wordcount1

Cmd 1

```
1 wordsList = ['cat', 'elephant', 'rat', 'rat', 'cat']
```

Command took 0.11 seconds -- by louhy1128@gmail.com at 2017-6-21 15:22:37 on My Cluster

Cmd 2

```
1 wordsRDD = sc.parallelize(wordsList, 4)
```

Command took 0.28 seconds -- by louhy1128@gmail.com at 2017-6-21 15:23:21 on My Cluster

Cmd 3

```
1 wordsRDD
```

```
Out[6]: ParallelCollectionRDD[0] at parallelize at PythonRDD.scala:475
```

Command took 0.04 seconds -- by louhy1128@gmail.com at 2017-6-21 15:24:48 on My Cluster

Cmd 4

```
1 wordsRDD.take(1)
```

▸ (1) Spark Jobs

```
Out[7]: ['cat']
```

Command took 0.07 seconds -- by louhy1128@gmail.com at 2017-6-21 15:24:52 on My Cluster

Cmd 5

```
1 print type(wordsRDD)
```

```
<class 'pyspark.rdd.RDD'>
```

Command took 0.07 seconds -- by louhy1128@gmail.com at 2017-6-21 15:25:21 on My Cluster

Cmd 6

```
1  def makePlural(word):
2      """Adds an 's' to `word`.
3
4      Note:
5          This is a simple function that only adds an 's'.  No attempt is
   made to follow proper
6          pluralization rules.
7
8      Args:
9          word (str): A string.
10
11     Returns:
12         str: A string with 's' added to it.
13     """
14     return word + 's'
```

Command took 0.04 seconds -- by louhy1128@gmail.com at 2017-6-21 15:26:33 on My Cluster

Cmd 7

```
1 makePlural('cat')
```

```
Out[10]: 'cats'
```

Command took 0.07 seconds -- by louhy1128@gmail.com at 2017-6-21 15:26:45 on My Cluster

Cmd 8

```
1 wordsRDD.map(makePlural).collect()
```

▸ (1) Spark Jobs

Out[12]: ['cats', 'elephants', 'rats', 'rats', 'cats']

Command took 0.17 seconds -- by louhy1128@gmail.com at 2017-6-21 15:29:41 on My Cluster

Cmd 9

```
1 pluralRDD = wordsRDD.map(makePlural)
```

Command took 0.04 seconds -- by louhy1128@gmail.com at 2017-6-21 15:30:01 on My Cluster

Cmd 10

```
1 pluralRDD = wordsRDD.map(lambda x: makePlural(x))
```

Command took 0.04 seconds -- by louhy1128@gmail.com at 2017-6-21 15:30:19 on My Cluster

Cmd 11

```
1 pluralRDD.map(len)
```

Out[15]: PythonRDD[5] at RDD at PythonRDD.scala:44

Command took 0.07 seconds -- by louhy1128@gmail.com at 2017-6-21 15:31:21 on My Cluster

Cmd 12

```
1 pluralRDD.map(len).collect()
```

▸ (1) Spark Jobs

Out[16]: [4, 9, 4, 4, 4]

Command took 0.12 seconds -- by louhy1128@gmail.com at 2017-6-21 15:32:04 on My Cluster

Cmd 13

```
1 pluralLengths = (pluralRDD.map(lambda x: len(x))
2                  .collect())
```

▸ (1) Spark Jobs

Command took 0.13 seconds -- by louhy1128@gmail.com at 2017-6-21 15:33:37 on My Cluster

Cmd 14

```
1 wordPairs = wordsRDD.map(lambda x: (x,1))
```

Command took 0.07 seconds -- by louhy1128@gmail.com at 2017-6-21 15:34:36 on My Cluster

Cmd 15

```
1 wordPairs.collect()
```

▸ (1) Spark Jobs

Out[20]: [('cat', 1), ('elephant', 1), ('rat', 1), ('rat', 1), ('cat', 1)]

Command took 0.13 seconds -- by louhy1128@gmail.com at 2017-6-21 15:34:47 on My Cluster

Cmd 16

```
1 wordCounts = wordPairs.reduceByKey(lambda x,y:x+y)
```

Command took 0.08 seconds -- by louhy1128@gmail.com at 2017-6-21 15:38:40 on My Cluster

Cmd 17

```
1 wordCounts.collect()
```

▸ (1) Spark Jobs

```
Out[23]: [('rat', 2), ('elephant', 1), ('cat', 2)]
```

Command took 0.48 seconds -- by louhy1128@gmail.com at 2017-6-21 15:38:45 on My Cluster

Cmd 18

```
1 wordCountsCollected = (wordsRDD.map(lambda x: (x,1))
2                        .reduceByKey(lambda x,y:x+y)
3                        .collect())
```

▸ (1) Spark Jobs

Command took 0.23 seconds -- by louhy1128@gmail.com at 2017-6-21 15:40:42 on My Cluster

Cmd 19

```
1 print wordCountsCollected
```

```
[('rat', 2), ('elephant', 1), ('cat', 2)]
```

Command took 0.07 seconds -- by louhy1128@gmail.com at 2017-6-21 15:40:56 on My Cluster

Cmd 20

```
1 uniqueWords = wordsRDD.distinct()
```

Command took 0.12 seconds -- by louhy1128@gmail.com at 2017-6-21 15:42:52 on My Cluster

Cmd 21

```
1 uniqueWords = wordsRDD.distinct().count()
```

▸ (1) Spark Jobs

Command took 0.22 seconds -- by louhy1128@gmail.com at 2017-6-21 15:43:32 on My Cluster

Cmd 22

```
1 print uniqueWords
```

```
3
```

Command took 0.04 seconds -- by louhy1128@gmail.com at 2017-6-21 15:43:42 on My Cluster

Cmd 23

```
1  def wordCount(wordListRDD):
2      """Creates a pair RDD with word counts from an RDD of words.
3
4      Args:
5          wordListRDD (RDD of str): An RDD consisting of words.
6
7      Returns:
8          RDD of (str, int): An RDD consisting of (word, count) tuples.
9      """
10     return wordListRDD.map(lambda x: (x,1)).reduceByKey(lambda x,y:x+y)
```

Command took 0.07 seconds -- by louhy1128@gmail.com at 2017-6-21 15:46:25 on My Cluster

Cmd 24

```
import re
def removePunctuation(text):
    """Removes punctuation, changes to lower case, and strips leading and trailing spaces.

    Note:
        Only spaces, letters, and numbers should be retained.  Other characters should should be
        eliminated (e.g. it's becomes its).  Leading and trailing spaces should be removed after
        punctuation is removed.

    Args:
        text (str): A string.

    Returns:
        str: The cleaned up string.
    """
    return re.sub(r'[^a-z0-9\s]','', text.lower()).strip()

```

Command took 0.07 seconds -- by louhy1128@gmail.com at 2017-6-21 15:46:58 on My Cluster

Cmd 25

```
print removePunctuation('Hi, you!')
```

```
hi you
```

Command took 0.04 seconds -- by louhy1128@gmail.com at 2017-6-21 15:47:21 on My Cluster

Cmd 26

```
print removePunctuation(' No under_score!')
```

```
no underscore
```

Command took 0.07 seconds -- by louhy1128@gmail.com at 2017-6-21 15:47:45 on My Cluster

Cmd 27

```
print removePunctuation(' *      Remove punctuation then spaces  * ')
```

```
remove punctuation then spaces
```

Command took 0.01 seconds -- by louhy1128@gmail.com at 2017-6-21 15:47:48 on My Cluster

Cmd 28

```
import os.path
baseDir = os.path.join('databricks-datasets')
inputPath = os.path.join('cs100', 'lab1', 'data-001', 'shakespeare.txt')
fileName = os.path.join(baseDir, inputPath)

```

Command took 0.04 seconds -- by louhy1128@gmail.com at 2017-6-21 15:48:36 on My Cluster

Cmd 29

```
print fileName
```

```
databricks-datasets/cs100/lab1/data-001/shakespeare.txt
```

Command took 0.07 seconds -- by louhy1128@gmail.com at 2017-6-21 15:49:11 on My Cluster

Cmd 30

```
1  shakespeareRDD = (sc
2                    .textFile(fileName, 8)
3                    .map(removePunctuation))
```

Command took 0.13 seconds -- by louhy1128@gmail.com at 2017-6-21 15:50:02 on My Cluster

Cmd 31

```
1  shakespeareRDD.take(100)
```

▸ (1) Spark Jobs

```
Out[39]:
[u'1609',
 u'',
 u'the sonnets',
 u'',
 u'by william shakespeare',
 u'',
 u'',
 u'',
 u'1',
 u'from fairest creatures we desire increase',
 u'that thereby beautys rose might never die',
 u'but as the riper should by time decease',
 u'his tender heir might bear his memory',
 u'but thou contracted to thine own bright eyes',
 u'feedst thy lights flame with selfsubstantial fuel',
 u'making a famine where abundance lies',
 u'thy self thy foe to thy sweet self too cruel',
 u'thou that art now the worlds fresh ornament',
 u'and only herald to the gaudy spring',
 u'within thine own bud buriest thy content',
```

Command took 0.17 seconds -- by louhy1128@gmail.com at 2017-6-21 15:52:57 on My Cluster

Cmd 32

```
1  shakespeareRDD.collect()
```

▸ (1) Spark Jobs

```
 u'when what i seek my weary travels end',
 u'doth teach that case and that repose to say',
 u'thus far the miles are measured from thy friend',
 u'the beast that bears me tired with my woe',
 u'plods dully on to bear that weight in me',
 u'as if by some instinct the wretch did know',
 u'his rider loved not speed being made from thee',
 u'the bloody spur cannot provoke him on',
```

```
u'that sometimes anger thrusts into his hide',
u'which heavily he answers with a groan',

u'more sharp to me than spurring to his side',
u'for that same groan doth put this in my mind',
u'my grief lies onward and my joy behind',
u'',
u'',
u'51',
u'thus can my love excuse the slow offence',
u'of my dull bearer when from thee i speed',
u'from where thou art why should i haste me thence',
u'till i return of posting is no need',
u'o what excuse will my poor beast then find'
```

Command took 0.79 seconds -- by louhy1128@gmail.com at 2017-6-21 15:53:08 on My Cluster

Cmd 33

```
print '\n'.join(shakespeareRDD
                .zipWithIndex()  # to (line, lineNum)
                .map(lambda (l, num): '{0}: {1}'.format(num, l))  # to
'lineNum: line'
                .take(15))
```

▸ (2) Spark Jobs

```
0: 1609
1:
2: the sonnets
3:
4: by william shakespeare
5:
6:
7:
8: 1
9: from fairest creatures we desire increase
10: that thereby beautys rose might never die
11: but as the riper should by time decease
12: his tender heir might bear his memory
13: but thou contracted to thine own bright eyes
14: feedst thy lights flame with selfsubstantial fuel
```

Command took 0.83 seconds -- by louhy1128@gmail.com at 2017-6-21 15:53:52 on My Cluster

Cmd 34

```
shakespeareWordsRDD = shakespeareRDD.flatMap(lambda x: x.split(" "))
```

Command took 0.07 seconds -- by louhy1128@gmail.com at 2017-6-21 15:54:26 on My Cluster

Cmd 35

```
shakespeareWordCount = shakespeareWordsRDD.count()
```

▸ (1) Spark Jobs

Command took 0.72 seconds -- by louhy1128@gmail.com at 2017-6-21 15:54:46 on My Cluster

Cmd 36

```
1 print shakespeareWordCount
```

```
927631
```

Command took 0.02 seconds -- by louhy1128@gmail.com at 2017-6-21 15:55:03 on My Cluster

Cmd 37

```
1 print shakespeareWordsRDD.top(5)
```

▸ (1) Spark Jobs

```
[u'zwaggerd', u'zounds', u'zounds', u'zounds', u'zounds']
```

Command took 0.73 seconds -- by louhy1128@gmail.com at 2017-6-21 15:55:27 on My Cluster

Cmd 38

```
1 shakeWordsRDD = shakespeareWordsRDD.filter(lambda x: len(x)>0)
```

Command took 0.04 seconds -- by louhy1128@gmail.com at 2017-6-21 15:56:57 on My Cluster

Cmd 39

```
1 shakeWordsRDD.take(10)
```

▸ (1) Spark Jobs

```
Out[47]:
[u'1609',
 u'the',
 u'sonnets',
 u'by',
 u'william',
 u'shakespeare',
 u'1',
 u'from',
 u'fairest',
 u'creatures']
```

Command took 0.22 seconds -- by louhy1128@gmail.com at 2017-6-21 15:57:38 on My Cluster

Cmd 40

```
1 wordsAndCounts = wordCount(shakeWordsRDD)
```

Command took 0.07 seconds -- by louhy1128@gmail.com at 2017-6-21 15:58:33 on My Cluster

Cmd 41

```
1 print wordsAndCounts.collect()
```

▸ (1) Spark Jobs

```
r', 54), (u'rolled', 1), (u'overlive', 1), (u'certes', 5), (u'lackbrain', 1),
 (u'unpleasantst', 1), (u'date', 19), (u'manycolourd', 1), (u'cheerful', 17),
 (u'bespeak', 6), (u'vestments', 2), (u'sufficeth', 7), (u'note', 148), (u'dul
y', 8), (u'andromache', 12), (u'claudios', 6), (u'shoot', 29), (u'resembling',
 4), (u'rivality', 1), (u'crowflowers', 1), (u'alligator', 1), (u'urgd', 26),
 (u'closer', 2), (u'redder', 1), (u'halting', 6), (u'syllable', 10), (u'blithil
d', 1), (u'unkindness', 23), (u'roused', 1), (u'jackalent', 2), (u'pedant', 36),
```

```
(u'sleekheaded', 1), (u'pair', 40), (u'furlongs', 2), (u'frightful', 2), (u'wind
pipes', 1), (u'luggd', 1), (u'resident', 2), (u'nole', 1), (u'foot', 169), (u'hu
lks', 1), (u'givet', 12), (u'occidental', 1), (u'ways', 68), (u'rowelhead', 1),
 (u'stockings', 14), (u'amendment', 4), (u'fleshly', 1), (u'shop', 13), (u'excep
ted', 6), (u'german', 10), (u'foola', 1), (u'hurld', 2), (u'tenderhearted', 1),

 (u'fervour', 3), (u'hopes', 57), (u'rattle', 1), (u'manywhom', 1), (u'caucasu
s', 2), (u'cockpigeon', 1), (u'countrvs', 1), (u'gnat', 4), (u'spoons', 3), (u'g
ourd', 1), (u'pede', 1), (u'staid', 1), (u'winged', 12), (u'retiring', 3), (u'de
ign', 6), (u'consulship', 1), (u'keenness', 1), (u'threatned', 8), (u'consumst',
1), (u'pajock', 1), (u'predict', 1), (u'incivility', 1), (u'whiles', 79), (u'wag
erd', 3), (u'streak', 1), (u'brazier', 1), (u'dyd', 3), (u'delivery', 4), (u'fra
nchisement', 1), (u'naiads', 1), (u'prime', 17), (u'oerturn', 2), (u'leapfrog',
 1), (u'davys', 2), (u'leather', 11), (u'crabbed', 3), (u'naked', 52), (u'extrem
ity', 27), (u'firmament', 6), (u'exhalst', 1), (u'tradesmans', 1), (u'alltyran
```

Command took 1.86 seconds -- by louhy1128@gmail.com at 2017-6-21 15:58:45 on My Cluster

Cmd 42

```
1 top15WordsAndCounts = wordCount(shakeWordsRDD).takeOrdered(15, key=lambda
  (w,c): -c)
```

▸ (1) Spark Jobs

Command took 1.12 seconds -- by louhy1128@gmail.com at 2017-6-21 16:02:04 on My Cluster

Cmd 43

```
1 print top15WordsAndCounts.collect()
```

```
AttributeError: 'list' object has no attribute 'collect'
```

Command took 0.07 seconds -- by louhy1128@gmail.com at 2017-6-21 16:02:19 on My Cluster

Cmd 44

```
1 print top15WordsAndCounts
```

```
[(u'the', 27361), (u'and', 26028), (u'i', 20681), (u'to', 19150), (u'of', 1746
3), (u'a', 14593), (u'you', 13615), (u'my', 12481), (u'in', 10956), (u'that', 10
890), (u'is', 9134), (u'not', 8497), (u'with', 7771), (u'me', 7769), (u'it', 767
8)]
```

Command took 0.04 seconds -- by louhy1128@gmail.com at 2017-6-21 16:02:33 on My Cluster

Cmd 45

```
1 fileName = "/FileStore/tables/ym8wxbpa1498079196208/98_0-059e3.txt"
2
```

Command took 0.02 seconds -- by louhy1128@gmail.com at 2017-6-21 16:09:24 on My Cluster

Cmd 46

```
1 fileName = "/FileStore/tables/ym8wxbpa1498079196208/98_0-059e3.txt"
2
```

Command took 0.07 seconds -- by louhy1128@gmail.com at 2017-6-21 16:18:03 on My Cluster

Cmd 47

```
1 print fileName
```

```
/FileStore/tables/ym8wxbpa1498079196208/98_0-059e3.txt
```

Command took 0.02 seconds -- by louhy1128@gmail.com at 2017-6-21 16:18:05 on My Cluster

Cmd 48

```
1 extRDD = (sc
2                   .textFile(fileName, 8)
3                   .map(removePunctuation))
```

Command took 0.04 seconds -- by louhy1128@gmail.com at 2017-6-21 16:18:07 on My Cluster

Cmd 49

```
1 print '\n'.join(extRDD
2                 .zipWithIndex()  # to (line, lineNum)
3                 .map(lambda (l, num): '{0}: {1}'.format(num, l))  # to
  'lineNum: line'
4                 .take(15))
```

▸ (2) Spark Jobs

```
0: the project gutenberg ebook of a tale of two cities by charles dickens
1:
2: this ebook is for the use of anyone anywhere at no cost and with
3: almost no restrictions whatsoever  you may copy it give it away or
4: reuse it under the terms of the project gutenberg license included
5: with this ebook or online at wwwgutenbergorg
6:
7:
8: title a tale of two cities
9: a story of the french revolution
10:
11: author charles dickens
12:
13: release date january 1994 ebook 98
14: posting date november 28 2009
```

Command took 0.63 seconds -- by louhy1128@gmail.com at 2017-6-21 16:18:09 on My Cluster

Cmd 50

```
1 extRDD.collect()
```

▸ (1) Spark Jobs

```
u'',
u'1',

u'from fairest creatures we desire increase',
u'that thereby beautys rose might never die',
u'but as the riper should by time decease',
u'his tender heir might bear his memory',
u'but thou contracted to thine own bright eyes',
u'feedst thy lights flame with selfsubstantial fuel',
u'making a famine where abundance lies',
u'thy self thy foe to thy sweet self too cruel',
```

```
u'thou that art now the worlds fresh ornament',
u'and only herald to the gaudy spring',
u'within thine own bud buriest thy content',
u'and tender churl makst waste in niggarding',
u'pity the world or else this glutton be',
u'to eat the worlds due by the grave and thee',
u'',
u'',
u'2',
u'when forty winters shall besiege thy brow',
u'and dig deep trenches in thy beauty's field'
```

Command took 0.63 seconds -- by louhy1128@gmail.com at 2017-6-21 16:14:01 on My Cluster

Cmd 51

```
1 extRDD.count()
```

▸ (1) Spark Jobs

Out[68]: 16272

Command took 0.28 seconds -- by louhy1128@gmail.com at 2017-6-21 16:21:21 on My Cluster

Cmd 52