CSCI 6515 - Machine Learning for Big Data

Assignment 2 – part 0

Note: Updated ex1.py file and all the generated output files are attached with this document.

Steps for Generating Sample 6 - 1

Transform RDD from Sample 4

1) Mapping the each entry of ((remoteIP, username), frequency) into (username, (remoteIP, frequency)

2) groupByKey -> will merge the values into an ResultIterable object

3) mapValues is used to convert ResultIterable objects in arrays.

4) Saving file to disk by calling saveAsTextFile

Code

```
usernameFrequencyWithIP = remoteIpAndUsernameFrequency\
    .map(lambda ((remoteIP, username), frequency): (username, ((remoteIP, frequency))))\
    .groupByKey()\
    .mapValues(list)
 usernameFrequencyWithIP.saveAsTextFile(os.path.join(output_folder,"s06-1-usernameFrequencyWithIP"))
```

Sample Output

```
(u'test1', [(u'198.204.240.42', 1)])
(u'jjastor', [(u'218.27.204.27', 1)])
(u'test3', [(u'58.58.179.52', 1)])
```

Steps for Generating Sample 6 - 2

1) join RDD#01 with RDD#06-01

   RDD#01 is in form of (username, usernameFrequency)

   RDD#06-01 is in form of (username, [(remoteIP1, frequency1), (remoteIP2, frequency2), ...])

   joining those RDDs will generate RDD#06-02 of form

   (username, (usernameFrequency, [(remoteIP1, frequency1), (remoteIP2, frequency2), ...]))

2) Sorting the RDD by username frequency.

3) Saving file to disk by calling saveAsTextFile

Code

```
usernamejoinedFrequencies = sortedUsernameFrequency.join(usernameFrequencyWithIP)\
        .sortBy((lambda (username, (usernameFrequency, remoteIP_frequency)): usernameFrequency), ascending=False)


usernamejoinedFrequencies.saveAsTextFile(os.path.join(output_folder,"s06-2-usernameJoinedFrequency"))
```

Sample Output

```
(u'root', (221987, [(u'96.57.3.115', 23), (u'87.106.219.6', 1), (u'80.242.123.194', 1), (u'58.58.179.52',
24), (u'58.218.211.166', 330), (u'43.255.191.164', 453), (u'43.255.191.160', 419), (u'43.255.191.149',
7593), (u'43.255.191.134', 5412), (u'43.255.191.133', 102), (u'43.255.191.130', 498),
(u'43.255.190.93', 1785), (u'43.255.190.92', 16803), (u'43.255.190.91', 1608), (u'43.255.190.90',
892), (u'43.255.190.89', 231), (u'43.255.190.191', 12387), (u'43.255.190.190', 279),
(u'43.255.190.189', 1832), (u'43.255.190.188', 1436), (u'43.255.190.187', 582), (u'43.255.190.186',
12051), (u'43.255.190.185', 252), (u'43.255.190.184', 733), (u'43.255.190.183', 796),
(u'43.255.190.182', 1018), (u'43.255.190.181', 717), (u'43.255.190.180', 8943), (u'43.255.190.179',
1094), (u'43.255.190.178', 11090), (u'43.255.190.177', 1277), (u'43.255.190.176', 1444),
(u'43.255.190.175', 14997), (u'43.255.190.174', 806), (u'43.255.190.173', 1313), (u'43.255.190.172',
620), (u'43.255.190.171', 1088), (u'43.255.190.170', 1264), (u'43.255.190.169', 5163),
(u'43.255.190.168', 1012), (u'43.255.190.167', 527), (u'43.255.190.166', 1661), (u'43.255.190.165',
1328), (u'43.255.190.164', 1464), (u'43.255.190.163', 1568), (u'43.255.190.162', 241),
(u'43.255.190.161', 1557), (u'43.255.190.160', 526), (u'43.255.190.159', 802), (u'43.255.190.158',
839), (u'43.255.190.157', 2384), (u'43.255.190.156', 2381), (u'43.255.190.155', 240),
(u'43.255.190.154', 1730), (u'43.255.190.153', 1799), (u'43.255.190.152', 1040), (u'43.255.190.151',
1132), (u'43.255.190.150', 1348), (u'43.255.190.149', 8907), (u'43.255.190.148', 940),
(u'43.255.190.147', 547), (u'43.255.190.146', 703), (u'43.255.190.145', 1144), (u'43.255.190.144',
2101), (u'43.255.190.143', 676), (u'43.255.190.142', 2079), (u'43.255.190.141', 733),
(u'43.255.190.140', 825), (u'43.255.190.139', 1312), (u'43.255.190.138', 1958), (u'43.255.190.137',
1058), (u'43.255.190.135', 933), (u'43.255.190.134', 1370), (u'43.255.190.133', 8644),
(u'43.255.190.132', 912), (u'43.255.190.131', 771), (u'43.255.190.130', 675), (u'43.255.190.126',
1402), (u'43.255.190.125', 1585), (u'43.255.190.124', 551), (u'43.255.190.123', 1025),
(u'43.255.190.122', 1434), (u'43.255.190.121', 793), (u'43.255.190.120', 6679), (u'43.255.190.119',
12282), (u'43.255.190.118', 319), (u'43.255.190.117', 1119), (u'43.255.190.116', 1324),
(u'43.255.190.115', 861), (u'27.112.8.214', 51), (u'222.187.223.214', 816), (u'222.186.129.101', 4),
(u'222.161.4.149', 27), (u'222.161.4.148', 18), (u'222.161.4.147', 24), (u'221.203.3.117', 819),
(u'218.87.111.118', 1180), (u'218.87.111.117', 66), (u'218.87.111.116', 135), (u'218.87.111.110',
1422), (u'218.87.111.109', 519), (u'218.87.111.108', 6), (u'218.87.111.107', 509), (u'218.87.109.62',
465), (u'218.87.109.60', 237), (u'218.65.30.73', 192), (u'218.65.30.61', 240), (u'218.65.30.23', 96),
(u'218.65.30.107', 847), (u'218.27.204.27', 1), (u'218.249.45.57', 1895), (u'198.55.103.208', 1309),
(u'193.104.41.53', 1), (u'186.46.85.2', 3), (u'186.112.230.147', 20), (u'182.100.67.115', 27),
(u'182.100.67.114', 49), (u'182.100.67.113', 708), (u'182.100.67.112', 616), (u'182.100.67.102', 27),
(u'176.96.241.80', 3), (u'144.0.0.200', 54), (u'125.208.3.28', 1), (u'119.147.137.94', 195),
(u'113.98.255.48', 41), (u'113.195.145.12', 18), (u'103.243.138.30', 2743), (u'1.30.20.148', 5)]))
```

```
(u'test', (268, [(u'80.242.123.194', 7), (u'58.58.179.52', 1), (u'27.112.8.214', 3), (u'218.249.45.57',
96), (u'198.204.240.42', 1), (u'190.210.182.225', 62), (u'113.98.255.48', 3), (u'103.243.138.30', 95)]))

(u'nagios', (136, [(u'58.58.179.52', 1), (u'27.112.8.214', 2), (u'218.27.204.27', 1), (u'218.249.45.57',
65), (u'198.204.240.42', 1), (u'113.98.255.48', 1), (u'103.243.138.30', 65)]))

(u'zabbix', (97, [(u'27.112.8.214', 5), (u'218.249.45.57', 46), (u'103.243.138.30', 46)]))

(u'guest', (92, [(u'222.186.129.101', 1), (u'218.249.45.57', 45), (u'176.96.241.80', 1),
(u'103.243.138.30', 45)]))

(u'www-data', (41, [(u'218.249.45.57', 20), (u'198.204.240.42', 1), (u'103.243.138.30', 20)]))

(u'zxin10', (40, [(u'218.249.45.57', 20), (u'103.243.138.30', 20)]))

(u'apache', (27, [(u'218.249.45.57', 13), (u'198.204.240.42', 1), (u'103.243.138.30', 13)]))

(u'zhaowei', (24, [(u'218.249.45.57', 12), (u'103.243.138.30', 12)]))
```