Dear Participant,

Please find below the Lab for Big Data on Spark course. Kindly make your submission before the deadline.

This Lab is of 10 marks where each question carries 2 marks.

**Problem Statement**

In this Lab, we will parse Apache logs to get some meaningful insights from the logs.

**Data set -**

File Name - access

HDFS Path - "home/gldata/ NASA/access"

Above dataset is access log of NASA Kennedy Space Center WWW server in Florida. The logs are an ASCII file with one line per request, with the following columns:

1. **host**- making the request. A hostname when possible, otherwise the Internet address if the name could not be looked up.
2. **timestamp**- in the format "DAY MON DD HH:MM:SS YYYY", where DAY is the day of the week, MON is the name of the month, DD is the day of the month, HH:MM:SS is the time of day using a 24-hour clock, and YYYY is the year. The timezone is -0400.
3. **request url**- Request URL.
4. **HTTP reply code**
5. **bytes returned by the server**

Note that from 01/Aug/1995:14:52:01 until 03/Aug/1995:04:36:13 there are no accesses recorded, as the Web server was shut down, due to Hurricane Erin.

Based on the above data, please answer following questions:

**Q1: Write spark code( using RDD) to find out top 10 requested URLs along with count of number of times they have been requested (This information will help company to find out most popular pages and how frequently they are accessed)**

**Sample output:**

URL                                                              Count

shuttle/missions/sts-71/mission-sts-71.html     549

shuttle/resources/orbiters/enterprise.html        145

**Q2: Write spark code to find out top 5 hosts / IP making the request along with count (This information will help company to find out locations where website is popular or to figure out potential DDoS attacks)**

**Sample output:**

URL                    Count

192.168.78.24     219

**Q3: Write spark code to find out top 5 time frame for high traffic (which day of the week or hour of the day receives peak traffic, this information will help company to manage resources for handling peak traffic load)**

**Sample Output:**

|     timeFrame    |req\_cnt|

|31/Nov/1995:10|   50|

**Q4: Write spark code to find out 5 time frames of least traffic (which day of the week or hour of the day receives least traffic, this information will help company to do production deployment in that time frame so that less number of users will be affected if some thing goes wrong during deployment)**

**Sample Output:**

|     timeFrame    |req\_cnt|

|31/Nov/1995:11|   5000|

**Q5: Write spark code to find out unique HTTP codes returned by the server along with count (this information is helpful for devops team to find out how many requests are failing so that appropriate action can be taken to fix the issue)**

**Sample output:**

HTTP code            Count

200 - 15400 404    324

***Hint -***

PATTERN = '^(\S+) (\S+) (\S+) \[([\w:/]+\s[+\-]\d{4})\] "(\S+) (\S+)(.\*)" (\d{3}) (\S+)'

#Example line

line = 'slppp6.intermind.net - - [01/Aug/1995:00:00:39 -0400] "GET /history/skylab/skylab-logo.gif HTTP/1.0" 200 3274'

def parseLogLine(log):

m = re.match(PATTERN, log)

if m:

return [Row(host=m.group(1), timeStamp=m.group(4),url=m.group(6), httpCode=int(m.group(8)))]

else:

return []

print(parseLogLine(line))