

Lab 3

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

|  |  |
| --- | --- |
| Name | Rebecca Cole |
| Date | 25/02/2021 |
| Student No | S00205102 |
| Student Email | S00205102@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 found the average of sepal width for the species Versicolor. I changed the species from any reference from Setosa to Versicolor however used the same column of data to average as I was still finding the average of the sepal width.  Output:    Code:  from mrjob.job import MRJob  from mrjob.step import MRStep  import re  DATA\_RE = re.compile(r"[\w.-]+")  class MRProb2\_3(MRJob):  def steps(self):  return [  MRStep(mapper=self.mapper\_get\_sepW\_versicolor,  reducer=self.reducer\_get\_avg)  ]  def mapper\_get\_sepW\_versicolor(self, \_, line):  # yield each petal width  data = DATA\_RE.findall(line)  if "Iris-versicolor" in data:  sep\_W = float(data[1])  yield ("sepal width", sep\_W)  def reducer\_get\_avg(self, key, values):  # get max of the petal widths  size, total = 0, 0  for val in values:  size += 1  total += val  yield ("versicolor sepal width avg", round(total,1) / size)  if \_\_name\_\_ == '\_\_main\_\_':  MRProb2\_3.run() |

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:

|  |
| --- |
| Birmingham Parking Dataset - <https://archive.ics.uci.edu/ml/datasets/Parking+Birmingham>  The Birmingham Parking Dataset is a simple beginner level dataset with 4 columns and 35718 rows. The dataset contains information about the number of cars parked in 30 parking areas around Birmingham at different times of the day between October to December 2016.  I found the average occupancy of carpark BHMBCCMKT01 between October to December 2016. |

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

|  |
| --- |
|  |