# Exercise 6

*Get started with Apache Spark and Python*

**Prior Knowledge**
Unix Command Line Shell
Simple Python

**Learning Objectives**
Understand the Spark system
Understand the Jupyter Notebook model
WordCount!

**Software Requirements**
(see separate document for installation of these)

- Apache Spark 3.0.0
- Python 3.8.x
- Jupyter notebooks

**Part A. Spark Python Shell (pySpark)**

1. We are going to do a wordcount against a set of books downloaded from Project Gutenberg. Wordcount is the definitive Big Data program (sort of `Hello World` for Big Data) and it is frankly embarrassing that we haven't done one yet.

2. Apache Spark has a useful Python shell, which we can use to interactively test and run code.

3. Let's make a directory for our code:

```
mkdir ~/pse
cd ~/pse
```

4. Now start the Spark Python command line tool –

```
pyspark
```

5. In the command-line you will see something like

```
[I 13:53:23.865 NotebookApp] Serving notebooks from local directory: /home/oxclo/pse
[I 13:53:23.866 NotebookApp] 0 active kernels
[I 13:53:23.866 NotebookApp] The Jupyter Notebook is running at:
http://localhost:8888/?token=fd655aab32ed4840ceb47b8b7392b1243a27f56350888a91
[I 13:53:23.866 NotebookApp] Use Control-C to stop this server and shut down all
kernels (twice to skip confirmation).
[C 13:53:23.868 NotebookApp]

    Copy/paste this URL into your browser when you connect for the first time,
    to login with a token:
        http://localhost:8888/?token=fd655aab32ed4840ceb47b8b7392b1243a27f56350888a91
```

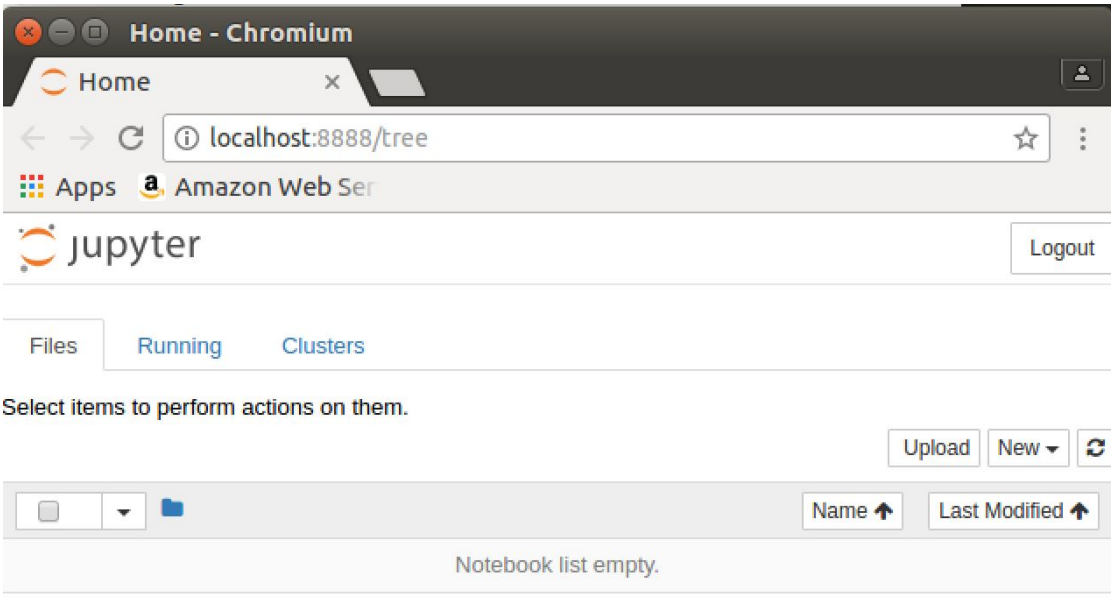


6. Jupyter is starting. This is because I have **already** preconfigured the following environment variables in `~/.bashrc`

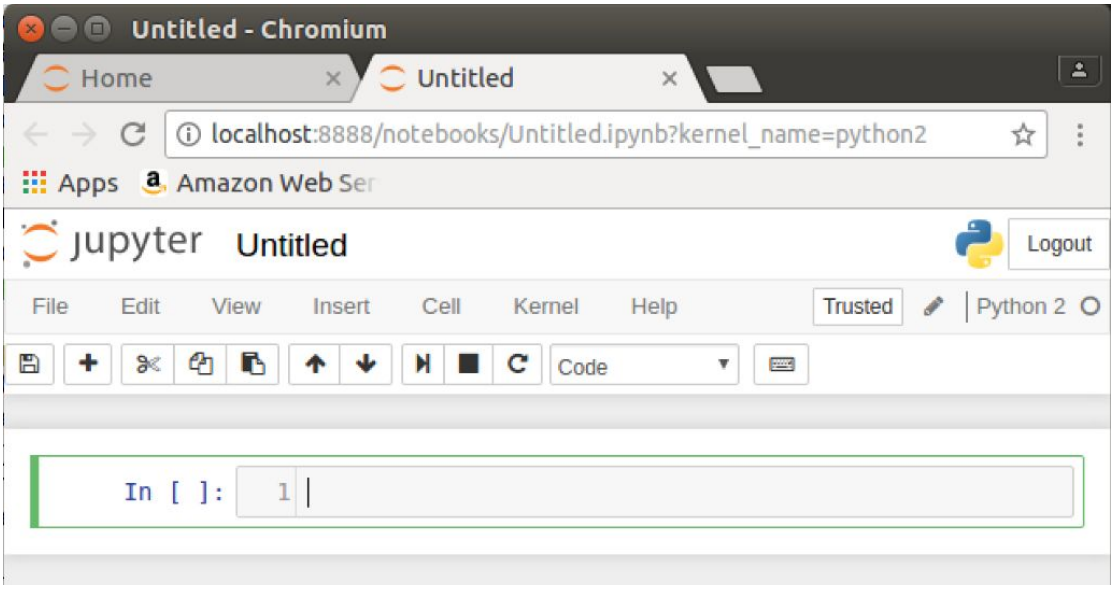
```
export PYSPARK_DRIVER_PYTHON=jupyter
export PYSPARK_DRIVER_PYTHON_OPTS='notebook'
```

*You don't need to do this!*

7. And then a browser window will pop up.

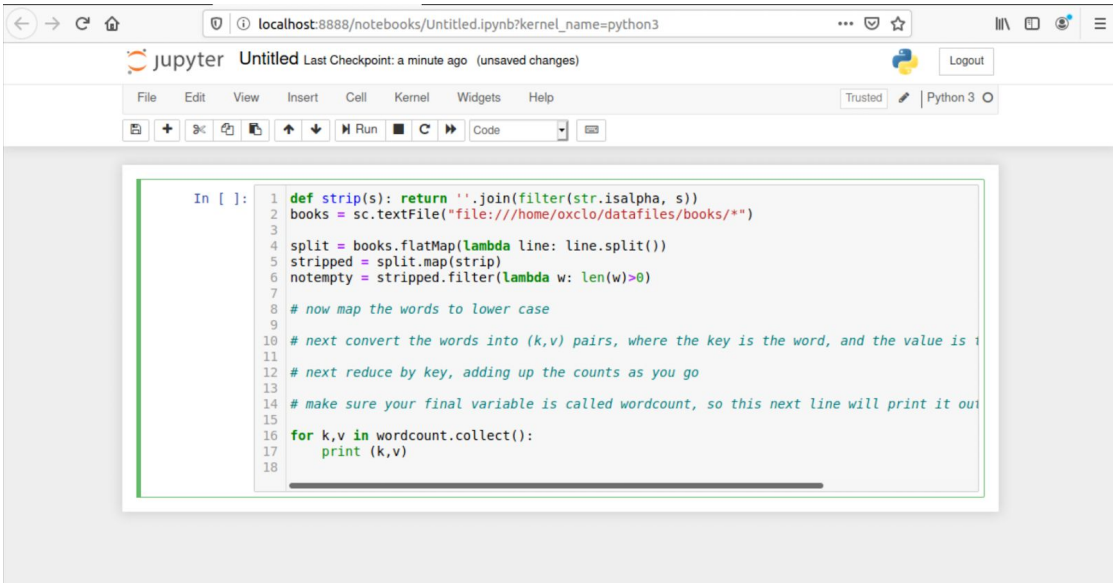



8. Use the **New** button to create a new Python3 notebook:




9. There is a starter of the code you need in the following URL: https://freo.me/first-notebook

   Paste that into the cell [1] so it looks like this:




10. There are some aspects that are not filled in that you need to write. Basically this is a data-processing pipeline (also a directed acyclic graph)

11. *Let's look at the parts that are there already.*

12. We already have a SparkContext object defined in the notebook (in a program you need to define one, which we will see later)

13. We want to remove any non-alphanumeric characters. This is a function that will do that:

```
def strip(s): return ''.join(filter(str.isalpha, s))
```

14. With the preliminaries over, the next line loads the data in:

```
books =
sc.textFile("file:///home/oxclo/datafiles/books/*")
```

15. Then splits the lines into separate words

```
split = books.flatMap(lambda line: line.split())
```

16. Removes non-alpha characters:

```
stripped = split.map(strip)
```

and removes empty items:

```
notempty = stripped.filter(lambda w: len(w)>0)
```

17. Now it is time for you to do something!

Convert all the words to lower case, using a map operation. In python, if



*str* is a string, then `str.lower()` is the same string in lower case.

18. Now you need to get ready for a reduce. In order to do a reduce, we need some form of *key, value* pairs. I recommend using *tuples* which are simply (k,v) in Python (the brackets group the items into a tuple).

19. Remembering how reduce works, we need each word to have a count. Before reducing, that count is 1. So we need a lambda that takes a word w and returns (w,1)

20. Now we can do a reduce that adds all those counts together.



21. Finally, we need to collect the results and print them. In Spark, they may be distributed across different RDD partitions on different machines, so the collect() method brings them together.

```
for k,v in wordcount.collect():
   print (k,v)
```

22. Try running the cell, by clicking 

23. Be patient. I suggest you look at the command window and wait until you see spark start working.

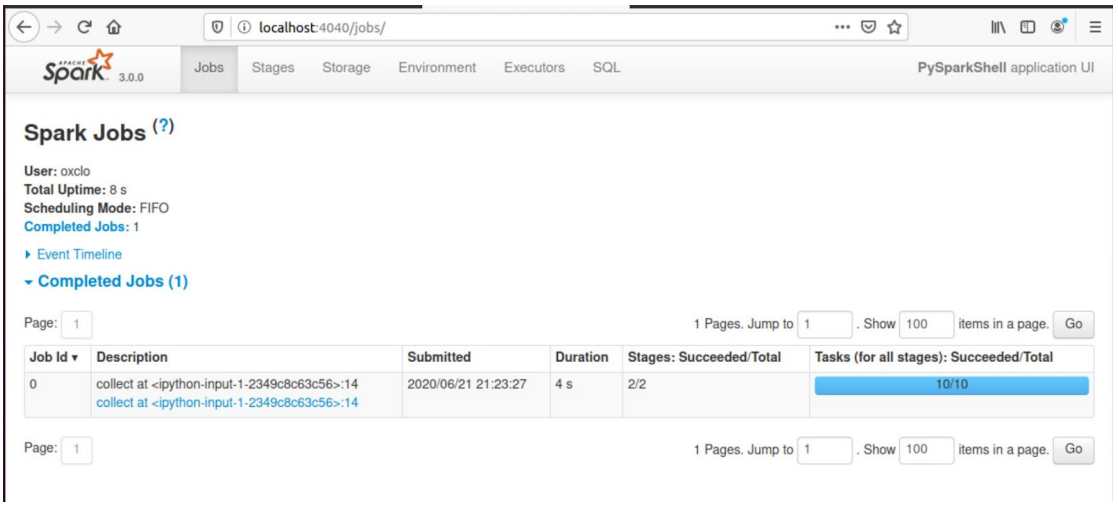
24. You should see a word count appear below cell 1:

```
systematic 7
parallelogram 1
sowell 1
presnya 1
four 265
conjuring 1
chamberagain 1
marching 32
sevens 4
awistocwacy 1
trotat 1
canes 1
shipmets 1
understandthat 2
lorn 16
lore 1
inwards 2
wickam 62
utterand 1
```

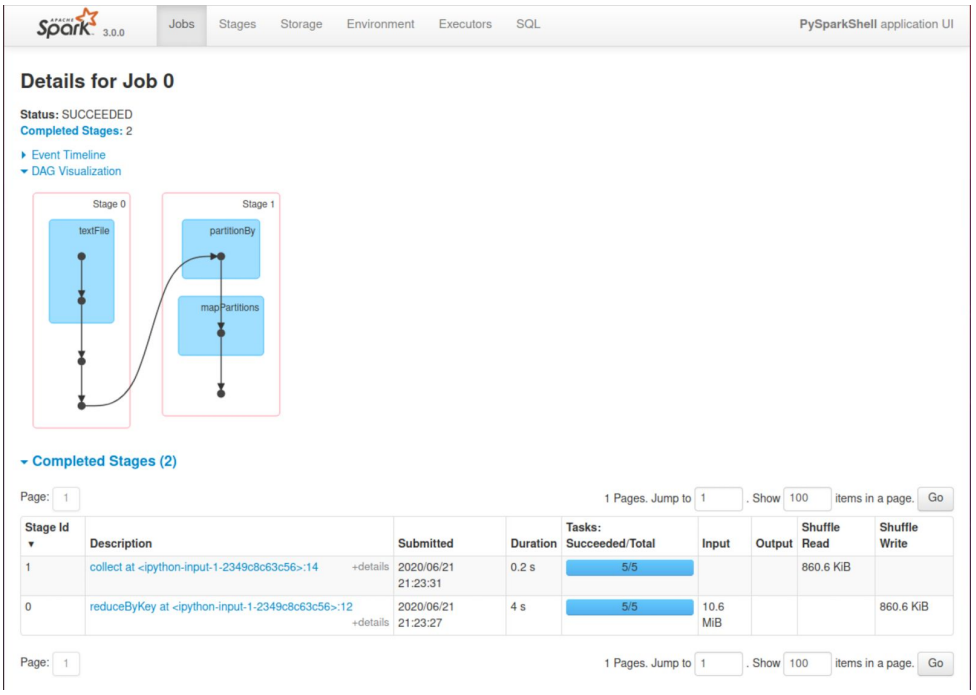
25. Congratulations!

26. While the pyspark is still running browse to http://localhost:4040

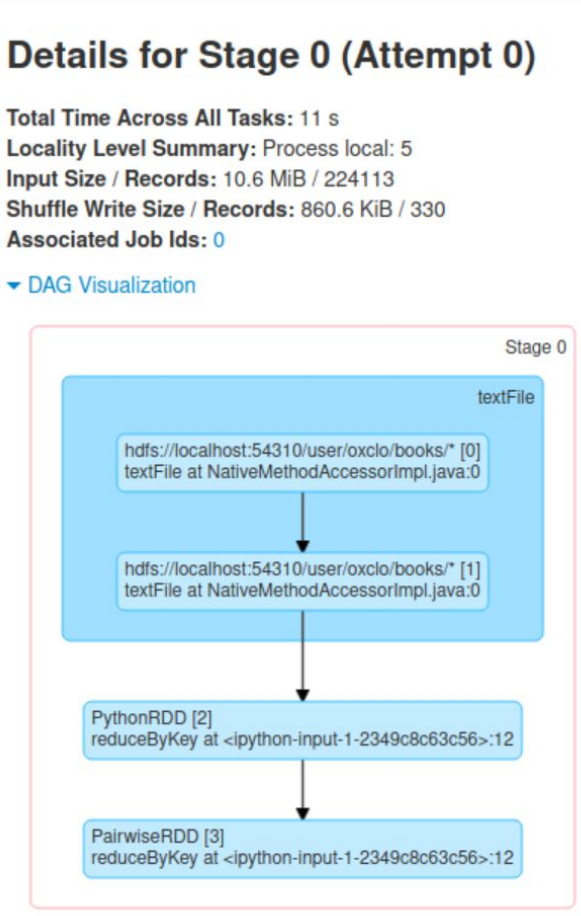
27. You will see the Spark web console:



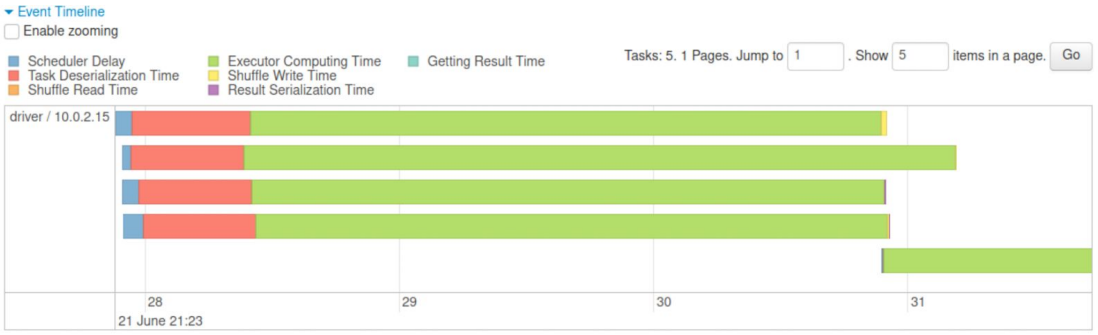
28. Click on the blue link “collect at ipython-input”
    This shows you how Spark converted your code into stages:




29. Click on Stage 0, or the latest Stage (Note that if you have done more than one run, it may not be Stage 0!)



30. And expand the Event Timeline:




**Summary Metrics for 5 Completed Tasks**

| Metric | Min | 25th percentile | Median | 75th percentile | Max |
|---|---|---|---|---|---|
| Duration | 0.8 s | 2 s | 2 s | 3 s | 3 s |
| GC Time | 0.0 ms | 10.0 ms | 10.0 ms | 10.0 ms | 10.0 ms |
| Input Size / Records | 1.8 MiB / 37272 | 1.8 MiB / 38588 | 1.9 MiB / 39566 | 2 MiB / 43680 | 3.1 MiB / 65007 |
| Shuffle Write Size / Records | 140.4 KiB / 65 | 168.1 KiB / 65 | 174.4 KiB / 65 | 176.4 KiB / 65 | 201.3 KiB / 70 |

31. Make sure your code is saved from the notebook.

32. Quit the notebook shell by typing Ctrl-C on the command line, and then Y
    Also close the notebook windows in the browser.

33. Now let's run the same code as a "job" instead of interactively.

34. From http://freo.me/oxclo-wc-py copy the code into a file `wc-job.py`

35. You will notice that there is a bunch of "setup" code that we didn't need in the pyspark command line tool. That is because pyspark assumes you want all this and does it for you.

36. The default environment in your terminal session is telling Pyspark to use Jupyter. We need to stop that. Type:

    ```
    unset PYSPARK_DRIVER_PYTHON
    unset PYSPARK_DRIVER_PYTHON_OPTS
    ```

    Note that this will reset when you start a new terminal window, so when you want to go back to using Jupyter, just kill this window and start another.

37. We run jobs locally on a single node directly on Spark:
    The local[*] indicates to use as many threads as you have cores on your system: *(all on one line)*
    ```
    ~/spark/bin/spark-submit --master local[*] wc-job.py
    "file:///home/oxclo/datafiles/books/*"
    ```

38. Congratulations, the lab is complete!

**Extension**

39. Re load the code into the Jupyter notebook and now improve it to show the wordcount in descending order, starting with the most common words.