CSP 554 Assignment 6

Atharva Tanaji Kadam (A20467229)

Magic Number:- 177757

Exercise 1 -

```
[hadoop@ip-172-31-68-38 ~]$ hdfs dfs -ls /user/hadoop/*177757*
-rw-r--r--   1 hadoop hadoop         59 2022-10-13 23:59 /user/hadoop/foodplaces177757.txt
-rw-r--r--   1 hadoop hadoop      17461 2022-10-13 23:59 /user/hadoop/foodratings177757.txt
[hadoop@ip-172-31-68-38 ~]$
```

```
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.4.3
      /_/

Using Python version 2.7.18 (default, May 30 2022 10:09:41)
SparkSession available as 'spark'.
>>> ex1RDD = sc.textFile('/user/hadoop/foodratings177757.txt')
>>> ex1RDD.take(5)
[Stage 0:>                                                          (
[Stage 0:>                                                          (
[Stage 0:============================================================(

[u'Sam,34,47,9,43,5', u'Joe,29,30,23,27,5', u'Sam,24,28,27,28,4', u'S
am,24,36,9,13,2', u'Jill,18,17,29,15,3']
>>>
```

ex1RDD = sc.textFile('/user/hadoop/foodratings177757.txt')
ex1RDD.take(5)

Exercises 2:

```
>>> ex2RDD = ex1RDD.map(lambda line: line.split(","))
>>> ex2RDD.take(5)
[Stage 1:>                                                          (
[Stage 1:>                                                          (

[[u'Sam', u'34', u'47', u'9', u'43', u'5'], [u'Joe', u'29', u'30', u'
23', u'27', u'5'], [u'Sam', u'24', u'28', u'27', u'28', u'4'], [u'Sam
', u'24', u'36', u'9', u'13', u'2'], [u'Jill', u'18', u'17', u'29', u
'15', u'3']]
>>>
```

ex2RDD = ex1RDD.map(lambda line: line.split(“,”))
ex2RDD.take(5)

Exercise 3:

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

[[u'Sam', u'34', 47, u'9', u'43', u'5'], [u'Joe', u'29', 30, u'23', u
'27', u'5'], [u'Sam', u'24', 28, u'27', u'28', u'4'], [u'Sam', u'24',
 36, u'9', u'13', u'2'], [u'Jill', u'18', 17, u'29', u'15', u'3']]
>>>
```

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

Exercise 4:

```
 36, u'9', u'13', u'2'], [u'Jill', u'18', 17, u'29', u'15', u'3']]
>>> ex4RDD = ex3RDD.filter(lambda line: int(line[2])<25)
>>> ex4RDD.take(5)
[Stage 3:>                                                          (
[Stage 3:>                                                          (

[[u'Jill', u'18', 17, u'29', u'15', u'3'], [u'Sam', u'20', 24, u'18',
 u'35', u'5'], [u'Joy', u'5', 23, u'18', u'50', u'2'], [u'Sam', u'33'
, 19, u'36', u'49', u'3'], [u'Sam', u'2', 17, u'9', u'15', u'5']]
>>>
```

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

Exercise 5:

```
>>> ex5RDD = ex4RDD.map(lambda line: [line[0], line])
>>> ex5RDD.take(5)
[Stage 4:>                                                          (
[Stage 4:>                                                          (

[[u'Jill', [u'Jill', u'18', 17, u'29', u'15', u'3']], [u'Sam', [u'Sam
', u'20', 24, u'18', u'35', u'5']], [u'Joy', [u'Joy', u'5', 23, u'18'
, u'50', u'2']], [u'Sam', [u'Sam', u'33', 19, u'36', u'49', u'3']], [
u'Sam', [u'Sam', u'2', 17, u'9', u'15', u'5']]]
>>>
```

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

Exercise 6:

```
>>> ex6RDD = ex5RDD.sortByKey()
[Stage 5:>
 [Stage 5:>

 >>> ex6RDD.take(5)
 [Stage 7:>

 [(u'Jill', [u'Jill', u'18', 17, u'29', u'15', u'3']), (u'Jill', [u'Ji
 ll', u'6', 4, u'7', u'27', u'4']), (u'Jill', [u'Jill', u'8', 8, u'24'
 , u'18', u'4']), (u'Jill', [u'Jill', u'42', 2, u'46', u'4', u'5']), (
 u'Jill', [u'Jill', u'12', 2, u'15', u'22', u'1'])]
 >>>
```

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