dataRDD = sc.textFile("file:/home/ubh01/sample.txt")

>>> dataRDD.count()

>>>dataRDD.getNumPartitions()

mapRDD = dataRDD.flatMap(lambda a : a.encode("ascii", "ignore").split(' '))

>>> for word in mapRDD.collect():

...     print(word)

...

mapRDD = dataRDD.flatMap(lambda a : a.split(' '))

>>> keybyword2 = mapRDD.map(lambda word : (word,1))

>>> for line in keybyword2.collect():

...     print(line)

>>> counts = keybyword2.reduceByKey(lambda a,b : a+b).sortByKey()

>>> for line in counts.collect():

...     print(line)

sortbyval = counts.sortBy(lambda a : -a[1]).collect()

>>> for line in sortbyval:

...     print(line)

>>> counts.saveAsTextFile("file:/home/ubh01/pyspark1")

>>> sc.parallelize(sortbyval).saveAsTextFile("file:/home/ubh01/pyspark2")