CSP 554 – Assignment #6

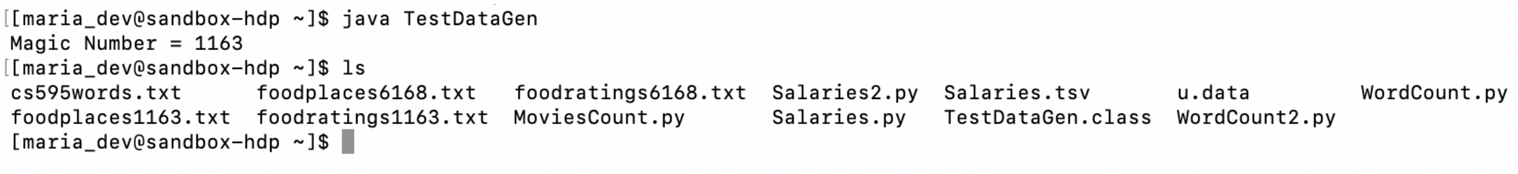
Name: Adarsh Mathad Vijayakumar
CWID: A20424847
Email ID: avijayakumar@hawk.iit.edu

Exercise 1) Use the TestDataGen program from previous assignments to generate a new foodratings<magic_number>.txt data file.
Copy the file to HDFS, say into the /user/maria_dev directory.

Read in the text file into an RDD named ex1RDD.

List the first five records of the RDD using the "take(5)" action and copy them and the "magic number to your assignment submission for this exercise.

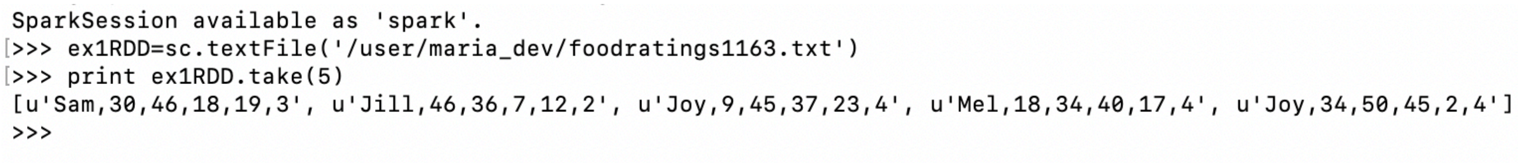
Using the TestDataGen program to generate a new foodratings<magic_number>.txt data file.
Command: java TestDataGen
Magic Number = 1163



```
[[maria_dev@sandbox-hdp ~]$ java TestDataGen
 Magic Number = 1163
[[maria_dev@sandbox-hdp ~]$ ls
 cs595words.txt      foodplaces6168.txt   foodratings6168.txt  Salaries2.py  Salaries.tsv       u.data         WordCount.py
 foodplaces1163.txt  foodratings1163.txt  MoviesCount.py       Salaries.py   TestDataGen.class  WordCount2.py
 [maria_dev@sandbox-hdp ~]$ █
```

Commands:
ex1RDD=sc.textFile('/user/maria_dev/foodratings1163.txt')
print ex1RDD.take(5)

Output:
[u'Sam,30,46,18,19,3', u'Jill,46,36,7,12,2', u'Joy,9,45,37,23,4', u'Mel,18,34,40,17,4', u'Joy,34,50,45,2,4']



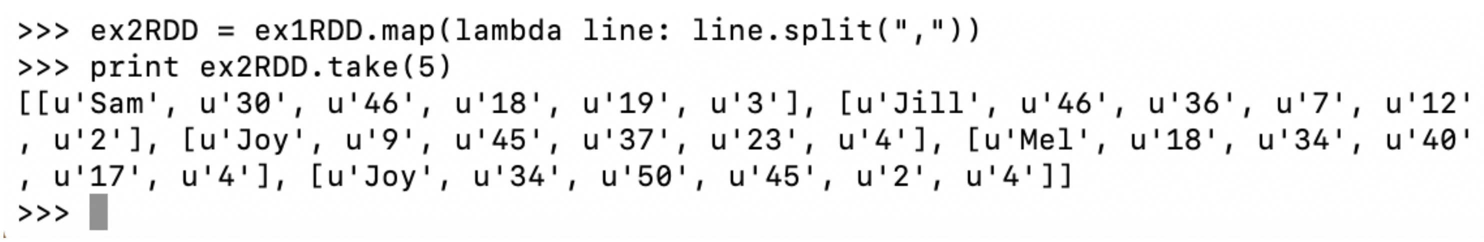
```
 SparkSession available as 'spark'.
[>>> ex1RDD=sc.textFile('/user/maria_dev/foodratings1163.txt')
[>>> print ex1RDD.take(5)
 [u'Sam,30,46,18,19,3', u'Jill,46,36,7,12,2', u'Joy,9,45,37,23,4', u'Mel,18,34,40,17,4', u'Joy,34,50,45,2,4']
 >>>
```

Exercise 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.

List the first five records of this RDD using the "take(5)" action and copy them to your assignment submission for this exercise.

Commands:
ex2RDD = ex1RDD.map(lambda line: line.split(","))
print ex2RDD.take(5)

Output:
[[u'Sam', u'30', u'46', u'18', u'19', u'3'], [u'Jill', u'46', u'36', u'7', u'12', u'2'], [u'Joy', u'9', u'45', u'37', u'23', u'4'], [u'Mel', u'18', u'34', u'40', u'17', u'4'], [u'Joy', u'34', u'50', u'45', u'2', u'4']]



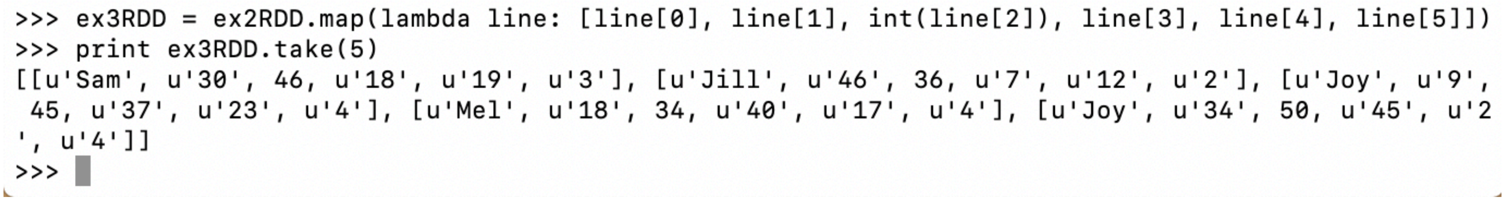
```
>>> ex2RDD = ex1RDD.map(lambda line: line.split(","))
>>> print ex2RDD.take(5)
[[u'Sam', u'30', u'46', u'18', u'19', u'3'], [u'Jill', u'46', u'36', u'7', u'12'
, u'2'], [u'Joy', u'9', u'45', u'37', u'23', u'4'], [u'Mel', u'18', u'34', u'40'
, u'17', u'4'], [u'Joy', u'34', u'50', u'45', u'2', u'4']]
>>> 
```

Exercise 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.

List the first five records of this RDD using the “take(5)” action and copy them to your assignment submission for this exercise.

Commands:
ex3RDD = ex2RDD.map(lambda line: [line[0],line[1],int(line[2]),line[3],line[4],line[5]])
print ex3RDD.take(5)

Output:
[[u'Sam', u'30', 46, u'18', u'19', u'3'], [u'Jill', u'46', 36, u'7', u'12', u'2'], [u'Joy', u'9', 45, u'37', u'23', u'4'], [u'Mel', u'18', 34, u'40', u'17', u'4'], [u'Joy', u'34', 50, u'45', u'2', u'4']]



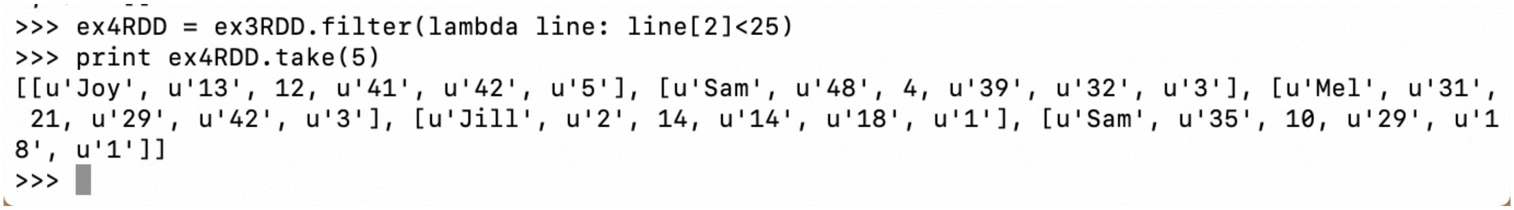
```
>>> ex3RDD = ex2RDD.map(lambda line: [line[0], line[1], int(line[2]), line[3], line[4], line[5]])
>>> print ex3RDD.take(5)
[[u'Sam', u'30', 46, u'18', u'19', u'3'], [u'Jill', u'46', 36, u'7', u'12', u'2'], [u'Joy', u'9',
 45, u'37', u'23', u'4'], [u'Mel', u'18', 34, u'40', u'17', u'4'], [u'Joy', u'34', 50, u'45', u'2
', u'4']]
>>> 
```

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

List the first five records of this RDD using the “take(5)” action and copy them to your assignment submission for this exercise.

Commands:
ex4RDD = ex3RDD.filter(lambda line: line[2]<25)
print ex4RDD.take(5)

Output:
[[u'Joy', u'13', 12, u'41', u'42', u'5'], [u'Sam', u'48', 4, u'39', u'32', u'3'], [u'Mel', u'31', 21, u'29', u'42', u'3'], [u'Jill', u'2', 14, u'14', u'18', u'1'], [u'Sam', u'35', 10, u'29', u'18', u'1']]



```
>>> ex4RDD = ex3RDD.filter(lambda line: line[2]<25)
>>> print ex4RDD.take(5)
[[u'Joy', u'13', 12, u'41', u'42', u'5'], [u'Sam', u'48', 4, u'39', u'32', u'3'], [u'Mel', u'31',
 21, u'29', u'42', u'3'], [u'Jill', u'2', 14, u'14', u'18', u'1'], [u'Sam', u'35', 10, u'29', u'1
8', u'1']]
>>> 
```

Exercise 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.

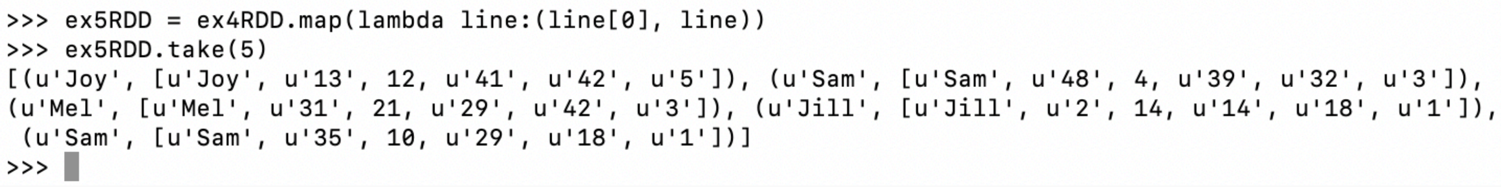
List the first five records of this RDD using the “take(5)” action and copy them to your assignment submission for this exercise.

Commands:

```
ex5RDD = ex4RDD.map(lambda line:(line[0], line))
ex5RDD.take(5)
```

Output:

```
[(u'Joy', [u'Joy', u'13', 12, u'41', u'42', u'5']), (u'Sam', [u'Sam', u'48', 4, u'39', u'32', u'3']), (u'Mel', [u'Mel', u'31',
21, u'29', u'42', u'3']), (u'Jill', [u'Jill', u'2', 14, u'14', u'18', u'1']), (u'Sam', [u'Sam', u'35', 10, u'29', u'18', u'1'])]
```



```
>>> ex5RDD = ex4RDD.map(lambda line:(line[0], line))
>>> ex5RDD.take(5)
[(u'Joy', [u'Joy', u'13', 12, u'41', u'42', u'5']), (u'Sam', [u'Sam', u'48', 4, u'39', u'32', u'3']),
(u'Mel', [u'Mel', u'31', 21, u'29', u'42', u'3']), (u'Jill', [u'Jill', u'2', 14, u'14', u'18', u'1']),
 (u'Sam', [u'Sam', u'35', 10, u'29', u'18', u'1'])]
>>> 
```

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

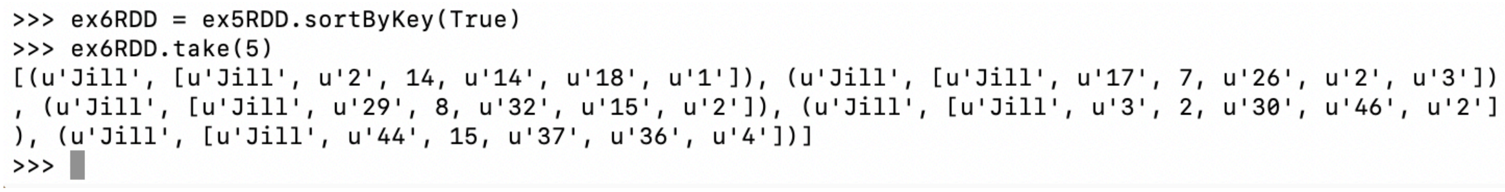
List the first five records of this RDD using the “take(5)” action and copy them to your assignment submission for this exercise.

Commands:

```
ex6RDD = ex5RDD.sortByKey(True)
ex6RDD.take(5)
```

Output:

```
[(u'Jill', [u'Jill', u'2', 14, u'14', u'18', u'1']), (u'Jill', [u'Jill', u'17', 7, u'26', u'2', u'3']), (u'Jill', [u'Jill', u'29', 8, u'32',
u'15', u'2']), (u'Jill', [u'Jill', u'3', 2, u'30', u'46', u'2']), (u'Jill', [u'Jill', u'44', 15, u'37', u'36', u'4'])]
```



```
>>> ex6RDD = ex5RDD.sortByKey(True)
>>> ex6RDD.take(5)
[(u'Jill', [u'Jill', u'2', 14, u'14', u'18', u'1']), (u'Jill', [u'Jill', u'17', 7, u'26', u'2', u'3'])
, (u'Jill', [u'Jill', u'29', 8, u'32', u'15', u'2']), (u'Jill', [u'Jill', u'3', 2, u'30', u'46', u'2']
), (u'Jill', [u'Jill', u'44', 15, u'37', u'36', u'4'])]
>>> 
```