KED2021 Assignment 1: Data Wrangling

Alex Flückiger | University of Lucerne

18 März 2021

Requirements

- Deadline: 25 March 2021, 23:59
- File format: executable shell script
- Naming schema: `SURNAME_KED2021_1.sh`
  Replace `SURNAME` with your surname.
- Use the shell template provided here.
- All tasks require shell commands unless stated otherwise.
- Submit your solutions on time via the respective exercise module on OLAT. The module is only open until midnight.
- Find solutions individually. When you are stuck, post your issue in the OLAT forum and ask friends. In terms of programming, Google may be your best friend.

Motivation

You learn how to perform basic shell commands and wrap them into a script to reproduce all steps.

Use your favorite text editor to assemble your script (e.g., `Atom`) . You may want to try out the commands directly in your shell and, after successfully running them, copy them over into your shell file. The command `history` shows the history of all used commands.

A template for your shell script can be found here.

1 Organize your project

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

As a 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:
   `KED2021_exercise_1`
2. Where did you create your project folder? In addition to the command, write the absolute path as `#comment` into your script.
3. In this folder, you make the following subfolders:
   `reports`, `src`, `data`, `data/raw`, `data/interim`
4. In a 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..2021}_{a..z}.txt
   touch data/raw/text_{2015..2021}_{1..12}_{1..30}.txt
   ```

   Don't forget to add these commands to your script.
5. The raw data should never be modified directly. Thus, create folders for each year (2015-2021) in `data/interim`. Copy the created `.txt` files from above into the folder with the corresponding year. Specifically, a file that has 2021 in its name goes into directory `2021`. Hint: Recall the expansion and wildcard operations.

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 reasonable organization of your data science project.

2 Report on file collection

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 may choose the directory and name for your output 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 the resulting number into a new file using operators. You may want to use the `wc` command in addition to a search command.
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.

3 Test your script

This task is a simple sanity check for your script. Your script has to pass this test, yet you don't need to include it in your submission.

Your deliverable has to be a runnable script comprising all commands to accomplish the tasks above. To test your script, run the commands below. Once you call the script, it executes all commands, one after another. Everything should be reconstructed accordingly in the test folder. If not, correct the script so that it runs without any issue.

```
mkdir test_script
cd test_script
bash PATH/SCRIPT_NAME.sh    # e.g. bash ../flueckiger_KED2021_1.sh
cd ..
rm -r test_script
```

4 Feedback

Please answer the following questions at the end of your script. Start your answers with the # symbol to make them comments that are ignored when running the script.

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.