**CSEE5590-0001/490-0003: Big Data Programming**

**Module2: Lesson Plan #5**

**ICP Feedback and Submission Link :** <https://forms.gle/wyFHFBL6LzM3xDGr5>

**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.

**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 Guidelines (In Class Students):**

1. ICP Submission is in pairs of two students.
2. Once completed, must be presented to TA or Instructor before the completion of the class
3. Submission after class is considered as late submission. (Check the late submission policy in the syllabus)
4. ICP Code with brief explanation should be pushed to GitHub. Submit Github Link through the Feedback Form - <https://forms.gle/wyFHFBL6LzM3xDGr5>

**Submission Guidelines (for online 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. Comment your code appropriately.
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

(ICP Submission Link#) : <https://forms.gle/wyFHFBL6LzM3xDGr5>

***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/)