Getting Started with Apache Spark Python REPL

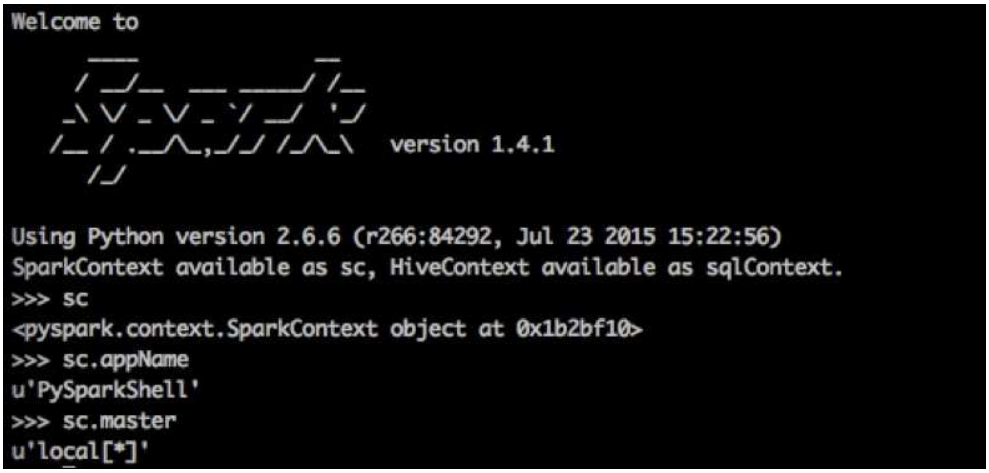
To open the spark shell, type the following:

For Python:

```
# pyspark
```

Take a look at the Spark context and some attributes:

```
>    sc
>    sc.appName
>    sc.master
```



2. From the Spark Shell, write the logic for counting all the words

Create an RDD from the file we just viewed above:

```
>>> baseRdd=sc.textFile("file:///home/notroot/lab/data/selfishgiant.txt")
```

Verify that you have created and RDD from the correct file using `take(1)`:

```
>>> baseRdd.take(1)
```

. Each element is currently a string, transform the string into arrays and examine the output:

```
>>> splitRdd = baseRdd.flatMap(lambda line: line.split(" "))
>>> splitRdd.take(5)
```

Map each element into a key/value pair, with the key being the word and the value being 1. Examine the output:

```
>>> mappedRdd = splitRdd.map(lambda line: (line,1))
>>> mappedRdd.take(5)
```

Reduce the key/value pairs to get the count of each word:

```
>>> reducedRdd = mappedRdd.reduceByKey(lambda a,b: a+b)
```

Run an action to get output:

```
>>>    reducedRdd.take(20)
>>> reducedRdd.collect()
```