Homework

2. Spark Homework (continued)

3. RDD operations

a. Count the number of requests from each user and save the result to an RDD named: `setupCountsRDD`

You'll need to use the user ID field - it is the third field in each line of the weblogs data.

Hint: Create a Pair RDD and use the WordCount approach covered in the RBDA course.

Your data will look something like this:

(useridA, 1)
(useridB, 1)
(useridA, 1)

b. Sum the values for each user ID and save the result to an RDD named: `requestCountsRDD`

Hint: Your RDD data will look something like this:

(useridA, 5)
(useridB, 7)
(useridC, 5)

c. Determine how many users visited once, twice, three times, etc. and save the result to an RDD named: `visitFrequencyTotalsRDD`

Generate data in this format: frequency:user-count pairs

Hint: The data shown in b. above produces the following -

(5:2)
(7:1)

d. Create an RDD where the user id is the key, and the value is the list of all the IP addresses that the user has connected from. Save the result to an RDD named: `validAcctsIpsFinalRDD`

You will need the accounts data in order to only output ip addresses for user IDs that appear in the accounts files.

Ensure that the output only contains user IDs of actual customers.

Input example: (useridX, 20.1.34.55)
(useridX, 74.125.239.98)

becomes: (useridX,List(20.1.34.55, 74.125.239.98))