Spark Cheatsheet

Contents

[Spark Cheatsheet 1](#_Toc440296388)

[Pair RDDs 1](#_Toc440296389)

[Count the occurrences of a word 1](#_Toc440296390)

[Count the number of unique words 1](#_Toc440296391)

[groupByKey() 1](#_Toc440296392)

[reduceByKey() 2](#_Toc440296393)

[Resilient Distributed Datasets (RDDs) 2](#_Toc440296394)

[Extract the first n data points from an RDD 2](#_Toc440296395)

[Number of elements in the RDD 2](#_Toc440296396)

[Read a file into an RDD 3](#_Toc440296397)

# Pair RDDs

## Count the occurrences of a word

In:

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

wordsRDD = sc.parallelize(wordsList, 4)

Out:

['cat', 'elephant', 'rat', 'rat', 'cat']

In:

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

Out:

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

In:

wordCounts = wordPairs.reduceByKey(lambda a,b : a+b)

Out:

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

## Count the number of unique words

In:

uniqueWords = len(wordCounts.collect())

print uniqueWords

Out:

3

## Get an RDD Containing Just the Values of a PairRDD

wordCounts = wordPairs.reduceByKey(lambda a,b: a+b)

print wordCounts.collect()

values = wordCounts.values().collect()

print values

Out:

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

[2, 1, 2]

## groupByKey()

print wordPairs.collect()

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

wordsGrouped = wordPairs.groupByKey()

for key, value in wordsGrouped.collect():

print '{0}: {1}'.format(key, list(value))

rat: [1, 1]

elephant: [1]

cat: [1, 1]

## reduceByKey()

>>> rdd = sc.parallelize([(1,2), (3,4), (3,6)])

>>> rdd.reduceByKey(lambda a, b: a + b)

RDD: [(1,2), (3,4), (3,6)] → [(1,2), (3,10)]

# Resilient Distributed Datasets (RDDs)

## Distinct Elements of the Source Dataset (transformation)

uniqueWords = wordsRDD.distinct()

.collect()

print wordsRDD.collect()

print uniqueWords

output:

['cat', 'elephant', 'rat', 'rat', 'cat']

['rat', 'elephant', 'cat']

## Extract the first n data points from an RDD

samplePoints = rawData.take(n)

## Number of elements in the RDD

numPoints = rawData.count()

print numPoints

## Read a file into an RDD

from test\_helper import Test

import os.path

baseDir = os.path.join('data')

inputPath = os.path.join('cs190', 'millionsong.txt')

fileName = os.path.join(baseDir, inputPath)

numPartitions = 2

rawData = sc.textFile(fileName, numPartitions)