KED2020 Exercise 1: Data Wrangling

Alex Flückiger

2020

Requirements

- Deadline: 26 March 2020, 23:59
- File format: executable shell script
- Naming schema: *SURNAME*_KED2020_ex_1.sh
  Replace *SURNAME* with your surname.
- Use the shell template provided here
- All tasks require shell commands unless stated otherwise.
- Please hand in your solutions via the respective exercise module on OLAT. Keep in mind to submit on time, as the module is only open until midnight.
- Find solutions individually. Ask friends or Google whenever you feel lost (in terms of programming, Google may be your best friend).

Organize your project

In this first task, you don't need to provide any interpretation, only the raw commands.

As the project grows over time, it is crucial to organize your work properly. Otherwise, you get lost or waste too much time to find a particular file.

1. Create a new project folder with the following name:
   `KED2020_exercise_1`
2. Where have you put your project folder? Show the absolute path for it.
3. In this folder, you make the following subfolders:
   `reports`, `src`, `data`, `data/raw`, `data/interim`
4. Create a new file called `script.sh` in your project folder. Copy the bash template from the KED2020 website and change the content as needed in this exercise. This file will be your final submission. Thus, you need to add the commands from the previous steps (1-3) to this script. For copying the template, you may use a text editor instead of the shell as this is a more advance operation.

5. In your project, you may have thousands of text files, which are named inconsistently. To simulate this, create empty files with the following commands:

   ```
   touch data/raw/speeches_{2015..2020}_{a..z}.txt
   touch data/raw/text_{2015..2020}_{1..30}.txt
   ```

   Don't forget to add these commands in your script.

6. The raw data should never be modified directly. Thus, create folders for each year (2015-2020) in `data/interim`. Copy the created `.txt` files from above into the folder with the corresponding year. Specifically, a file that has 2020 in its name needs to go into the directory `2020`. Hint: Recall the expansion and wildcard operations.

7. Rename the file from `script.sh` to `SURNAME_KED2020_ex_1.sh` and move it into `src`. By the way, `src` is a short name for *source*, meaning *source code*.

Beyond this toy project, you may want to learn more about how to organize your research project. The cookie cutter website is a great resource that provides useful information on the reasonable organization of your data science project.

Make report of your file

In this second task, please give a short explanation accompanying your command.

What files do you have on your computer? Let's create some reports. You are free to choose the location and names for the files. Yet, please recall the conventions that help others to understand the purpose of your scripts and outputs.

1. Count the total number of all `.pdf`, `.txt`, and `.docx` files on your computer and write this number into a new file using operators. You may want to check the man page.
2. Write a single command (piping) to get the file names of the 20 oldest `.pdf` in your document folder, including any subfolder, and write them into a new file.

Test your script

This task is a a simple sanity check. You don't need to include it in your submission.

Your deliverable has to be a runnable script comprising all the commands to accomplish the tasks above. Once you start the script again, it executes all the commands, one after another. To test it, you can create a new folder, copy

the script into that folder, and execute it. Everything should be reconstructed accordingly. If not, correct the script so that it runs through without any issue.

Feedback

1. Do you have any questions concerning the exercise or the commands?
2. How long did it take to solve this exercise? Give a fair estimation.