Research Software Engineer, IT

`ITRES1001`, University of Leeds

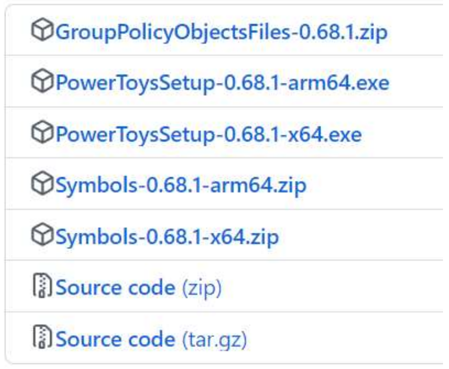
Problem 1: Batch / Bash Re-naming files

- 1. Use PowerToy (https://github.com/microsoft/PowerToys/releases/tag/v0.68.1).
- Microsoft official instructions can be found at https://learn.microsoft.com/en-us/windows/powertoys/powerrename

A summary

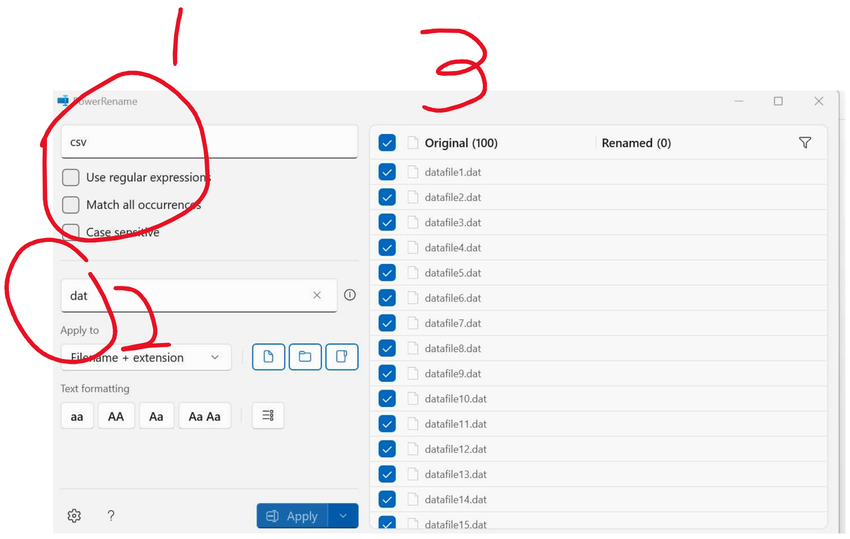
1. Go to the PowerToy Github download the executable (remember to check the hash, if the user has a concern on security).

- Windows users can choose PowerToysSetup-x.xx.x-x64.exe.
- Mac and Linux users can choose Source code (zip or tar.gz).



GroupPolicyObjectsFiles-0.68.1.zip
PowerToysSetup-0.68.1-arm64.exe
PowerToysSetup-0.68.1-x64.exe
Symbols-0.68.1-arm64.zip
Symbols-0.68.1-x64.zip
Source code (zip)
Source code (tar.gz)

1. Enter the string, a user wants to replace. Here it is “csv” (See 1).
2. Enter the target string that the user wants to change to. Here is it “dat” (See 2).
3. The result is showed in 3.
4. The task can also be done by using Windows cmd (or powershell) bash script, Python script and R script.



csv

Use regular expressions

Match all occurrences

Case sensitive

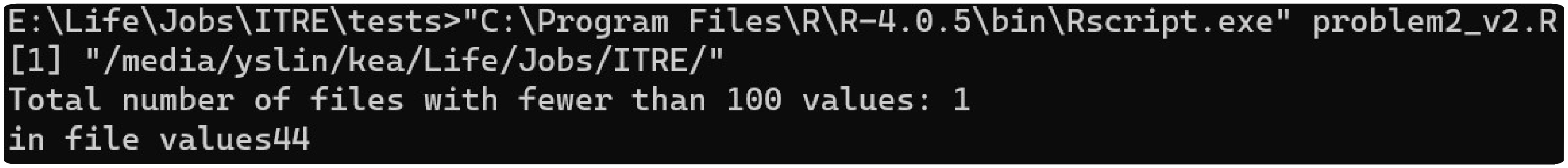
dat

Apply to: Filename + extension

Text formatting

Original (100) | Renamed (0)

datafile1.dat
datafile2.dat
datafile3.dat
datafile4.dat
datafile5.dat
datafile6.dat
datafile7.dat
datafile8.dat
datafile9.dat
datafile10.dat
datafile11.dat
datafile12.dat
datafile13.dat
datafile14.dat
datafile15.dat



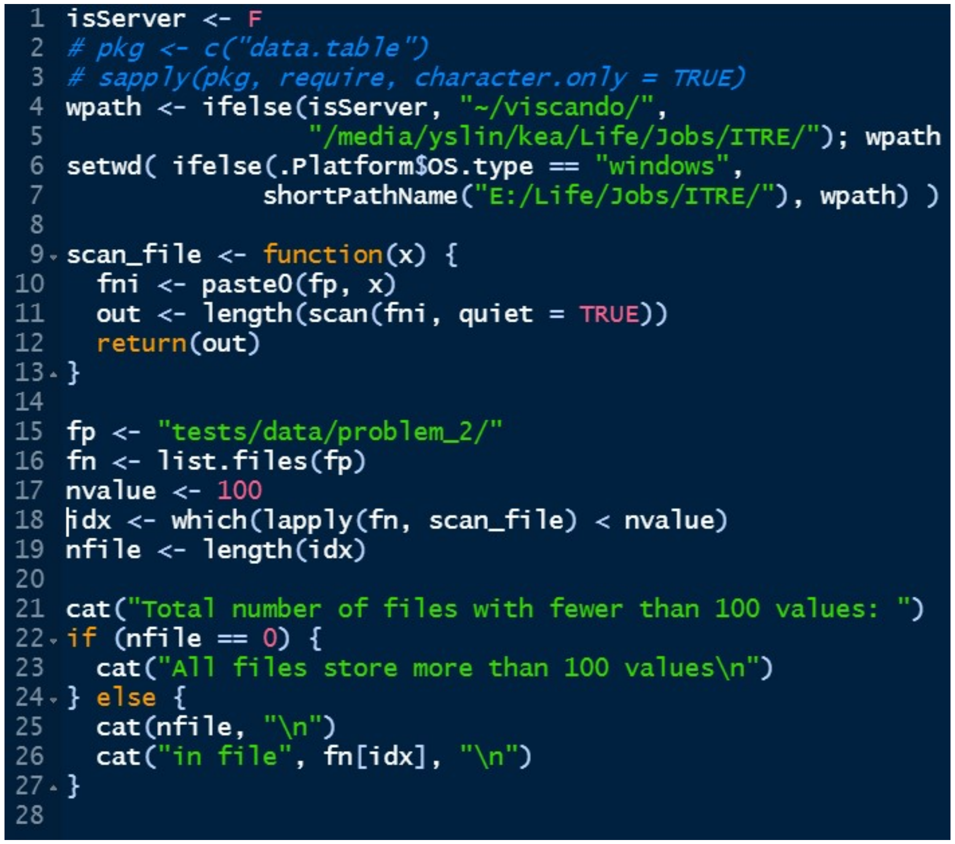
```
E:\Life\Jobs\ITRE\tests>"C:\Program Files\R\R-4.0.5\bin\Rscript.exe" problem2_v2.R
[1] "/media/yslin/kea/Life/Jobs/ITRE/"
Total number of files with fewer than 100 values: 1
in file values44
```

Problem 2: Examine files / their contents filling some criteria

- R solution. Execute the script, “problem2_v2.R”, using the tool, Rscript.exe

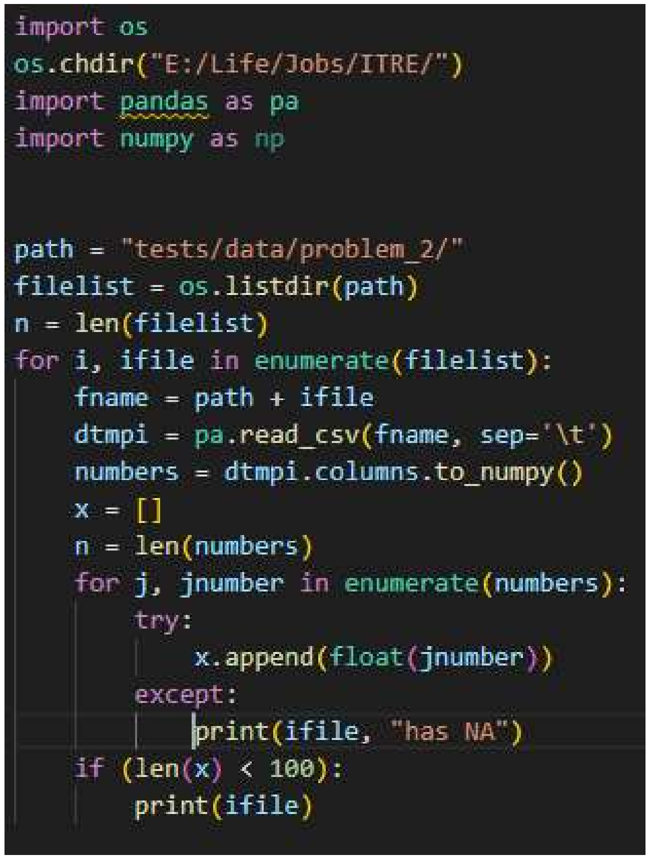
Details

- Line 1 to 7 sets up the working directory, depending on whether the user is on a server, a Windows or a Linux (local) environment.
- Line 9 to 13 is a function to scan over all the data files.
- Line 15 to 19 locates the data and applies the scan_file function to find the individual file. Line 21 to 27 reports the result.
- The variable nvalue sets the expected number of value (in this case 100).



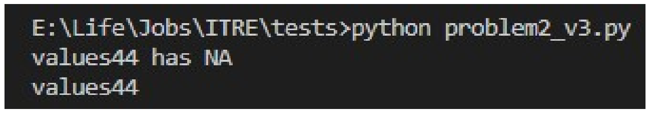
```
 1  isServer <- F
 2  # pkg <- c("data.table")
 3  # sapply(pkg, require, character.only = TRUE)
 4  wpath <- ifelse(isServer, "~/viscando/",
 5                  "/media/yslin/kea/Life/Jobs/ITRE/"); wpath
 6  setwd( ifelse(.Platform$OS.type == "windows",
 7               shortPathName("E:/Life/Jobs/ITRE/"), wpath) )
 8
 9  scan_file <- function(x) {
10    fni <- paste0(fp, x)
11    out <- length(scan(fni, quiet = TRUE))
12    return(out)
13  }
14
15  fp <- "tests/data/problem_2/"
16  fn <- list.files(fp)
17  nvalue <- 100
18  idx <- which(lapply(fn, scan_file) < nvalue)
19  nfile <- length(idx)
20
21  cat("Total number of files with fewer than 100 values: ")
22  if (nfile == 0) {
23    cat("All files store more than 100 values\n")
24  } else {
25    cat(nfile, "\n")
26    cat("in file", fn[idx], "\n")
27  }
28
```

Python solution



```python
import os
os.chdir("E:/Life/Jobs/ITRE/")
import pandas as pa
import numpy as np


path = "tests/data/problem_2/"
filelist = os.listdir(path)
n = len(filelist)
for i, ifile in enumerate(filelist):
    fname = path + ifile
    dtmpi = pa.read_csv(fname, sep='\t')
    numbers = dtmpi.columns.to_numpy()
    x = []
    n = len(numbers)
    for j, jnumber in enumerate(numbers):
        try:
            x.append(float(jnumber))
        except:
            print(ifile, "has NA")
    if (len(x) < 100):
        print(ifile)
```



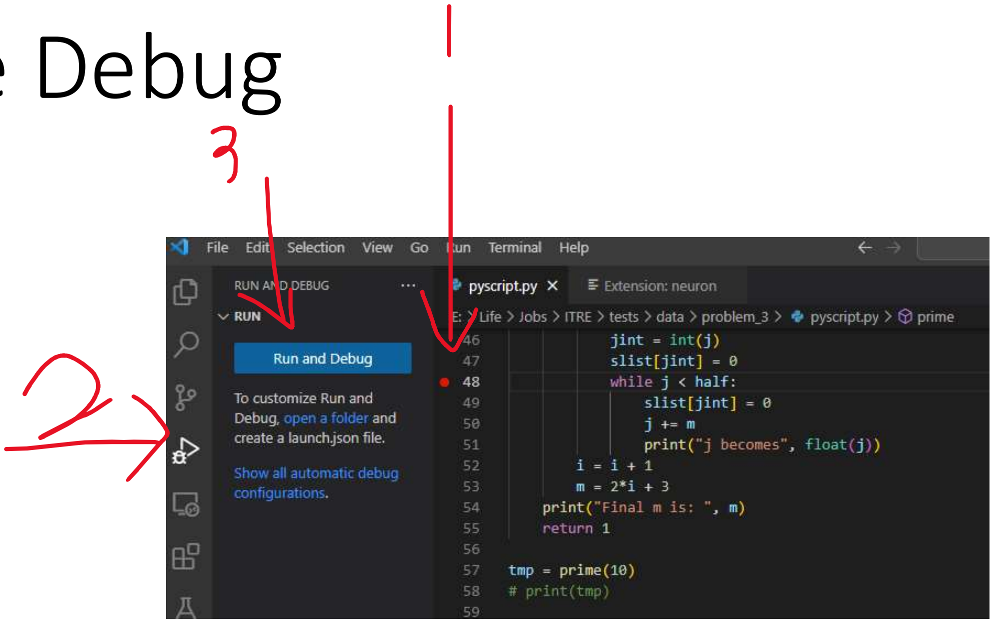
```
E:\Life\Jobs\ITRE\tests>python problem2_v3.py
values44 has NA
values44
```

Problem 3

1. Python is a scripting language, so its background engine reads and executes its code line by line (essentially, but not exactly; Python3 use bytecode JIT, too).
2. The while loop (esp. the inner one, “while j < half:”) increases the line that Python must read and execute. When the input (i.e., *n*) increases (e.g., 10 to 100), the *prime* function must repeatedly execute *the while loop and anything inside (e.g., from once to 8 times)*
3. *One way to find out the step-by-step process is to use debug tools.*

Using VS Code Debug

1. Use a mouse click on the line where the second while loop starts. A small red dot should show up to mark the debugging point.
2. Click the icon on 2.
3. Click the “Run and Debug” to start debugging (and maybe a Enter after).

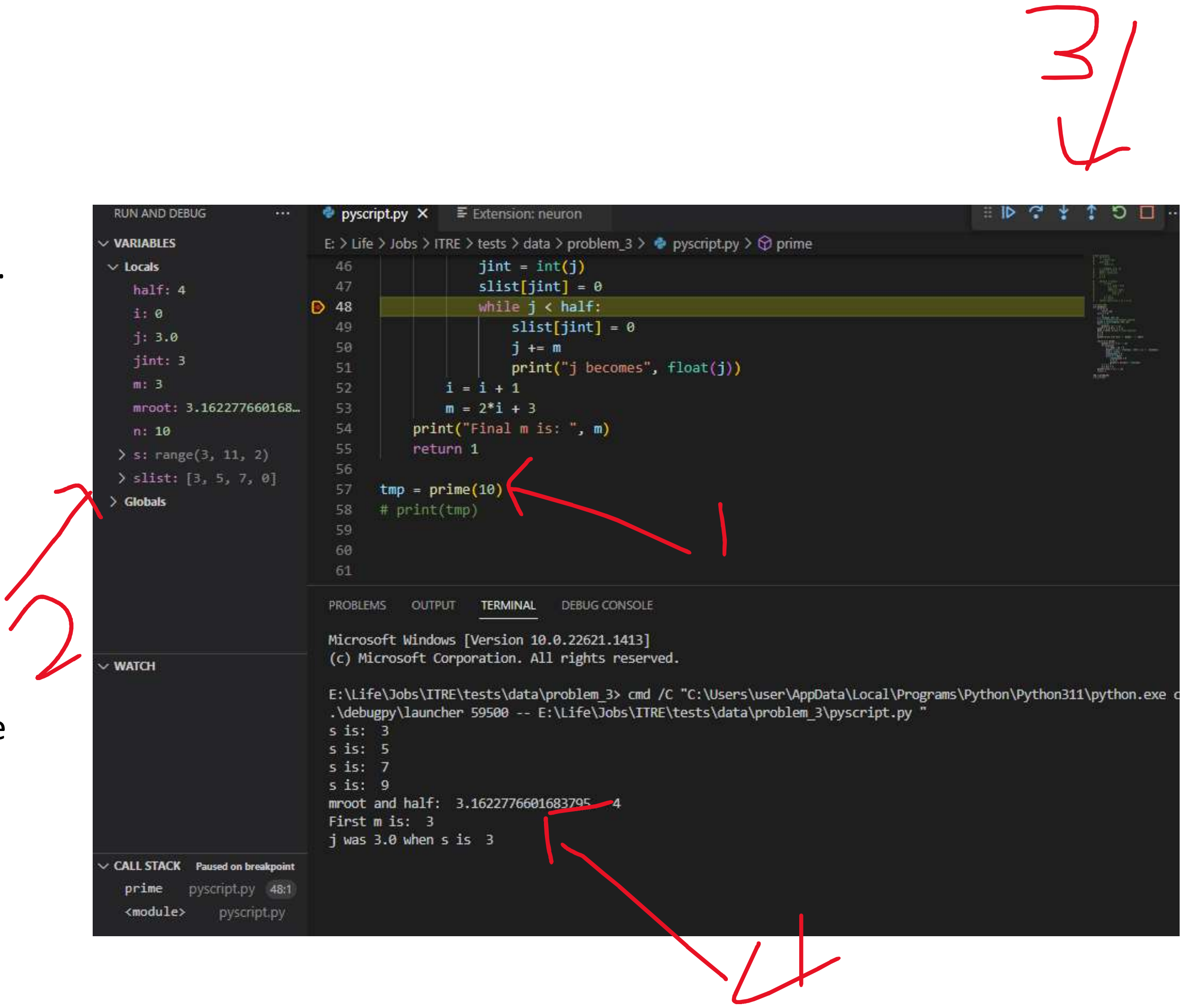

3
2
File Edit Selection View Go Run Terminal Help
RUN AND DEBUG
RUN
Run and Debug
Show all automatic debug configurations.
pyscript.py
Extension: neuron
E: > Life > Jobs > ITRE > tests > data > problem_3 > pyscript.py > prime

To customize Run and Debug, open a folder and create a launch.json file.

```
            jint = int(j)
            slist[jint] = 0
            while j < half:
                slist[jint] = 0
                j += m
                print("j becomes", float(j))
        i = i + 1
        m = 2*i + 3
    print("Final m is: ", m)
    return 1

tmp = prime(10)
# print(tmp)
```

1. Increase the input from, e.g., 10, to a larger number, e.g., 30.
2. Watch the change every time Python executes the while loop.
3. Use the clickable buttons to go over the loop. The user should / might find out that the larger the input, the longer and more clicks, she needs to step over the loop.
4. An alternative method is to use the traditional printing method.



```
half: 4
i: 0
j: 3.0
jint: 3
m: 3
mroot: 3.162277660168…
n: 10
s: range(3, 11, 2)
slist: [3, 5, 7, 0]
```

```python
            jint = int(j)
            slist[jint] = 0
            while j < half:
                slist[jint] = 0
                j += m
                print("j becomes", float(j))
        i = i + 1
        m = 2*i + 3
    print("Final m is: ", m)
    return 1

tmp = prime(10)
# print(tmp)
```

```
Microsoft Windows [Version 10.0.22621.1413]
(c) Microsoft Corporation. All rights reserved.

E:\Life\Jobs\ITRE\tests\data\problem_3> cmd /C "C:\Users\user\AppData\Local\Programs\Python\Python311\python.exe c
.\debugpy\launcher 59500 -- E:\Life\Jobs\ITRE\tests\data\problem_3\pyscript.py "
s is:  3
s is:  5
s is:  7
s is:  9
mroot and half:  3.1622776601683795  4
First m is:  3
j was 3.0 when s is  3
```

```
prime     pyscript.py  48:1
<module>  pyscript.py
```

Problem 3 – Requirement Analysis

- Costing questions
  - How many individual in this group would be? (a rough estimate with a confident range)
  - Both Facebook and Twitter? Or either Facebook or Twitter?
  - Can the client provide his/her own text parsing tools / results? Are customised tools needed for automated processing?
  - How many (minimal) features are needed to extract from these comments, if automated processing is chosen?
  - Whether these features are well established? Or yet-to-be established and need technical assistant to establish? (The former costs less than latter)
  - For example, separately extract and analyze prepositions and particles, like “the”, “a”, “and” etc from the subjects, verbs, nouns etc.
  - Does the client require assistance of statistical analyses? Descriptive or inferential statistics or both?

- Advance costing questions
  - If the client also want to have inference statistics? Frequentist or Bayesian approach or both?
  - Involving complex computational statistics? For example, bootstraping, jackknife, Markov chains, and Bayesian inference, quantitative modelling etc.
  - Involving to build or tweak machine-learning model(s)?
  - Whether data visusalization is needed? Basic and insight visualization?
  - How much and how long relevant computation resources are needed to allocate to the project? CPU (and/or GPU), memory and disk space.
  - Is there a preference in working (or the format of results) environment, Windows, MacOS, Linux or other operation systems?

Not requirement but very important questions

- Legal questions
  - Does the client need to acquire (or purchase) the service providers’ (i.e. Facebook and Twitter) permission (hence perhaps some API) to conduct research there?
- Privacy questions
  - Has the client clear regional privacy issues both in UK and in Indian and has appropriate measures to protect the potential individual’s rights?
- Ethics questions:
  - Has the client acquired / applied for research ethics reviews in UK and Indian?
  - Do the ethics committees the client use apply a good standard of research practice?
- If mostly answers to the above are unclear or No, can the service of the research software engineer(s) be exempted from any legal liability?

Problem 5 – Question 1

- Can you determine the content of the fname_out variable?
  - Yes. The context of *fname_out is from a concatenation of three strings: Home*, *MetaData*, *And .png*
  - *The first two are two variables. Home was not shown in the script, so my best guess is that it was set up in another script, not shown but in the same script, in a user-defined package, or a global variable. It probably is a string indicating the user's home directory.*
  - *MetaData is also a concatenation of four strings: VarName, TimeStamp, Frame and "_".*

- VarName and TimeStamp are given.
- Frame is converted from a variable, storing possibly simulation hour (as simhour), which is perhaps an integer or a floating point number.
- Frame variable is to take four digits before the decimal mark.

Question 2

- What information is missing?
- The variables, simhour and Home, was not shown in the script.