Linux Tutorial: Assignment 1

Due date: June 10, 2019

You can work in groups of up to 2 persons

Assignment description

Part 1: Answer the questions and write your answers into ***"answers.txt"*** file.

Part 2: Follow the steps to perform basic dataset analysis and transformations. All work should be done on the ***lnx.cs.smu.ca*** server.

As a result, you should get an **archive with files** from the steps.

Also, you should create a **text file** ***"answers.txt"*** with commands that you have used on **each step** along with **answers and/or command output (if applicable)**. The text file should look like this:

```
Part 1:

   1. answer1

   2. answer2

   .

   .

Part 2:

   1. ls -l
   2. cat textfile.txt | grep ABC > output.txt
      wc -l output.txt
      answer: 35
   3. rm output.txt

   .

   .

   10. touch file1
       answer: 2339
```

Please submit both the **arc.tar.xz** and **answers.txt** via Brightspace. The **arc.tar.xz** file should be in your home directory as well. If it's not found in your home directory, you'll get lower marks.

In case of any questions please send an email to **nikita.neveditsin@smu.ca**

Part 1: warm-up questions (max 15 pts)

1. Can ~ be equal to . ? If yes then give an example
2. Can ~ be equal to .. ? If yes then give an example
3. Can . be equal to .. ? If yes then give an example
4. Can ~ be equal to /? If yes then give an example

Part 2: practice (max 85 pts)

1. Download a dataset from *https://archive.ics.uci.edu/ml/machine-learning-databases/00235/household_power_consumption.zip*

Dataset attribute Information:

1.date: Date in format dd/mm/yyyy
2.time: time in format hh:mm:ss
3.global_active_power: household global minute-averaged active power (in kilowatt)
4.global_reactive_power: household global minute-averaged reactive power (in kilowatt)
5.voltage: minute-averaged voltage (in volt)
6.global_intensity: household global minute-averaged current intensity (in ampere)
7.sub_metering_1: energy sub-metering No. 1 (in watt-hour of active energy). It corresponds to the kitchen, containing mainly a dishwasher, an oven and a microwave (hot plates are not electric but gas powered).
8.sub_metering_2: energy sub-metering No. 2 (in watt-hour of active energy). It corresponds to the laundry room, containing a washing-machine, a tumble-drier, a refrigerator and a light.
9.sub_metering_3: energy sub-metering No. 3 (in watt-hour of active energy). It corresponds to an electric water-heater and an air-conditioner.

2. Uncompress the archive to get a text file
3. Get the exact number of records in the dataset (excluding header)
4. Show the top 10 (with **max** values) records by `global_active_power`. Write the output to a file **gap_max.txt** (and attach output in the *answers.txt*)
5. Replace the semicolons by commas in the original file (household_power_consumption.txt) to get a .csv file. Save it as **hpc.csv** (and attach output in the *answers.txt*)
6. Transform the dataset **hpc.csv** to get a new dataset with the following fields:
    - `year`
    - `global_active_power`
    - `global_reactive_power`
    - `global_intensity`

    The new dataset should look like this:

    ```
    Year,Global_active_power,Global_reactive_power,Global_intensity
    2006,4.216,0.418,18.400
    2006,5.360,0.436,23.000
    2006,5.374,0.498,23.000
    ```

    Save the resulting dataset as **transformed.csv** (and attach output in the *answers.txt*)

7. Find all records from transformed.csv where `Global_active_power = 2.042` like this:

   ```
   2006,2.042,0.090,8.400
   2006,2.042,0.066,9.000
   2006,2.042,0.000,8.800
   2006,2.042,0.182,8.600
   2006,2.042,0.156,9.200
   2007,2.042,0.000,8.400
   ```

   Write it to file **2p042.csv** (and attach output in the *answers.txt*)

   How many records are there? (write your answer to *answers.txt*)

8. Create a new file **tr_no2007.csv** as a copy of the **transformed.csv** without records from the year of 2007
9. Remove the header from the **tr_no2007.csv**, shuffle it and split into 3 files. Make sure that output files are correct (sum of number of lines in output files equals to number of lines in **tr_no2007.csv** and lines are intact)
10. Create an archive with name **arc.tar.xz** using LZMA/LZMA2 compression with the following files in the archive:
    1. gap_max.txt
    2. hpc.csv
    3. transformed.csv
    4. 2p042.csv
    5. tr_no2007.csv
    6. Three output files from step 9