Data Proc 2024: Assignment 08

- Due: 13 Dec 2024 (before class: 1315 PM)
- Complete the assignment on the server using your user account.
  - You must create the scripts in the correct location in your home directory.

Assignment 08: Goals

- Regular expressions in:
  - grep
  - sed
  - python
  - (extra credit: find)

References for HW 8

- See lecture slides for lecture 8 (regex)

Copy files from $DP24/assignments/08

to your home: $HOME/dataproc2024/assignments/08

```
riveale@dataproc2023vm:/usr/share/dataproc/2024/assignments/08$ ls
exfile2.txt  problem1.sh  problem3.sh  problem5.py
problem0.sh  problem2.sh  problem4.sh
```

Assignment 08

- Create a directory in your home:
  - $HOME/dataproc2024/assignments/08/
- You will complete
  - $HOME/dataproc2024/assignments/08/problem0.sh
  - $HOME/dataproc2024/assignments/08/problem1.sh
  - $HOME/dataproc2024/assignments/08/problem2.sh
  - $HOME/dataproc2024/assignments/08/problem3.sh
  - $HOME/dataproc2024/assignments/08/problem4.sh
  - $HOME/dataproc2024/assignments/08/problem5.py

Problem 0 (problem0.sh)

```
File Edit Options Buffers Tools Sh-Script Help
#!/bin/bash

## Problem 0 for HW 8
## Regular expression (grep)
## Make this regular expression grep only print lines of students
## at Kyoto University
## (it currently will print all students in the input file exfile2.txt)

## NOTE EXFILE2.TXT HAS HEADER REMOVED!!!
grep -E '^[a-zA-Z]+,[a-zA-Z]+,[0-9]+,[a-zA-Z0-9]+,[a-zA-Z ]+$' exfile2.t\
xt

-UU-:---  F1  problem0.sh     All   L1      (Shell-script[bash]) ----------
```

Problem 1 (problem1.sh)

```
File Edit Options Buffers Tools Sh-Script Help
#!/bin/bash

## Problem 1 HW 8
## RegExp (grepping)

#Print only lines of students whose family names are longer than 7 lette\
rs in input file exfile2.txt (right now it recognizes all students)
grep -E '^[a-zA-Z]+,[a-zA-Z]+,[0-9]+,[a-zA-Z0-9]+,[a-zA-Z ]+$' exfile2.t\
xt
-UU-:---  F1  problem1.sh     All   L1      (Shell-script[bash]) ----------
```

Problem 2 (problem2.sh)

```
#!/bin/bash

## Change this to print only lines of people whose FAMILY NAME begins
## with an 'S' from an inputfile formatted like exfile2.txt

## It should read the first command line argument as the input file to r\
ead (e.g. bash problem2.sh newfilename.txt)

grep '#### FILL IN HERE #####'  FILLHERE
```

```
-UU-:---   F1  problem2.sh     All   L1      (Shell-script[bash]) ----------
```

Problem 3 (problem3.sh)

```
File Edit Options Buffers Tools Sh-Script Help
#!/bin/bash

## Fix the below SED command to
## Replace all spaces " " of input file exfile2.txt
## with underscores "_"

## E.g. Kyoto University -> Kyoto_University

sed -E 's/ / /g' exfile2.txt

-UU-:---  F1  problem3.sh     All   L1      (Shell-script[bash]) ----------
```

Problem 4 (problem4.sh)

```
File Edit Options Buffers Tools Sh-Script Help
#!/bin/bash

## HW 8 problem 4 (sed regex)

## Fix the following SED command to remove the initial letter
## in some student IDs (i.e. B3039200291 -> 3039200291;
## a3810202901 -> 3810202901)

## Note sed is not grep, if it does not match the pattern, it simply
## prints as-is...

## Here is a SED command to print the same thing back directly.
## Hint: modify what is in one of the capture groups (XXX)
sed -E 's/^([a-zA-Z]+,[a-zA-Z]+,[0-9]+,)([a-zA-Z0-9]+)(,[a-zA-Z ]+)$/\1\\
2\3/g' exfile2.txt


-UU-:---   F1  problem4.sh     All   L1      (Shell-script[bash]) ----------
```

Problem 5 (problem5.py)

```
## PYTHON SCRIPT
## HW8 problem 5

## Read in a file with format like exfile2.txt
## Use regular expressions to print only the last names to standard out
## (one per line) of EVERY PERSON (not just students)

import re
import sys
# HINT: this matches only students...
repattern = '^[a-zA-Z]+,[a-zA-Z]+,[0-9]+,[a-zA-Z0-9]+,[a-zA-Z ]+$';

with open(infname, 'r') as inf:
    for line in inf:
        result = re.match(repattern, line);

        #HINT: I already specify that you will capture some group
        ## (the last name), and print it here...
        print(result.group(1));
    pass;
```