tasks. A single task could consist of multiple analytical queries.
      - One task must be implemented via RDD transformations and actions only.
        - Must NOT be a simple word count (e.g. most used language).
      - The other two tasks must be implemented via Spark SQL and DFs.
      - One of your analytical queries must use the input file trends.txt.
  - Extra Requirement:
    - Implement a graphical user interface that enable the user to dynamically execute your analytical tasks and provide a visual representation of the results.
      - Flow: the user selects an analytical task, the task is executed in the back-end, the results are returned and displayed to the user in a visual representation (e.g. pie chart).
- Prerequisite Skills:
  - Create and transform an RDD.
  - Create an SQL query.
  - Write a basic word count using Spark RDDs.
  - Load a file into a DF and query it using Spark SQL.

---

[*] All requirements must be implemented in a program(s) and NOT via shell commands.

- Input file trends.txt layout:
  - One JSON object per line.
  - The example below only shows an object with two trending hashtags #LEmissionPolitique and #FillonStrasbourg. However, each object in the file contains 50 trending hashtags or topics.
  - JSON Object Example:

```json
{
  "timestamp": 1491511802.291998,
  "details": [
    {
      "created_at": "2017-04-06T20:48:18Z",
      "trends": [
        {
          "url": "http://twitter.com/search?q=%23LEmissionPolitique",
          "query": "%23LEmissionPolitique",
          "tweet_volume": 92636,
          "name": "#LEmissionPolitique",
          "promoted_content": null
        },
        {
          "url": "http://twitter.com/search?q=%23FillonStrasbourg",
          "query": "%23FillonStrasbourg",
          "tweet_volume": 16795,
          "name": "#FillonStrasbourg#FillonStrasbourg",
          "promoted_content": null
        }
      ],
      "as_of": "2017-04-06T20:49:56Z",
      "locations": [
        {
          "woeid": 1,
          "name": "Worldwide"
        }
      ]
    }
  ]
}
```