Bhushan Kothari

bkkothari@scu.edu

Abstract

Implementation of selection of top-N values using PySpark

Top – N Pyspark

**>>> r1 = sc.textFile(‘Documents/url\_IP’)**

**>>> r1.collect()**[u'url1,2', u'url2,3', u'url3,4', u'url4,65', u'url5,7', u'url6,7', u'url7,6', u'url8,5', u'url9,45', u'url10,2', u'url1,34', u'url2,89', u'url3,56', u'url4,12', u'url5,34', u'url6,56', u'url7,90', u'url8,89', u'url9,67', u'url10,45', u'url1,90', u'url2,89', u'url3,56', u'url4,45', u'url5,23', u'url6,11', u'url7,1', u'url8,4', u'url9,4', u'url10,5', u'url1,6', u'url2,6', u'url3,5', u'url4,3', u'url5,3', u'url6,44', u'url7,5', u'url8,87', u'url9,88', u'url10,57', u'url1,6', u'url2,4', u'url3,6', u'url4,4', u'url5,35', u'url6,56', u'url7,4', u'url8,67', u'url9,34', u'url10,1']

**>>> r2 = r1.map(lambda x:(x.split(',')))  
>>> r2.collect()**[[u'url1', u'2'], [u'url2', u'3'], [u'url3', u'4'], [u'url4', u'65'], [u'url5', u'7'], [u'url6', u'7'], [u'url7', u'6'], [u'url8', u'5'], [u'url9', u'45'], [u'url10', u'2'], [u'url1', u'34'], [u'url2', u'89'], [u'url3', u'56'], [u'url4', u'12'], [u'url5', u'34'], [u'url6', u'56'], [u'url7', u'90'], [u'url8', u'89'], [u'url9', u'67'], [u'url10', u'45'], [u'url1', u'90'], [u'url2', u'89'], [u'url3', u'56'], [u'url4', u'45'], [u'url5', u'23'], [u'url6', u'11'], [u'url7', u'1'], [u'url8', u'4'], [u'url9', u'4'], [u'url10', u'5'], [u'url1', u'6'], [u'url2', u'6'], [u'url3', u'5'], [u'url4', u'3'], [u'url5', u'3'], [u'url6', u'44'], [u'url7', u'5'], [u'url8', u'87'], [u'url9', u'88'], [u'url10', u'57'], [u'url1', u'6'], [u'url2', u'4'], [u'url3', u'6'], [u'url4', u'4'], [u'url5', u'35'], [u'url6', u'56'], [u'url7', u'4'], [u'url8', u'67'], [u'url9', u'34'], [u'url10', u'1']]

**>>> r3 = r2.reduceByKey(lambda x,y:(int(x)+int(y)))  
>>> r3.collect()**[(u'url10', 110), (u'url5', 102), (u'url4', 129), (u'url7', 106), (u'url6', 174), (u'url1', 138), (u'url3', 127), (u'url2', 191), (u'url9', 238), (u'url8', 252)]

**>>> r4 = r3.takeOrdered(3, lambda x: -x[0])  
>>> r4**[(u'url8', 252), (u'url9', 238), (u'url2', 191)]

Another method can be using sortBy() followed by take()

**>>> r4 = r3.sortBy(lambda x:x[1],False)  
>>> r4.collect()**[(u'url8', 252), (u'url9', 238), (u'url2', 191), (u'url6', 174), (u'url1', 138), (u'url4', 129), (u'url3', 127), (u'url10', 110), (u'url7', 106), (u'url5', 102)]

**>>> r5 = r4.take(3)  
>>> r5**[(u'url8', 252), (u'url9', 238), (u'url2', 191)]