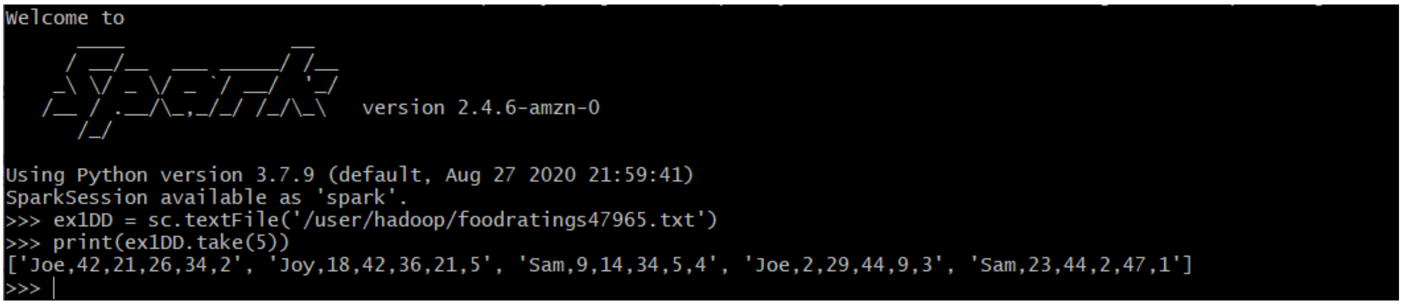
Chirag Khandhar

A20438926 | CSP 554 - Big Data Technologies | Fall 2020 | Assignment 6

Magic Number = **47965**

Q 1. Copy the file to HDFS, say into the /user/hadoop directory. Read in the text file into an RDD named ex1RDD.

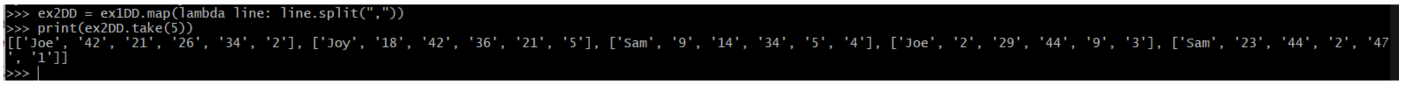
```
ex1DD = sc.textFile('/user/hadoop/foodratings47965.txt')
print(ex1DD.take(5))
```



```
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.4.6-amzn-0
      /_/

Using Python version 3.7.9 (default, Aug 27 2020 21:59:41)
SparkSession available as 'spark'.
>>> ex1DD = sc.textFile('/user/hadoop/foodratings47965.txt')
>>> print(ex1DD.take(5))
['Joe,42,21,26,34,2', 'Joy,18,42,36,21,5', 'Sam,9,14,34,5,4', 'Joe,2,29,44,9,3', 'Sam,23,44,2,47,1']
>>> |
```

Q 2. Create another RDD called ex2RDD where each record of this new RDD has 6 fields, each a string, by splitting apart each record on "," boundaries from the ex1RDD.

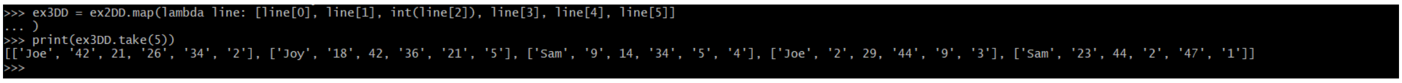
```
ex2DD = ex1DD.map(lambda line: line.split(","))
print(ex2DD.take(5))
```



```
>>> ex2DD = ex1DD.map(lambda line: line.split(","))
>>> print(ex2DD.take(5))
[['Joe', '42', '21', '26', '34', '2'], ['Joy', '18', '42', '36', '21', '5'], ['Sam', '9', '14', '34', '5', '4'], ['Joe', '2', '29', '44', '9', '3'], ['Sam', '23', '44', '2', '47
', '1']]
>>> |
```

Q 3. Create another RDD called ex3RDD from ex2RDD where each record of this new RDD has its third column converted from a string to an integer.

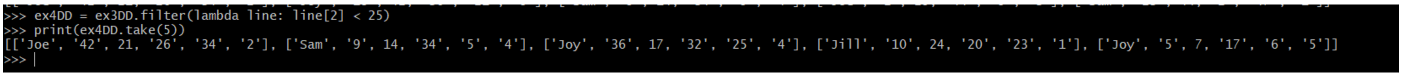
```
ex3DD = ex2DD.map(lambda line: [line[0], line[1], int(line[2]), line[3],
line[4], line[5]])

print(ex3DD.take(5))
```



```
>>> ex3DD = ex2DD.map(lambda line: [line[0], line[1], int(line[2]), line[3], line[4], line[5]]
... )
>>> print(ex3DD.take(5))
[['Joe', '42', 21, '26', '34', '2'], ['Joy', '18', 42, '36', '21', '5'], ['Sam', '9', 14, '34', '5', '4'], ['Joe', '2', 29, '44', '9', '3'], ['Sam', '23', 44, '2', '47', '1']]
>>>
```

Q 4. Create another RDD called ex4RDD from ex3RDD where each record of this new RDD is allowed to have a value for its third field that is less than 25 (<25).

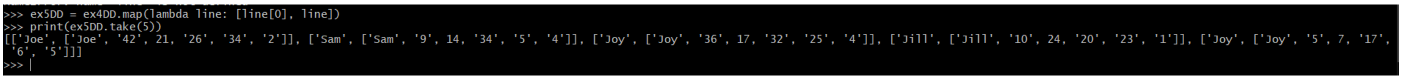
```
ex4DD = ex3DD.filter(lambda line: line[2] < 25)
print(ex4DD.take(5))
```



```
>>> ex4DD = ex3DD.filter(lambda line: line[2] < 25)
>>> print(ex4DD.take(5))
[['Joe', '42', 21, '26', '34', '2'], ['Sam', '9', 14, '34', '5', '4'], ['Joy', '36', 17, '32', '25', '4'], ['Jill', '10', 24, '20', '23', '1'], ['Joy', '5', 7, '17', '6', '5']]
>>> |
```

Q 5. Create another RDD called ex5RDD from ex4RDD where each record is a key value pair where the key is the first field of the record and the value is the entire record.

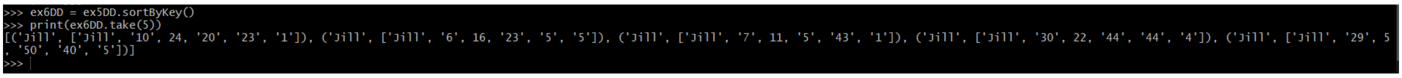
```
ex5DD = ex4DD.map(lambda line: [line[0], line])
print(ex5DD.take(5))
```



```
>>> ex5DD = ex4DD.map(lambda line: [line[0], line])
>>> print(ex5DD.take(5))
[['Joe', ['Joe', '42', 21, '26', '34', '2']], ['Sam', ['Sam', '9', 14, '34', '5', '4']], ['Joy', ['Joy', '36', 17, '32', '25', '4']], ['Jill', ['Jill', '10', 24, '20', '23', '1']], ['Joy', ['Joy', '5', 7, '17',
 '6', '5']]]
>>> |
```

Q 6. Create another RDD called ex6RDD from ex5RDD where the records are organized in ascending order by key.

```
ex6DD = ex5DD.sortByKey()
print(ex6DD.take(5))
```



```
>>> ex6DD = ex5DD.sortByKey()
>>> print(ex6DD.take(5))
[('Jill', ['Jill', '10', 24, '20', '23', '1']), ('Jill', ['Jill', '6', 16, '23', '5', '5']), ('Jill', ['Jill', '7', 11, '5', '43', '1']), ('Jill', ['Jill', '30', 22, '44', '44', '4']), ('Jill', ['Jill', '29', 5
, '50', '40', '5'])]
>>> 
```