**CSEE5590/CS490: Big Data Programming**

**LAB ASSIGNMENT #2**

4/28/2020 - 11:59 PM  
Submit a lab report including screenshots and source code & data to GitHub. Post your GitHub URL through Lab 2 form at   
<https://forms.gle/7bJhJYWkPAhsV2cj9>

1. **Hadoop MapReduce Algorithm**
2. Implement MapReduce algorithm for finding Facebook common friends problem and run the MapReduce job on Apache Spark.
3. ***Use the following dataset to run your program:***

<https://umkc.box.com/s/w3kjvikyxm7nerkp75sdnqfvnj7rhs50>

1. Write a report including your algorithm and result screenshots.

**Finding Facebook common friends**: Facebook has a list of friends (note that friends are a bi-directional thing on Facebook. If I'm your friend, you're mine). They also have lots of disk space and they serve hundreds of millions of requests every day. They have decided to pre-compute calculations when they can to reduce the processing time of requests. One common processing request is the "You and Joe have 230 friends in common" feature. When you visit someone's profile, you see a list of friends that you have in common. We're going to use MapReduce so that we can calculate everyone's common friends once a day and store those results. Later on, it's just a quick lookup. We've got lots of disk, it's cheap.

**Example (What is the Key/Value Pair?)**

Assume the friends are stored as Person -> [List of Friends], our friends list is then:

A -> B C D

B -> A C D E

C -> A B D E

D -> A B C E

E -> B C D

The result after reduction is:

(A B) -> (C D)

(A C) -> (B D)

(A D) -> (B C)

(B C) -> (A D E)

(B D) -> (A C E)

(B E) -> (C D)

(C D) -> (A B E)

(C E) -> (B D)

(D E) -> (B C)

When D visits B's profile, we can quickly look up (B D) and see that they have three friends in common, (A C E).

Ref: <https://snap.stanford.edu/data/egonets-Facebook.html>

1. **Spark Data Frames**

**Datasets:**

1. **FIFA World Cup:**

<https://www.kaggle.com/abecklas/fifa-world-cup#WorldCupMatches.csv>

1. **Kickstarter Projects**

<https://www.kaggle.com/kemical/kickstarter-projects>

* 1. Create a Spark DataFrame using one of datasets and try to use all different StructType.
  2. Perform 10 intuitive questions in Dataset (e.g.: pattern recognition, topic discussion, most important terms, etc.). Use your innovation to think out of box.
  3. Perform any 5 queries in Spark RDD’s and Spark Data Frames. Compare the results.( for e.g. Selecting no of times Brazil won the World Cup , Selecting the Argentina World Cup statistics etc)
  4. Every query used should differentiate itself with other queries being used. A query cannot be same except change in the column and data.

1. **Spark Streaming Task**

**Perform Word-Count on Twitter Streaming Data using Spark.**

[**https://www.linkedin.com/pulse/apache-spark-streaming-twitter-python-laurent-weichberger/**](https://www.linkedin.com/pulse/apache-spark-streaming-twitter-python-laurent-weichberger/)

[**https://github.com/stefanobaghino/spark-twitter-stream-example**](https://github.com/stefanobaghino/spark-twitter-stream-example)

1. **Spark Graphx Task**

**Use one of the following datasets**

* + **Nashville-meetup Dataset**

[**https://www.kaggle.com/stkbailey/nashville-meetup**](https://www.kaggle.com/stkbailey/nashville-meetup)

* + **Word Game Dataset**

[**https://www.kaggle.com/anneloes/wordgame**](https://www.kaggle.com/anneloes/wordgame)

* + **Cyber Crime Motive**

[**https://www.kaggle.com/sunilkumarsv/indiacybercrimestats2013**](https://www.kaggle.com/sunilkumarsv/indiacybercrimestats2013)

**Perform the following tasks**

* 1. Perform Page Rank
  2. State importance of using graphx on the chosen dataset.

**Submission Guidelines (for all students):**

1. Submit your source code and documentation to GitHub and represent the work through wiki page properly (submit your screenshots as well. The screenshot should have both the code and the output)
2. Source code has to be prroperly commented. Coding standards must be ensured.
3. Submit a brief demo video 2-3 min showing your assignment with a voice over explaining your work through the Submission Link.
4. Use the following google link to submit your assignment

<https://forms.gle/7bJhJYWkPAhsV2cj9>

***Cheating, plagiarism, disruptive behavior and other forms of unacceptable conduct are subject to strong sanctions in accordance with university policy. See detailed description of university policy at the following URL:*** [*https://catalog.umkc.edu/special-notices/academic-honesty/*](https://catalog.umkc.edu/special-notices/academic-honesty/)