

Assignment 1

January 28, 2022

1 Warmup Assignment

In this exercise you will just submit a ping to the grader in order to make sure the exercise and grading environment is setup correctly.

We have to install a little library in order to submit to coursera

```
[1]: !rm -f rklib.py
      !wget https://raw.githubusercontent.com/IBM/coursera/master/rklib.py
```

```
--2022-01-28 04:26:19--
https://raw.githubusercontent.com/IBM/coursera/master/rklib.py
Resolving raw.githubusercontent.com (raw.githubusercontent.com)...
185.199.110.133, 185.199.108.133, 185.199.109.133, ...
Connecting to raw.githubusercontent.com
(raw.githubusercontent.com)|185.199.110.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 2540 (2.5K) [text/plain]
Saving to: 'rklib.py'

rklib.py          100%[=====>]    2.48K  --.-KB/s    in 0s

2022-01-28 04:26:19 (18.9 MB/s) - 'rklib.py' saved [2540/2540]
```

Please provide your email address and obtain a submission token on the grader's submission page in coursera, then execute the cell

```
[2]: from rklib import submit
      import json

      key = "RrIb4SHNEeiLcw7AkKxwA"
      part = "zetaj"
      email = "f20191315@hyderabad.bits-pilani.ac.in"
      token = "04JmZ1r3JRAqhYmYA"

      submit(email, token, key, part, [part], json.dumps(23))
```

Submission successful, please check on the coursera grader page for the status

```
{"elements":[{"itemId":"MV2rL","id":"sUpST4RAEeawAApvKZgcCQ~MV2rL~lrb7P3_yEeyAAg  
oQFU4ePQ","courseId":"sUpST4RAEeawAApvKZgcCQ"}],"paging":{"linked":{}}
```

Assignment 1

January 28, 2022

1 Warmup Assignment 2

Below you see some ApacheSpark code written in Python. You don't have to change code now, the only thing we want you to do is to make sure that you have a proper Apache Spark Notebook environment available for this course

This notebook is designed to run in a IBM Watson Studio default runtime (NOT the Watson Studio Apache Spark Runtime as the default runtime with 1 vCPU is free of charge). Therefore, we install Apache Spark in local mode for test purposes only. Please don't use it in production.

In case you are facing issues, please read the following two documents first:

<https://github.com/IBM/skillsnetwork/wiki/Environment-Setup>

<https://github.com/IBM/skillsnetwork/wiki/FAQ>

Then, please feel free to ask:

<https://coursera.org/learn/machine-learning-big-data-apache-spark/discussions/all>

Please make sure to follow the guidelines before asking a question:

<https://github.com/IBM/skillsnetwork/wiki/FAQ#im-feeling-lost-and-confused-please-help-me>

If running outside Watson Studio, this should work as well. In case you are running in an Apache Spark context outside Watson Studio, please remove the Apache Spark setup in the first notebook cells.

```
[1]: from IPython.display import Markdown, display
def printmd(string):
    display(Markdown('# <span style="color:red">' + string + '</span>'))

if ('sc' in locals() or 'sc' in globals()):
    printmd('<<<<<!!!! It seems that you are running in a IBM Watson Studio_
    ↳Apache Spark Notebook. Please run it in an IBM Watson Studio Default Runtime_
    ↳(without Apache Spark) !!!!!>>>>>')
```

2 «««<!!!! It seems that you are running in a IBM Watson Studio Apache Spark Notebook. Please run it in an IBM Watson Studio Default Runtime (without Apache Spark) !!!!»»»>

```
[2]: !pip install pyspark==2.4.5
```

Collecting pyspark==2.4.5



Downloading pyspark-2.4.5.tar.gz (217.8 MB)



[illegible]



```
76.4 MB 22.5 MB/s eta 0:00:07 i£;i£;äÜLaÜLäÜLäÜLäÜLäÜLäÜLäÜLäÜLäÜL 216.7 MB
→93.0 MB/s eta 0:00:01
```



Collecting py4j==0.10.7


Downloading py4j-0.10.7-py2.py3-none-any.whl (197 kB)




























































































































































































































Building wheels for collected packages: pyspark

```
Building wheel for pyspark (setup.py) ... done
```

```
Created wheel for pyspark: filename=pyspark-2.4.5-py2.py3-none-any.whl
```

```
size=218257928
```

```
sha256=9fc9e83e10fb78552c6b46a5ebb7692fb2ae1f2fd4b4d48323b4ca6003dbe0a3
```

Stored in directory: /home/spark/shared/.cache/pip/wheels/40/1b/56/aa5b76fa0c55166784b69226bcf72e3480d08d248a16f0b15c

Successfully built pyspark

```
Installing collected packages: py4j, pyspark
```

```
Attempting uninstall: py4j
```

```
Found existing installation: py4j 0.10.9
```

Uninstalling py4j-0.10.9:

```
ERROR: Could not install packages due to an OSError: [Errno 13] Permission
denied: 'WHEEL'
```

Consider using the `--user` option or check the permissions.

```
[3]: try:
      from pyspark import SparkContext, SparkConf
      from pyspark.sql import SparkSession
    except ImportError as e:
      printmd('<<<<!!!! Please restart your kernel after installing Apache Spark_
      <!!!!>>>>')
```

```
[4]: sc = SparkContext.getOrCreate(SparkConf().setMaster("local[*]"))
```

```
spark = SparkSession \
    .builder \
    .getOrCreate()
```

```
[5]: def assignment1(sc):
      rdd = sc.parallelize(list(range(100)))
      return rdd.count()
```

```
[6]: print(assignment1(sc))
```

100

```
[7]: !rm -f rklib.py
      !wget https://raw.githubusercontent.com/IBM/coursera/master/rklib.py
```

--2022-01-28 04:49:12--

https://raw.githubusercontent.com/IBM/coursera/master/rklib.py
Resolving raw.githubusercontent.com (raw.githubusercontent.com)...
185.199.111.133, 185.199.109.133, 185.199.108.133, ...
Connecting to raw.githubusercontent.com
(raw.githubusercontent.com)|185.199.111.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 2540 (2.5K) [text/plain]
Saving to: 'rklib.py'

rklib.py 100%[=====>] 2.48K --.-KB/s in 0s

2022-01-28 04:49:12 (50.4 MB/s) - 'rklib.py' saved [2540/2540]

Please provide your email address and obtain a submission token on the grader's submission page in coursera, then execute the cell

```
[8]: from rklib import submit
      import json

      key = "R1eDmiHNEei9kxIYdin0mA"
      part = "fnFg7"
      email = "f20191315@hyderabad.bits-pilani.ac.in"
      token = "ioD9fEv18gqCXsVU"

      submit(email, token, key, part, [part], json.dumps(assignment1(sc)))
```

Submission successful, please check on the coursera grader page for the status

```
-----
{"elements": [{"itemId": "D277m", "id": "sUpST4RAEeawAApvKZgcCQ~D277m~pAqDQH_1EeyeFxLRn7pabw", "courseId": "sUpST4RAEeawAApvKZgcCQ"}], "paging": {}, "linked": {}}
```

