Spark Cheatsheet

Contents

[Spark Cheatsheet 1](#_Toc448475725)

[DataFrames 2](#_Toc448475726)

[Display Data Frame as json 2](#_Toc448475727)

[Display a Spark DataFrame schema 2](#_Toc448475728)

[Pair RDDs 3](#_Toc448475729)

[Count the occurrences of a word 3](#_Toc448475730)

[Count the number of unique words 3](#_Toc448475731)

[Extract the Keys and Values from an RDD 3](#_Toc448475732)

[Return as a list 3](#_Toc448475733)

[Find the n most common words 3](#_Toc448475734)

[groupByKey() 5](#_Toc448475735)

[Keep only unique values in a Key-Value PairRDD 5](#_Toc448475736)

[Sorting by Keys and Values 6](#_Toc448475737)

[Resilient Distributed Datasets (RDDs) 6](#_Toc448475738)

[Extract the first n data points from an RDD 6](#_Toc448475739)

[Filter an RDD 6](#_Toc448475740)

[Get Distinct Elements in an RDD 6](#_Toc448475741)

[Number of elements in the RDD 7](#_Toc448475742)

[Read a file into an RDD 7](#_Toc448475743)

# DataFrames

## Display Data Frame as json

import json

def makeJson(someDf):

aDict = json.loads(someDf)

return json.dumps(aDict, indent=2)

def printResultsAsJson(someDf):

coll = someDf.toJSON().map(makeJson).collect()

for element in coll:

print element

queryResult = sqlContext.sql('''

SELECT smfPayloadData.events.DeliveryLocationArrival

FROM mapped

where smfPayloadData.events.DeliveryLocationArrival is not null

limit 10

''')

# queryResult.show()

printResultsAsJson( queryResult)

## Display a Spark DataFrame schema

# Print the schema in a tree format

df.printSchema()

## root

## |-- age: long (nullable = true)

## |-- name: string (nullable = true)

## Join on two DataFrames

results\_df = sqlContext.sql('''

SELECT mapped.smfPayloadData.paired.shipmentId as shipment\_id,

sap\_idoc.smfPayloadData.SAPShipment.SAPShippingPoint as origin

FROM mapped

INNER JOIN sap\_idoc ON sap\_idoc.smfPayloadData.SAPShipment.SAPShipmentID =

mapped.smfPayloadData.paired.shipmentId

limit 10

''')

print type(results\_df)

results\_df.printSchema()

output:

<class 'pyspark.sql.dataframe.DataFrame'>

root

|-- shipment\_id: string (nullable = true)

|-- origin: string (nullable = true)

# 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

## Extract the Keys and Values from an RDD

keys()

values()

### Return as a list

daysWithHosts = dailyHosts.keys().collect()

## Find the n most common words

endpointCountPairTuple = not200.map(lambda log: (log.endpoint, 1))

endpointSum = endpointCountPairTuple.reduceByKey(lambda a,b : a + b)

# these endpointSum tuples are like

# [(u'/images/NASA-logosmall.gif', 8761), (u'/images/KSC-logosmall.gif', 7236),…]

print endpointSum.takeOrdered(10, lambda s: -1 \* s[1])

## groupByKey()

This outputs a ResultIterable object , which can be cast to a list using a .mapValues() function

dayAndHostTuple = access\_logs.map(lambda log: (log.date\_time.day, log.host))

print dayAndHostTuple.take(3)

groupedByDay = dayAndHostTuple.groupByKey()

print groupedByDay.take(3)

output

[(1, u'in24.inetnebr.com'), (1, u'uplherc.upl.com'), (1, u'uplherc.upl.com')]

[(8, <pyspark.resultiterable.ResultIterable object at 0xb0904bec>), (12, <pyspark.resultiterable.ResultIterable object at 0xb0904a6c>), (4, <pyspark.resultiterable.ResultIterable object at 0xb0902a2c>)]

If you want the value of the PairRDD to be a list, use:

groupedByDay = (dayAndHostTuple.groupByKey()

.mapValues(lambda x: list(x)))

output:

[(8, [u'pc7567.dialup.rwth-aachen.de', u'mtrsppp46.epix.net',…]), (9, […]), …]

## Keep only unique values in a Key-Value PairRDD

Using list(set(x)) converts the iterable result of groupByKey() to a list containing only the distinct values.

dayGroupedHosts = (dayToHostPairTuple.groupByKey()

.mapValues(lambda x: list(set(x))))

## Sorting by Keys and Values

dayHostCount = dayGroupedHosts.mapValues(lambda host\_list: len(host\_list))

# This looks like [(8, 60142), (12, 38070), (4, 59554), (16, 56651), (20, 32963)]

**dailyHosts = dayHostCount.sortByKey()**

print dailyHosts.take(3)

output: [(1, 33996), (3, 41387), (4, 59554)]

Sort by keys (ascending):

RDD.takeOrdered(5, key = lambda x: x[0])

Sort by keys (descending):

RDD.takeOrdered(5, key = lambda x: -x[0])

Sort by values (ascending):

RDD.takeOrdered(5, key = lambda x: x[1])

Sort by values (descending):

RDD.takeOrdered(5, key = lambda x: -x[1])

# pyspark

## pyspark.sql

<https://spark.apache.org/docs/1.6.1/api/python/pyspark.sql.html>

### pyspark.sql.DataFrame

<https://spark.apache.org/docs/1.6.1/api/python/pyspark.sql.html#pyspark.sql.DataFrame>

#### Filter a pyspark.sql.DataFrame

<https://spark.apache.org/docs/1.6.1/api/python/pyspark.sql.html#pyspark.sql.DataFrame>

shipment\_df = mapped\_recent.filter("smfpayloadData.paired.shipmentId = '0305775487'")

print shipment\_df.count()

output: 1

# Resilient Distributed Datasets (RDDs)

## Extract the first n data points from an RDD

samplePoints = rawData.take(n)

## Filter an RDD

The lambda filters in the rows which evaluate to true.

badRecords = (access\_logs.filter(lambda row: row.response\_code == 404))

print 'Found %d 404 URLs' % badRecords.count()

output: Found 6185 404 URLs

## Get Distinct Elements in an RDD

Returns an RDD containing the unique subset

badUniqueEndpoints = badEndpoints.distinct()

## 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)