CORK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA CHORCAÍ

Computing@CIT

BIG DATA PROCESSING

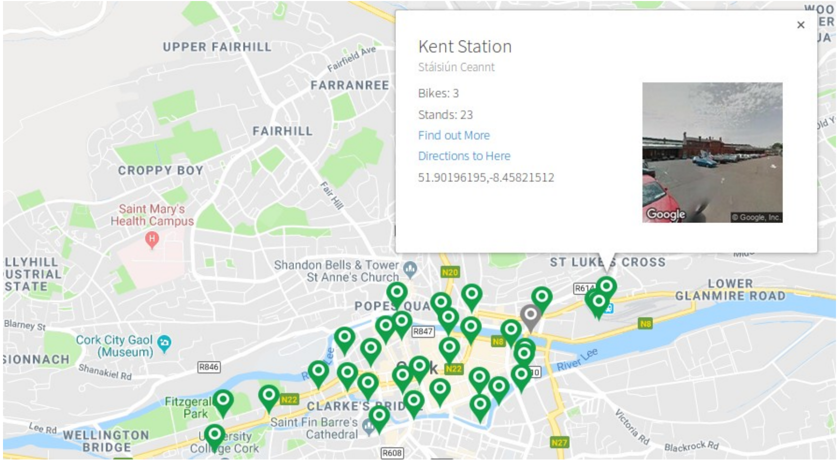
ASSIGNMENT 1: SPARK CORE & SPARK SQL

BACKGROUND.

Cork Smart Gateway is the home of open data for Cork: http://data.corkcity.ie/
At the moment it contains 8 datasets in open format. One of them is the "Coca-Cola Zero Bikes" (http://data.corkcity.ie/dataset/coca-cola-zero-bikes) which provides data related to the bike-sharing rental service supported by the city council.

While historic information on the state of the 31 bike stations in Cork city is not available, real-time information can be obtained via:

- The Coca-Cola Zero Bikes website: Visual Format
  https://www.bikeshare.ie/cork.html



Kent Station
Stáisiún Ceannt

Bikes: 3
Stands: 23
Find out More
Directions to Here
51.90196195,-8.45821512

- The Open Data API Endpoint: JSON Format.
  https://data.bikeshare.ie/dataapi/resources/station/data/list

{"schemeId":2, "schemeShortName":"Cork", "stationId":2032,
"name":"Kent Station", "nameIrish":"Stáisiún Ceannt", "docksCount":30,
"bikesAvailable":3, "docksAvailable":23, "status":0,
"latitude":51.902, "longitude": -8.458, "dateStatus":"15-10-2018 15:32:30" }

MY_DATASET FOLDER

My former colleague Michael O’Keefe collected data from mid January 2017 to late September 2017 by creating a service quering the API every 5 minutes from 6am to midnight and gathering all entries of a day into a file.

I have selected the files for the period 01/02/2017 – 31/08/2017 and provided the entries in the following csv format:
status ; name ; longitude ; latitude ; dateStatus ; bikesAvailable ; docksAvailable
For example, the aforementioned entry for Kent Station would be represented as follows:
0;Kent Station;-8.45821512;51.90196195;15-10-2018 15:32:30;3;23

In total, the dataset for the selected period contains 1,339,200 entries for 43,200 API requests over 200 days. It is provided in the folder “my_dataset” and has to be uploaded to the Databricks FileSystem (DBFS) via the databricks command line interface described in the lecture notes.

MY_CODE FOLDER

The assignment is divided into 4 parts:

- Part1: Introductory Data Analysis with Spark Core.
- Part2: Advanced Data Analysis with Spark Core.
- Part3: Introductory Data Analysis with Spark SQL.
- Part4: Advanced Data Analysis with Spark SQL.

The four parts are provided in the folder “my_code”.
Each part provides a Python file (A01_Part_Number.py) with 5 independent exercises: ex1, ex2, ex3, ex4, ex5.
Each exercise consists on a Python function to be completed.

MY_RESULT FOLDER

The correct results to be obtained from each exercise are provided in the folder “my_result”.

MARK BREAKDOWN.

The assignment is worth 50 marks. It consists on 4 parts with 5 exercises each: 20 exercises. All exercises are worth the same: 2.5 marks per exercise.

The submission of each exercise is optional (it is perfectly fine to not submit some exercises). Marks will be given to each submitted exercise in the following basis:

A. Exercise is correct and the code is efficient -> 100% of marks.
B. Exercise is correct but the code has some efficiencies -> 75% of marks.
C. Exercise is not correct due to a single error in the code -> 50% of marks.
D. Exercise is not correct due to multiple errors in the code -> 25% of marks.
E. Exercise has not been attempted or has been barely attempted -> 0% of marks.

SUBMISSION DETAILS.

Submission deadline: Sunday 17th of November, 11:59pm.

Please submit to Canvas (folder 5. Submissions) the following files:

- Part 1 -> A01_Part1.py
- Part 2 -> A01_Part2.py
- Part 3 -> A01_Part3.py
- Part 4 -> A01_Part4.py

Assignment Demo.

A demo for the assignment will take place from Week 10.
I will organise a brief individual interviews with each student to run the code and discuss about it (student is expected to explain the approach followed when tackling each exercise and explain the code being submitted).

The demo is mandatory for the assignment to be evaluated.

ASSIGNMENT 1 – PART 1 (SPARK CORE) (Week 5)

Use only RDD-based operations.

Exercise 1: Total amount of entries in the dataset.

Exercise 2: Number of Coca-cola bikes stations in Cork.

Exercise 3: List of Coca-Cola bike stations.

Exercise 4: Sort the bike stations by their longitude (East to West).

Exercise 5: Average number of bikes available at Kent Station.

ASSIGNMENT 1 – PART 2 (SPARK CORE) (Week 6)

Use only RDD-based operations.

Exercise 1: Number of times each station ran out of bikes (sorted decreasingly by station).

Exercise 2: Pick one busy day with plenty of ran outs -> Sunday 28th August 2017
Average amount of bikes per station and hour window
(e.g. [9am, 10am), [10am, 11am), etc.)

Exercise 3: Pick one busy day with plenty of ran outs -> Sunday 28th August 2017
Get the different ran-outs to attend.
Note: n consecutive measurements of a station being ran-out of bikes has to be considered a single ran-out, that should have been attended when the ran-out happened in the first time.

Exercise 4: Pick one busy day with plenty of ran outs -> Sunday 28th August 2017
Get the station with biggest number of bikes for each ran-out to be attended.

Exercise 5: Total number of bikes that are taken and given back per station
(Sort the results in decreasing order in the sum of bikes taken + bikes given back).

ASSIGNMENT 1 – PART 3 (SPARK SQL) (Week 7)

Use only DataFrame-based operations.

Exercise 1: Total amount of entries in the dataset.

Exercise 2: Number of Coca-cola bikes stations in Cork.

Exercise 3: List of Coca-Cola bike stations.

Exercise 4: Sort the bike stations by their longitude (East to West).

Exercise 5: Average number of bikes available at Kent Station.

ASSIGNMENT 1 – PART 4 (SPARK SQL) (Week 8)

Use only DataFrame-based operations.

Exercise 1: Number of times each station ran out of bikes (sorted decreasingly by station).

Exercise 2: Pick one busy day with plenty of ran outs -> Sunday 28th August 2017
Average amount of bikes per station and hour window
(e.g. [9am, 10am), [10am, 11am), etc.)

Exercise 3: Pick one busy day with plenty of ran outs -> Sunday 28th August 2017
Get the different ran-outs to attend.
Note: n consecutive measurements of a station being ran-out of bikes has to be considered a single ran-out, that should have been attended when the ran-out happened in the first time.

Exercise 4: Pick one busy day with plenty of ran outs -> Sunday 28th August 2017
Get the station with biggest number of bikes for each ran-out to be attended.

Exercise 5: Total number of bikes that are taken and given back per station
(Sort the results in decreasing order in the sum of bikes taken + bikes given back).