MTU
Ollscoil Teicneolaíochta na Mumhan
Munster Technological University

A02 – REPORT ON ADDITIONAL SPARK LIBRARIES.

Besides the Spark libraries covered in this semester < Spark Core, Spark SQL and Spark Real-Time Libs (Spark Streaming & Spark Structured Streaming) > Spark has also:

- A library especifially devoted to Graph algorithms
    - Spark GraphX (for working with RDDs)
    - Spark GraphFrames (for working with DataFrames)
- A library especifially devoted to Machine Learning algorithms
    - Spark MLlib (for working with RDDs)
    - Spark ML (for working with DataFrames)

Write a report of up to 1,000 words where you present and discuss:

- A novel exercise to be included in the data analysis of the Dublin Bus dataset involving the Spark Graph and/or Machine Learning libraries.

There is no need to implement the new exercise, you just need to discuss it in terms of:

- Its originality - It has to be different from the 4 exercises proposed in Assignments 1 and 2.
- Its relevance - Include a potential use-case derived from the exercise you are proposing.
- Its viability:
    - Do not implement the exercise, but briefly discuss in natural language (English and/or psudocode) the main steps that would be needed so as to implement it.
    - Include in the discussion whether, if you had to implement it, you would choose to implement it using the library version for working with RDDs or DataFrames. Justify your selection.
    - Position the new exercise in terms of difficulty with respect to the other four exercises proposed in this assignment.