Using rdd

from \_\_future\_\_ import print\_function

import os.path

from pyspark.sql.functions import split

import sys

from operator import add

from pyspark.sql import SparkSession

from pyspark import SparkContext

filename\_apple = "/FileStore/tables/apple\_daily.csv"

fileRDD\_apple = sc.textFile(filename\_apple)

Joins

joined1 = highs\_apple.join(highs\_yahoo)

Ascending and descending order

numbersCount = numbers.count()

# use take order to order this in ascending order

print(numbers.takeOrdered(numbersCount))

# count in descending order

print(numbers.takeOrdered(numbersCount, key=lambda x:-x))

Describe what the code dose

* Sortby -> orderRdd.reduceByKey(add).sortBy(lambda x: -x[0]), will sort in descending order according to x[0]
* sortByKey -> orderRdd.reduceByKey(add).sortByKey(False), will sort in descending order (i.e. False) with the key being the x[0] by default.
* map -> fileRDD\_yahoo\_data = fileRDD\_yahoo.map(lambda line: line.split(splitit)).filter(lambda row: row[0] != 'Date') , brings out each line as a list
* flatMap -> fileRDD\_yahoo\_data = fileRDD\_yahoo.flatMap(lambda line: line.split(splitit)).filter(lambda row: row[0] != 'Date') brings out each data entry individually
* ReduceByKey -> occurance=wordListRDD.map(lambda x: (x, 1)).reduceByKey(lambda x, y: x + y), will bring together the data in the way that you want, if you are doing a word count it will add up all the entry’s to count the number of words

Create a list

cityListRdd=sc.parallelize(["Leeds", "Liverpool","York","Exeter", "Bath"])

How we do zip

x = sc.parallelize(range(0,5))

y = sc.parallelize(range(1000, 1005))

print(x.zip(y).collect())

What is Val and VAR in Spark?

val means immutable and var means mutable.

Union

append

How to print nth number

nth=6

numbersIs = numbers.take(numbers.count())

print(numbers.collect())

print(numbersIs[nth-1])