CORK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA CHORCAÍ

Computing@CIT

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.

Spark GraphFrames This is graphical representation of spark

We can use GraphFrames using the Vertex and Edge

- Vertex DataFrame: It has column ID and each vertex specify the unique ID
- Edge DataFrame: It has two elemenets 1. Src (Source) and 2. dst (Destination)

Let's assume you live in Dublin , and today is Saturday morning. Your boss call you in the morning and said that we have meeting at 2pm and you have to come to office. Now I need to go by bus line No 40 and now you have the destination of the bus. So you need to find out the timings of the busline

The following is the implementation of GraphFrames because

It provides uniform API for all languages such as Python,Java and Scala

It allows Powerful queries like Spark SQL

Fully supported DataFrame data sources

Set the vertices

```
path1 = pyspark.sql.Row(id=1, value=40)
path2 = pyspark.sql.Row(id=2, value=50)
path3 = pyspark.sql.Row(id=3, value=54)
```

Set the Edges

```
edge1 = pyspark.sql.Row(src=1, dst=4, value="10-40")
edge2 = pyspark.sql.Row(src=2, dst=8, value="20-80")
```

send this edges into dataframe and create the dataframe

edgesDF = spark.createDataFrame([edge1, edge1])

create dataframe by passing vertices and edges

myGF = graphframes.GraphFrame(my_verticesDF, my_edgesDF)

We revert back to Svertices
solSvertices = edgesDF.vertices

We display the content of **solSvertices**
solSvertices.show()

Now similar to Edges

7. Operation T2: We revert back to soledgesDataF
soledgesDataF = myGF.edges

8. Operation A2: We display the content of sol_edgesDF
soledgesDataF.show()