
# Lab 2

Click here to Register Attendance

| | |
|---|---|
| Name | Mick Dobbs |
| Date | 18th February 2022 |
| Student No | S00217208 |
| Student Email | S00217208@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)
3. Download the python file here called MapReduceIris.py 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:

```
(base) C:\Users\m033\MSc\Big_Data>python MapReduceIris.py iris.data
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory C:\Users\m033\AppData\Local\Temp\MapReduceIris.m033.20220221.201850.748788
Running step 1 of 1...
job output is in C:\Users\m033\AppData\Local\Temp\MapReduceIris.m033.20220221.201850.748788\output
Streaming final output from C:\Users\m033\AppData\Local\Temp\MapReduceIris.m033.20220221.201850.748788\output...
"setosa sepal width avg"        3.418
Removing temp directory C:\Users\m033\AppData\Local\Temp\MapReduceIris.m033.20220221.201850.748788...
```

10. 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:

```
(base) C:\Users\m033\MSc\Big_Data>python MapReduceIris.py iris.data
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory C:\Users\m033\AppData\Local\Temp\MapReduceIris.m033.20220221.212654.832722
Running step 1 of 1...
job output is in C:\Users\m033\AppData\Local\Temp\MapReduceIris.m033.20220221.212654.832722\output
Streaming final output from C:\Users\m033\AppData\Local\Temp\MapReduceIris.m033.20220221.212654.832722\output...
"virginica sepal width avg"     2.9739999999999998
Removing temp directory C:\Users\m033\AppData\Local\Temp\MapReduceIris.m033.20220221.212654.832722...
```

I made simple changes to the code to look for a different regular expression pattern. The new code is in MapReduceIris-1.py

I created a new function called mapper_get_sepW_species. This takes in a variable which can be set to one of the 3 species of iris. The new function evaluates whichever one of the species is searched for and gives the appropriate output.

To make this work, I needed to just have the mapper value changed to call the new function.

The original code is still in MapReduceIris.py

11. 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:

I chose the Adult Data Set Index of /ml/machine-learning-databases/adult (uci.edu). This was used for a prediction task is to determine whether a person makes over 50K a year but I am going to use it to find the average age of people with the occupation “Exec-managerial”.

The new code is in MapReduceIris-2.py. I had to ensure I got the first column of data which is where the age is held. I have included the adult.name file as reference. I changed the search to data[0] from data[1] in the original code. The rest of it was simply changing the values being looked for and the output.

```
(base) C:\Users\m033\MSc\Big_Data>python MapReduceIris-2.py adult.data
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory C:\Users\m033\AppData\Local\Temp\MapReduceIris-2.m033.20220221.223724.586130
Running step 1 of 1...
job output is in C:\Users\m033\AppData\Local\Temp\MapReduceIris-2.m033.20220221.223724.586130\output
Streaming final output from C:\Users\m033\AppData\Local\Temp\MapReduceIris-2.m033.20220221.223724.586130\output...
"Exec-managerial average age is"        42.16920806689621
Removing temp directory C:\Users\m033\AppData\Local\Temp\MapReduceIris-2.m033.20220221.223724.586130...
```

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

| |
|---|
| MickDobbsKildavin2/secondrepo: For Programming for Big Data Lab 1 (github.com) |