CSEE5590-0001/490-0003: Big Data Programming

Lesson Plan # 12

ICP Feedback and Submission Link: https://forms.gle/xMAmr3zATrtMG5cX7

Lesson Title: *Graph Frames and GraphX*

Lesson Description: *Distributed Collection of Data*

Lesson Overview:

- Graph frames
- GraphX vs Graph frames
- Pyspark and Scala environment setup
- Basic Commands on for creation of data frames
- Basic commands of graph frame algorithms
- Loading and saving data to file
- Implementations
- References

In Class Exercise

Dataset:
https://umkc.box.com/s/1drojp9ndqhlpee0gdvuvwuygk8phdyb

Graph Frames in Pyspark / Scala
Consider the datasets attached above

Part – 1:

1. Import the dataset as a csv file and create data frames directly on import than create graph out of the data frame created.
2. Concatenate chunks into list & convert to Data Frame
3. Remove duplicates
4. Name Columns
5. Output Data Frame
6. Create vertices
7. Show some vertices
8. Show some edges
9. Vertex in-Degree
10. Vertex out-Degree
11. Apply the motif findings.

12. Apply Stateful Queries.
13. Subgraphs with a condition.

Bonus

1.Vertex degree
2. What are the most common destinations in the dataset from location to location?
3. What is the station with the highest ratio of in degrees but fewest out degrees. As in, what station acts as almost a pure trip sink. A station where trips end at but rarely start from.
4.Save graphs generated to a file.

ICP Submission Guidelines:

1. ICP Submission is individual however, it can be completed as a Team during session.
2. If completed, should be presented to TA or Instructor before the completion of the class
3. Submission after the deadline is considered as late submission. (Check the late submission policy in the syllabus)
4. ICP Code with brief explanation should be pushed to GitHub.
5. Submit your screenshots as well to GitHub and documentation. The screenshot should have both the code and the output.
6. Submit a demo video 2-3 min showing your assignment with a voice over explaining your work if you are unable to complete ICP within the deadline due to genuine reason.
7. Provide the video submission link through the GitHub and submission form https://forms.gle/xMAmr3zATrtMG5cX7

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/*