

Lab 2

[Click here to Register Attendance](https://forms.gle/pesShFDe1ibVfS3L7)

|  |  |
| --- | --- |
| Name | Andre McQuaid |
| Date | 21/2/2022 |
| Student No | S00216958 |
| Student Email | S00216958@mail.itsligo.ie |

### **Running a MapReduce Job on your local machine**

1. Go to the UCI Machine Learning Data Repository and explore the datasets available
2. Download a dataset of your choice and extract the files (I used the [Iris dataset](https://archive.ics.uci.edu/ml/datasets/iris))
3. Download the python [file here called MapReduceIris.py](https://github.com/marloftitsligo/ProgrammingForBigData/tree/main/MapReduceIris) and place in the same folder as your dataset
4. Open a Command Line (assuming you have Python installed)
5. Run the following command (replace iris.data with your data filename):  
   python MapReduceIris.py iris.data
6. You will likely get an error message as this python script uses a library called mrjob.   
   To install the library type:  
   pip install mrjob  
   You should get a confirmation message: Successfully installed mrjob-0.7.4
7. Try running the original command again (replace iris.data with your data filename):  
   python MapReduceIris.py iris.data
8. You should get the following in the output:   
   "setosa sepal width avg" 3.418
9. Post a screenshot of the output here:

|  |
| --- |
|  |

1. Try to adjust the Reducer calculation to find the average of other Species  
   Describe the rationales for your changes, your code and screenshot your output here:

|  |
| --- |
| I removed the species name check in the mapper function so that it would process all three species. I concatenated the species name to the “ sepal width” text in the mapper ‘yield’ line so that the key would be species-specific.  In the reducer function I concatenated the key and the “ avg” text so that there would be three named species-specific averages output at the end of the job. |

1. Find another dataset to use with this MapReduce code and make necessary changes to the code to perform an analysis of your choice.   
   Insert the name and a link to the dataset here and a short description of the analyses you performed:

|  |
| --- |
| Forest fires data set from UCI ML dataset: [Index of /ml/machine-learning-databases/forest-fires (uci.edu)](https://archive.ics.uci.edu/ml/machine-learning-databases/forest-fires/)   1. Determined the maximum temperature for fires in each month. August had the highest temperature when a fire occurred. 2. Determined the total number of fires on particular days. Most occurred on Sundays, the least occurred on Wednesdays. 3. Compared the average area burnt when it did not rain compared to when it did rain. The area burnt was 8 times greater when it did not rain.   The repo detailed in section 12 below contains three screenshots, one for each of the three analyses listed. |

1. Push the code and screenshots from step 11 above to Github and post the link here:

|  |
| --- |
| [andremcq/BigDataLab2 (github.com)](https://github.com/andremcq/BigDataLab2) (see the forestfires folder in this repo too)  i.e. [BigDataLab2/ForestFires at main · andremcq/BigDataLab2 (github.com)](https://github.com/andremcq/BigDataLab2/tree/main/ForestFires)  The main folder [andremcq/BigDataLab2 (github.com)](https://github.com/andremcq/BigDataLab2) contains the code and data for questions 1 through 10.  The ForestFires subfolder [BigDataLab2/ForestFires at main · andremcq/BigDataLab2 (github.com)](https://github.com/andremcq/BigDataLab2/tree/main/ForestFires) contains the code, data and screenshots for questions 11 through 12. |