Himanshu Goyal

Assigment 3

1.

```
In [3]: import json
from pyspark import SparkContext, SparkConf
sc = SparkContext.getOrCreate()
jsonRDD = sc.wholeTextFiles("C:\\Users\\himan\\Downloads\\Spark\\spark-2.2.1-bin-hadoop2.7\\examples\\src\\main\\resources\\peopl
#jsonRDD = sc.wholeTextFiles("people.json").map(lambda x: x[1])
print(jsonRDD.collect())
```

```
['{"name":"Michael"}\n{"name":"Andy", "age":30}\n{"name":"Justin", "age":19}\n']
```

2.

```
In [2]: import json
import pyspark
from pyspark import SparkConf, SparkContext
from pyspark.sql import SQLContext
conf = SparkConf().setMaster("local").setAppName("Load_Json_PYspark")
sc = SparkContext.getOrCreate()
sqlContext = SQLContext(sc)
p=sqlContext.read.json("C:\\Users\\himan\\Downloads\\Spark\\spark-2.2.1-bin-hadoop2.7\\examples\\src\\main\\resources\\people.jso
p.printSchema()
p.registerTempTable('people')
sqlContext.sql("Select distinct name from people").show()
```

```
root
 |-- age: long (nullable = true)
 |-- name: string (nullable = true)

+-------+
|   name|
+-------+
|Michael|
|   Andy|
| Justin|
+-------+
```

3.

In [51]:

```
import csv
import pyspark
from pyspark import SparkConf, SparkContext
from operator import add
conf = SparkConf().setMaster("local").setAppName("Load_Csv")
sc = SparkContext.getOrCreate()

with open("C:\\Users\\himan\\Downloads\\Spark\\spark-2.2.1-bin-hadoop2.7\\examples\\src\\main\\resources\\people.txt") as csvfile
    readCSV = csv.reader(csvfile, delimiter=',')
    print(type(readCSV))

    for row in readCSV:
        print(row)
print(sc.textFile("C:\\Users\\himan\\Downloads\\Spark\\spark-2.2.1-bin-hadoop2.7\\examples\\src\\main\\resources\\people.txt").ma
```

```
<class '_csv.reader'>
['Michael', ' 29']
['Andy', ' 30']
['Justin', ' 19']
[['Michael', ' 29'], ['Andy', ' 30'], ['Justin', ' 19']]
```