ASSIGNMENT - 6

**CSP - 554 BIG DATA TECHNOLOGIES**

**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.

This RDD should now have records each consisting of a single string having 6 comma-separated parts something like the following:

u'Joe,44,33,41,1,5'

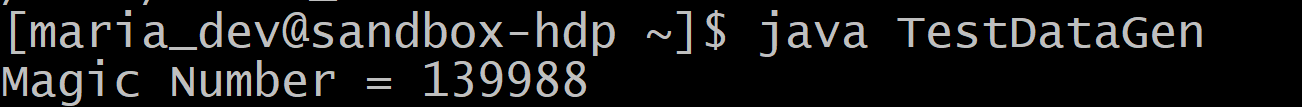
u'Mel,13,33,30,50,6'

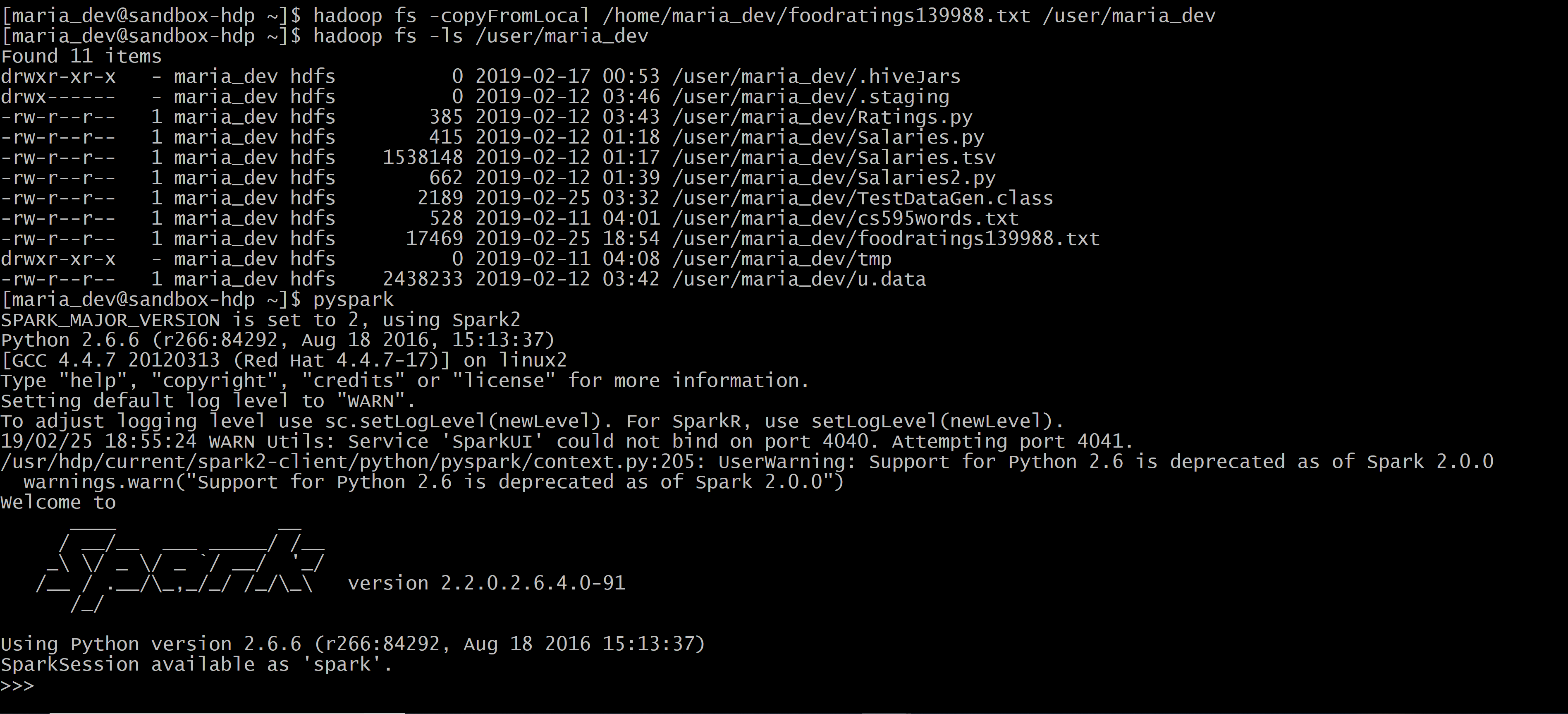
u'Mel,12,40,30,42,1'

u'Sam,15,28,28,39,2'

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.

Magic Number = 139988





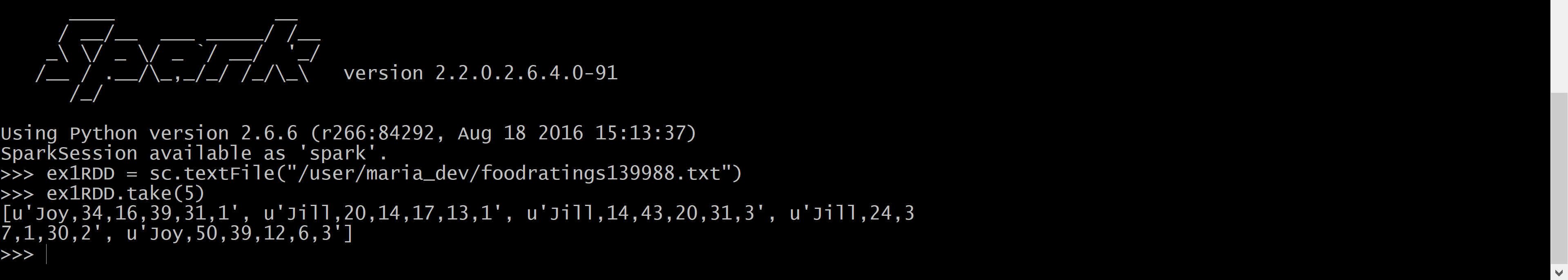
Read in the text file into an RDD named ex1RDD.

**Command:**

ex1RDD = sc.textFile("/user/maria\_dev/foodratings139988.txt")

ex1RDD.take(5)

[u'Joy,34,16,39,31,1', u'Jill,20,14,17,13,1', u'Jill,14,43,20,31,3', u'Jill,24,3,7,1,30,2', u'Joy,50,39,12,6,3']



**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.

The records of the new RDD should look something like:

u'Joe', u'44', u'33', u'41', u'1‘, u’5’

u‘Mel', u'13', u'33', u'30', u'50, u’6’‘

u‘Mel', u'12', u'40', u'30', u'42‘, u’1’

u'Sam', u'15', u'28', u'28', u'39‘, u’3’

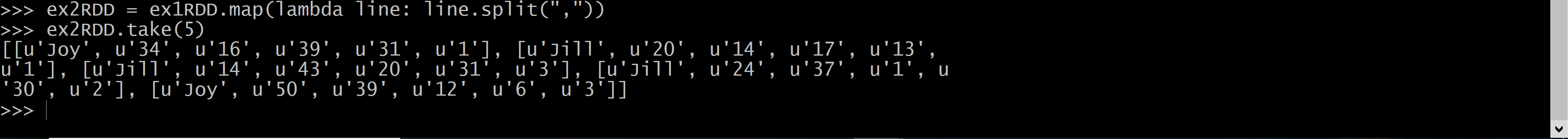
List the first five records of this RDD using the “take(5)” action and copy them to your assignment submission for this exercise.

**Command:**

ex2RDD = ex1RDD.map(lambda line: line.split(“,”))

ex2RDD.take(5)

[[u'Joy', u'34', u'16', u'39', u'31', u'1'], [u'Jill', u'20', u'14', u'17', u'13', u'1'], [u'Jill', u'14', u'43', u'20', u'31', u'3'], [u'Jill', u'24', u'37', u'1', u'30', u'2'], [u'Joy', u'50', u'39', u'12', u'6', u'3']]



**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.

The records of the new RDD should look something like:

u'Joe', u'44', 33, u'41', u'1‘, u’1’

u‘Mel', u'13', 33, u'30', u'50‘, u’2’

u‘Mel', u'12', 40, u'30', u'42‘, u’3’

u'Sam', u'15', 28, u'28', u'39‘, u’4’

Hint: Use a lambda function something like the following:

lambda line : [line[0], line[1], int(line[2]), line[3], line[4], line[5]]

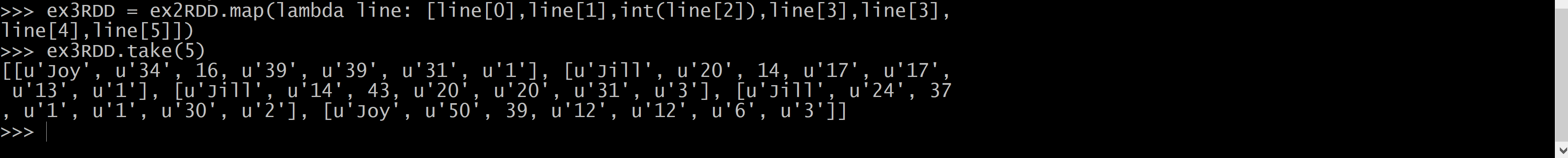
List the first five records of this RDD using the “take(5)” action and copy them to your assignment submission for this exercise.

**Command:**

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

ex3RDD.take(5)

[[u'Joy', u'34', 16, u'39', u'39', u'31', u'1'], [u'Jill', u'20', 14, u'17', u'17', u'13', u'1'], [u'Jill', u'14', 43, u'20', u'20', u'31', u'3'], [u'Jill', u'24', 37, u'1', u'1', u'30', u'2'], [u'Joy', u'50', 39, u'12', u'12', u'6', u'3']]



**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.

The records of the new RDD should look something like:

u'Joe', u'44', 21, u'41', u'1‘, u’6’

u‘Mel', u'13', 3, u'30', u'50‘, u’1’

u‘Mel', u'12', 4, u'30', u'42‘, u’4’

u'Sam', u'15', 8, u'28', u'39‘, u’5’

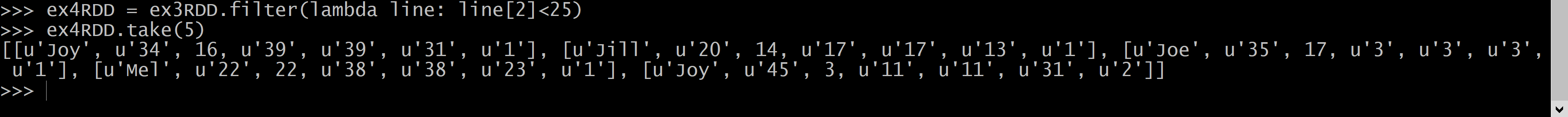
List the first five records of this RDD using the “take(5)” action and copy them to your assignment submission for this exercise.

**Command:**

ex4RDD = ex3RDD.filter(lambda line: line[2]<25)

ex4RDD.take(5)

[[u'Joy', u'34', 16, u'39', u'39', u'31', u'1'], [u'Jill', u'20', 14, u'17', u'17', u'13', u'1'], [u'Joe', u'35', 17, u'3', u'3', u'3', u'1'], [u'Mel', u'22', 22, u'38', u'38', u'23', u'1'], [u'Joy', u'45', 3, u'11', u'11', u'31', u'2']]



**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

The records of the new RDD should look something like:

(u’Joe’, (u'Joe', u'44', 21, u'41', u'1‘, u’1’))

(u’Mel’, (u‘Mel', u'13', 3, u'30', u'50‘, u’6’))

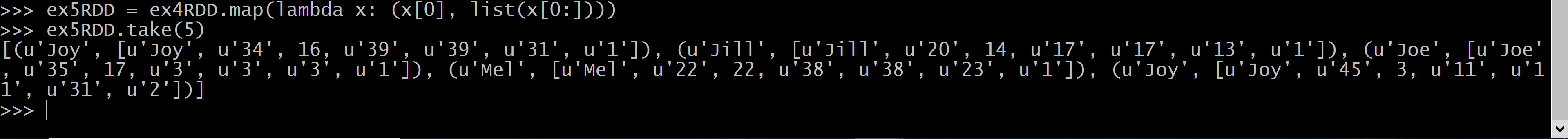
List the first five records of this RDD using the “take(5)” action and copy them to your assignment submission for this exercise.

**Command:**

ex5RDD = ex4RDD.map(lambda x: (x[0], list(x[0:])))

ex5RDD.take(5)

[(u'Joy', [u'Joy', u'34', 16, u'39', u'39', u'31', u'1']), (u'Jill', [u'Jill', u'20', 14, u'17', u'17', u'13', u'1']), (u'Joe', [u'Joe', u'35', 17, u'3', u'3', u'3', u'1']), (u'Mel', [u'Mel', u'22', 22, u'38', u'38', u'23', u'1']), (u'Joy', [u'Joy', u'45', 3, u'11', u'11', u'31', u'2'])]



**Exercise 6**

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

The records of the new RDD should look something like:

(u’Joe’, (u'Joe', u'44', 21, u'41', u'1‘, u’4’))

(u’Mel’ , (u‘Mel', u'13', 3, u'30', u'50‘, u’3’))

(u’Sam’ , (u‘Sam', u'23', 3, u'40', u'20‘, u’7’))

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

**Command:**

ex6RDD = ex5RDD.sortByKey(True)

ex6RDD.take(5)

[(u'Jill', [u'Jill', u'20', 14, u'17', u'17', u'13', u'1']), (u'Jill', [u'Jill', u'34', 17, u'27', u'27', u'4', u'4']), (u'Jill', [u'Jill', u'48', 16, u'2', u'2', u'22', u'5']), (u'Jill', [u'Jill', u'15', 17, u'20', u'20', u'32', u'3']), (u'Jill', [u'Jill', u'37', 23, u'33', u'33', u'41', u'3'])]

