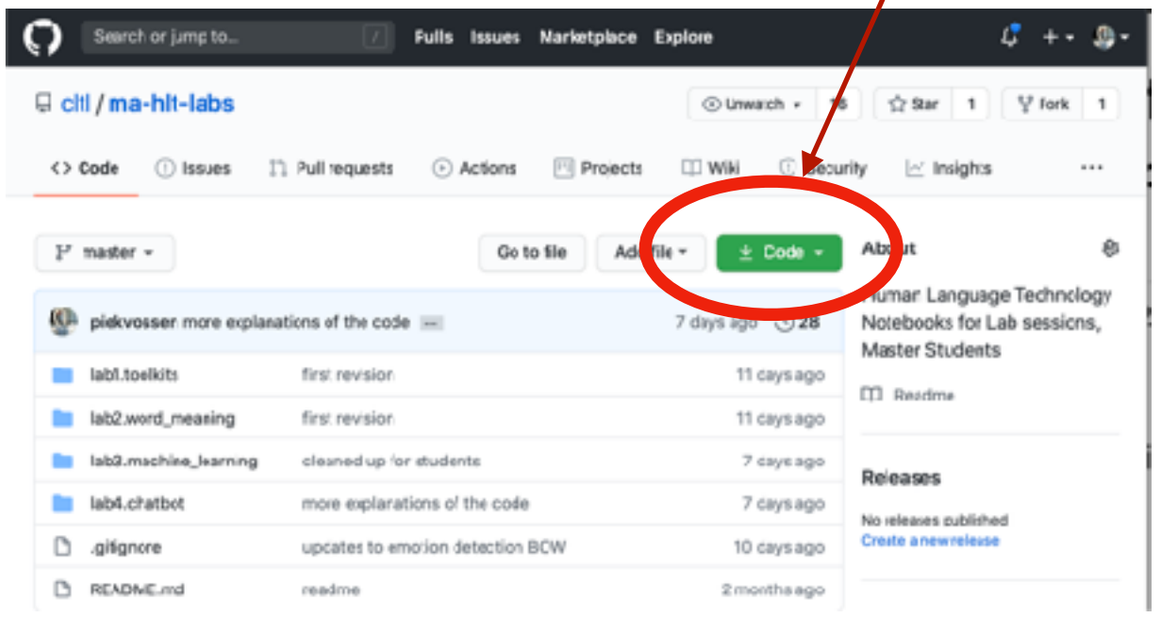
Lab1 Getting started with NP

Introduction to Human language Technology

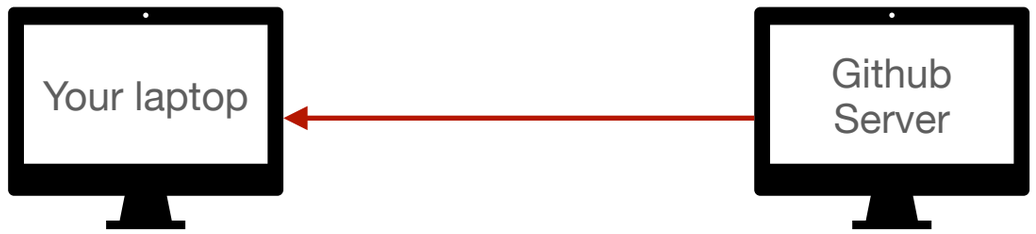
Piek Vossen

Getting the Lab notebooks

Check for updates



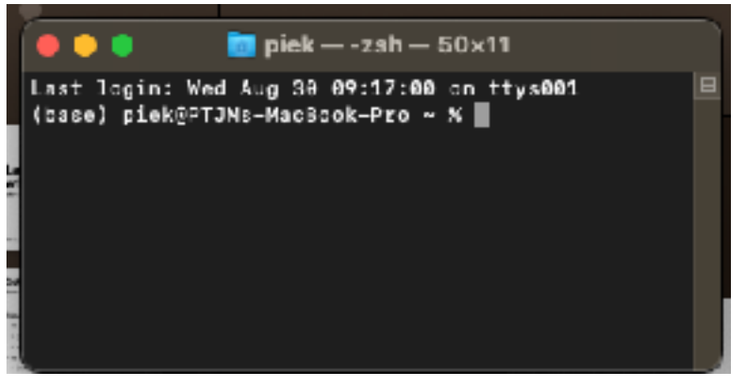
- https://github.com/cltl/ma-hlt-labs
  - Git installed:
    - Clone with ssh:
    - > git clone git@github.com:cltl/ma-hlt-labs.git
  - No local Git installed:
    - Download ZIP file
    - Unpack anywhere



Preparations

- Hardware
  - Bring your own laptop (Windows, Mac or Linux are all fine) with rights to install software and data.
  - At least 8GB of memory and 500GB of disk capacity.
  - Linux or MacBooks because these are most compatible with the environment of the staff and other researchers in the field. As a Linux system, Ubuntu is recommended, check if your laptop supports it. You can find lists of compatible laptops online (e.g. here: https://ubuntu.com/certified/laptops)
  - Windows works for most things but it is a little more difficult.
- Python
  - If you follow the Python for Text Analysis course parallel to this course, you will have the basic skills to attend.
  - Otherwise, install Anaconda on your local machine which also installs Python. We work with Python 3.10 or higher. You can follow the instructions to install anaconda given here:
    - https://www.anaconda.com/download
  - The installer will also install a graphical interface with various tools among which Jupyter notebooks and Jupyter lab. We will not use this but work from a Terminal or Command line.

The Terminal



- Terminal or Command line
  - The Terminal or Command line lets you type in basic commands that will be carried out by the computer.
  - Working with a terminal gives you more control over your code, is more efficient, and is the only way to run code on remote servers that do not come with a graphical interface.
  - Linux, Mac (Unix) users can launch a terminal:
    - In macOS: Open "Finder" ⇒ Go ⇒ Utilities ⇒ Select “Terminal".
    - In Ubuntu: Open "Dash" ⇒ type "Terminal"; or choose "Applications" ⇒ “Installed” ⇒ Select "Terminal".
  - Windows also has a terminal, which can be installed and activated as explained here:
    - https://docs.microsoft.com/en-us/windows/terminal/get-started
  - You can also simulate a linux terminal in Windows through Git bash:
    - https://www.atlassian.com/git/tutorials/git-bash
  - Useful links:
    - https://www.ibm.com/developerworks/aix/library/au-unixtext/
    - https://tldp.org/LDP/intro-linux/html/intro-linux.html
    - https://www3.ntu.edu.sg/home/ehchua/programming/howto/Unix_SurvivalGuide.html

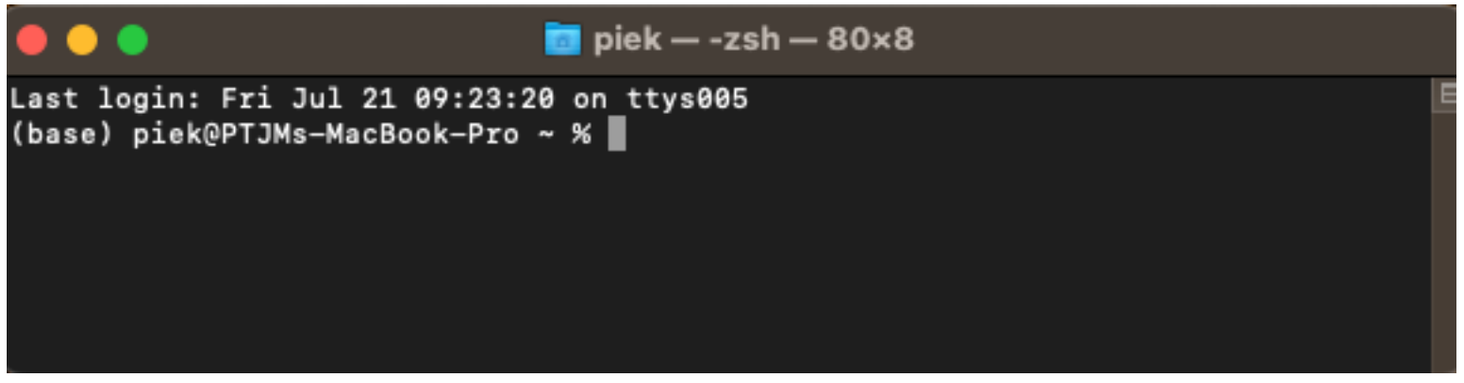
Using a terminal

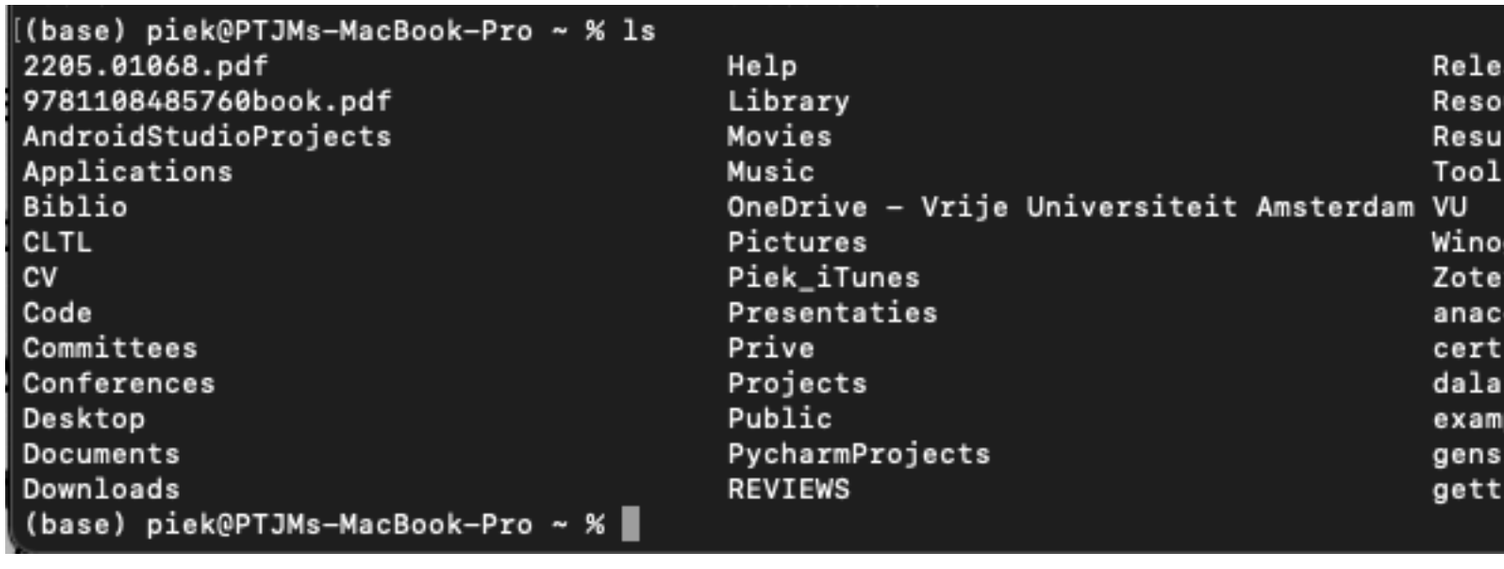
- A terminal gives you access to a so-called Unix/Linux *shell*.
    - Working with shell commands is extremely fast and efficient.
    - A taste of the power of shell commands "Unix for Poets" by Kenneth Church:
        - https://www.cs.upc.edu/~padro/Unixforpoets.pdf
        - How to count words, sort wordlists, make n-grams and make concordances for large amounts of text just using shell commands
- Basic commands for most of the classes (some will not work for Windows, for some there is a Windows alternative):
    - **pwd**: gives the path to the current directory
    - **ls** (Mac/Linux), **dir** (Windows): gives a list of what is stored in the current directory
    - **cd**: change directory, either going up to the parent or going in a subdirectory
    - **mkdir**: create a new directory in the current directory
    - **touch**: create a new empty file in the current directory
    - **echo** & **>**: return any text between quotes which can be redirected as a stream, e.g. echo “hello world” > file.txt
    - **cat:** (type Windows): print the content of a file to the screen
    - **mv**: move a file or rename a file
    - **rmdir**: remove a subfolder when empty
    - **rm**: permanently remove files and folders

Using a terminal

pwd, ls, dir

- When you open a terminal or command line box, you will be somewhere on your hard disk.
- You will see a prompt (% or >) after which you can type a command to the computer.
- Where are you on your disk?
  - In your terminal, type “pwd” directly after the prompt (%) and hit enter. My computer echos “/Users/piek” which is my home directory
- What is in my home dir?
  - Type “ls” (Mac/Linux) or “dir” (Windows) and you get a listing of files and subdirectories in the directory that you are now
  - My home directory has familiar subdirectories such as “Desktop”, “Documents” and “Downloads” and lots of other stuff





Using a terminal

cd

- The “cd” command stands for change directory and expects the name of a subdirectory
- Try any subdirectory that is listed, here we move into the “Downloads” subdirectory “cd Downloads”:
    - TIP: type “cd Downl” and press the TAB key. You will see that it is autocompleted to “cd Downloads”. This comes in handy when you have to type long names or pathes.
- Before the prompt “%”, we now see “Downloads” appear. Let’s use “pwd” again to check the path.
    - “/Users/piek/Downloads” so we went down into a subdirectory from my home dir.
- Using the command “ls” we can see what is in Downloads.
    - We see here folder with the name “old” and the folder “ma-hlt-labs-master” with the lab sessions that we downloaded from Github.
    - We can use “cd” again to enter it. Type “cd ma-hlt” and use the TAB key to complete.
- We use “ls” to get a listing. The option “-l” also make the terminal show details such as creation date/time, size and ownership. The list shows the different subfolders for the lab sessions of this course. Use “pwd” to see the full path.
- So far we went down but you also want to go up to a parent directory or grandparent. For this we use “cd ..”, where the double periods stand for one-level up.
    - Using “cd ..”, we go back up to “Downloads”, again to “piek”, next to “Users” and finally to “/“ which is the top root of the disk (different on Windows).
    - After going up, we go down again to “/Users/piek”, where we started.

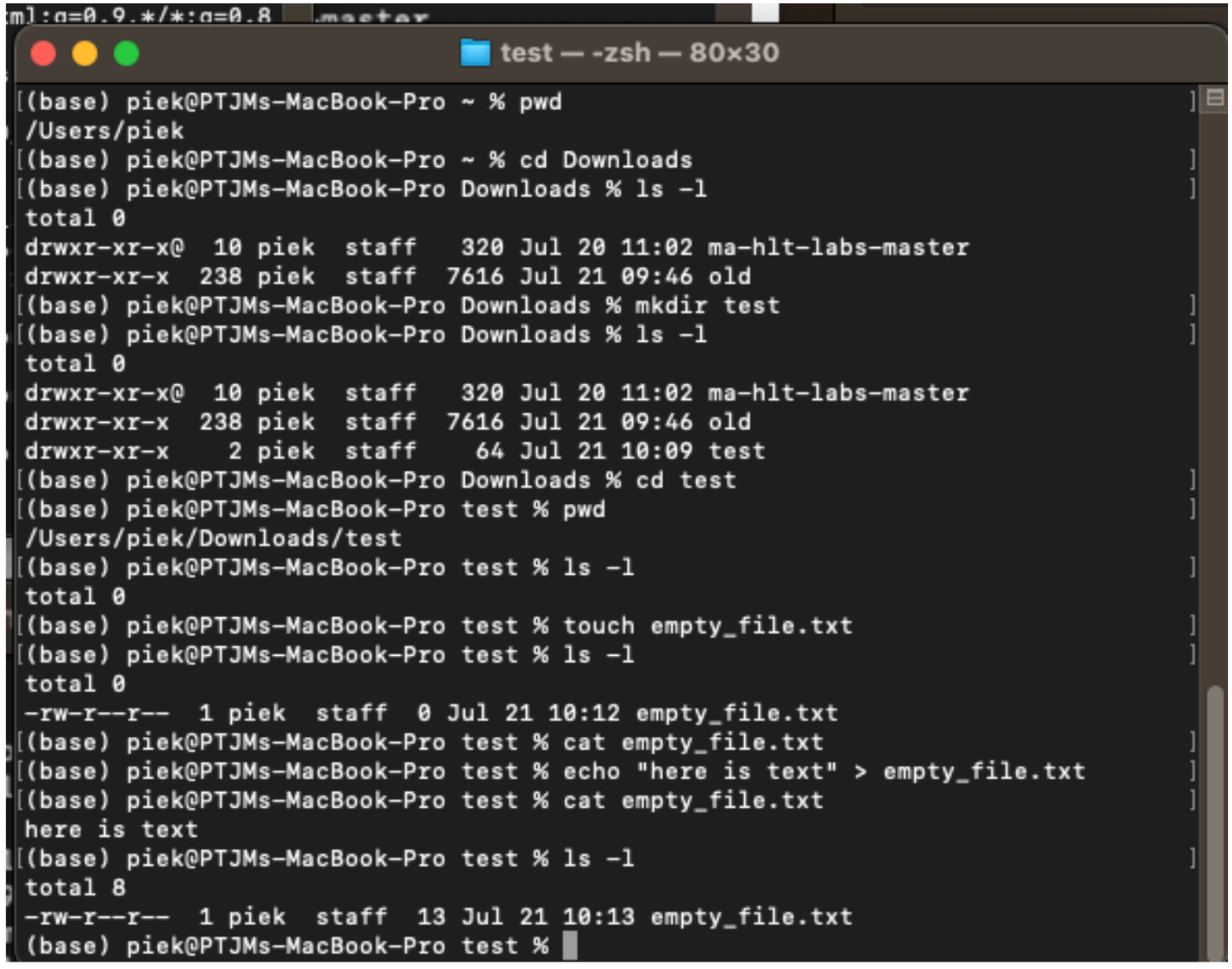
```
PycharmProjects
(base) piek@PTJMs-MacBook-Pro ~ % cd Downloads
(base) piek@PTJMs-MacBook-Pro Downloads % pwd
/Users/piek/Downloads
(base) piek@PTJMs-MacBook-Pro Downloads % ls
ma-hlt-labs-master      old
(base) piek@PTJMs-MacBook-Pro Downloads % cd ma-hlt-labs-master
(base) piek@PTJMs-MacBook-Pro ma-hlt-labs-master % ls -l
total 8
-rw-r--r--@  1 piek  staff  2520 Jul 20 11:02 README.md
drwxr-xr-x@  9 piek  staff   288 Jul 20 11:02 data-formats
drwxr-xr-x@ 10 piek  staff   320 Jul 20 11:02 lab1.toolkits
drwxr-xr-x@ 18 piek  staff   576 Jul 20 11:02 lab2.word_meaning
drwxr-xr-x@ 16 piek  staff   512 Jul 20 11:02 lab3.machine_learning
drwxr-xr-x@  9 piek  staff   288 Jul 20 11:02 lab4.contextualized-models
drwxr-xr-x@ 15 piek  staff   480 Jul 20 11:02 lab5.final_assignment
(base) piek@PTJMs-MacBook-Pro ma-hlt-labs-master % pwd
/Users/piek/Downloads/ma-hlt-labs-master
(base) piek@PTJMs-MacBook-Pro ma-hlt-labs-master %
```

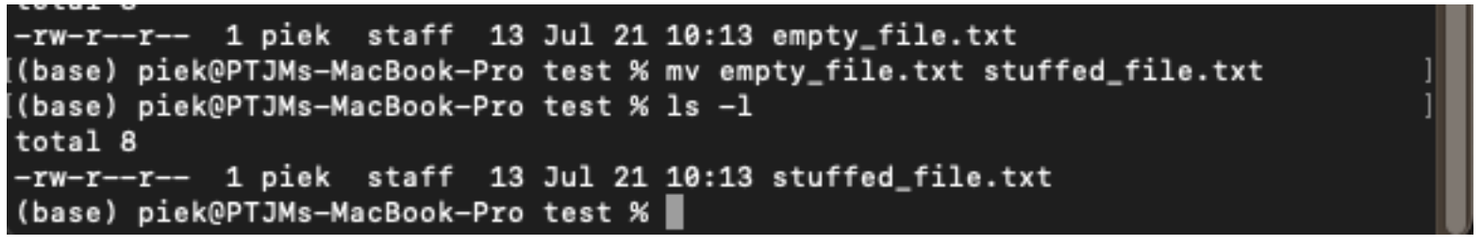
```
drwxr-xr-x@ 15 piek  staff   480 Jul 20 11:02 lab5.final_assignment
(base) piek@PTJMs-MacBook-Pro ma-hlt-labs-master % pwd
/Users/piek/Downloads/ma-hlt-labs-master
(base) piek@PTJMs-MacBook-Pro ma-hlt-labs-master % cd ..
(base) piek@PTJMs-MacBook-Pro Downloads % pwd
/Users/piek/Downloads
(base) piek@PTJMs-MacBook-Pro Downloads % cd ..
(base) piek@PTJMs-MacBook-Pro ~ % pwd
/Users/piek
(base) piek@PTJMs-MacBook-Pro ~ % cd ..
(base) piek@PTJMs-MacBook-Pro /Users % pwd
/Users
(base) piek@PTJMs-MacBook-Pro /Users % cd ..
(base) piek@PTJMs-MacBook-Pro / % pwd
/
(base) piek@PTJMs-MacBook-Pro / % cd Users
(base) piek@PTJMs-MacBook-Pro /Users % cd piek
(base) piek@PTJMs-MacBook-Pro ~ % pwd
/Users/piek
(base) piek@PTJMs-MacBook-Pro ~ %
```

Using a terminal

mkdir, touch, echo, cat. mv

- In addition to navigating, you can also create directories and files.
- **mkdir** makes directories:
    - In the terminal screen dump, we navigated back to Downloads and used “mkdir test” to make a new directory.
    - Using “cd” we can enter it and get a listing, which shows it is empty: “total 0”.
- **touch** makes files:
    - We can use “touch” (Mac/Linux) to make a new file inside the “test” folder.
    - When we get a listing for “test”, we now see the file but it is 0 kb in size (empty).
- **cat** (type Windows) shows the content
    - The “cat” command (type on Windows) prints the content which is empty.
    - We use the “echo” command to return a text “here is a text” which we next redirect using “>” as a stream into the empty file.
    - We use “cat empty_file.txt” again to print its content. It is not longer empty as is also shown when we get a listing of the “test” directory: 13 kb.
- **mv** moves or renames a file:
    - The file is no longer empty so let’s rename it.
    - For this, we use the “mv” command (Mac/Linux), which can be used for moving files as well as renaming if the target directory is the same but the filename is different. Try “mv empty_file.txt stuffed_file.txt” and get a new listing.





Using a terminal

rmdir, rm

- We created a “test” folder but now we want to clean up the mess.
- Removing by command line is efficient but also risky because there is no Trash to recover from.
    - “**rmdir**” removes a directory but only works when it is empty
    - “**rm**” removes files but with the option “**-r**” for recursively it also remove any subdirectory while emptying it first (everything from that point on is gone).
- We first try to remove the directory “test”. We navigate up to Downloads and type “rmdir test”:
    - We get the message: rmdir: test: Directory not empty, so this failed
- To remove it, we first need to empty it, so we navigate back in and use “rm stuffed_file.txt” to get rid of the file.
- Now we can go back up and use “rmdir test” from Downloads. The listing shows that it worked.
- We could have been more efficient by using “rm -r test” from Downloads, which would remove the content of “test” (both files and any subdirectories) and “test” itself.
    - However, this is very risky. Before doing this, check the content carefully using a listing, also check any subdirectories.
    - Doing this from the root (the top directory of your disk), will empty the full disk permanently

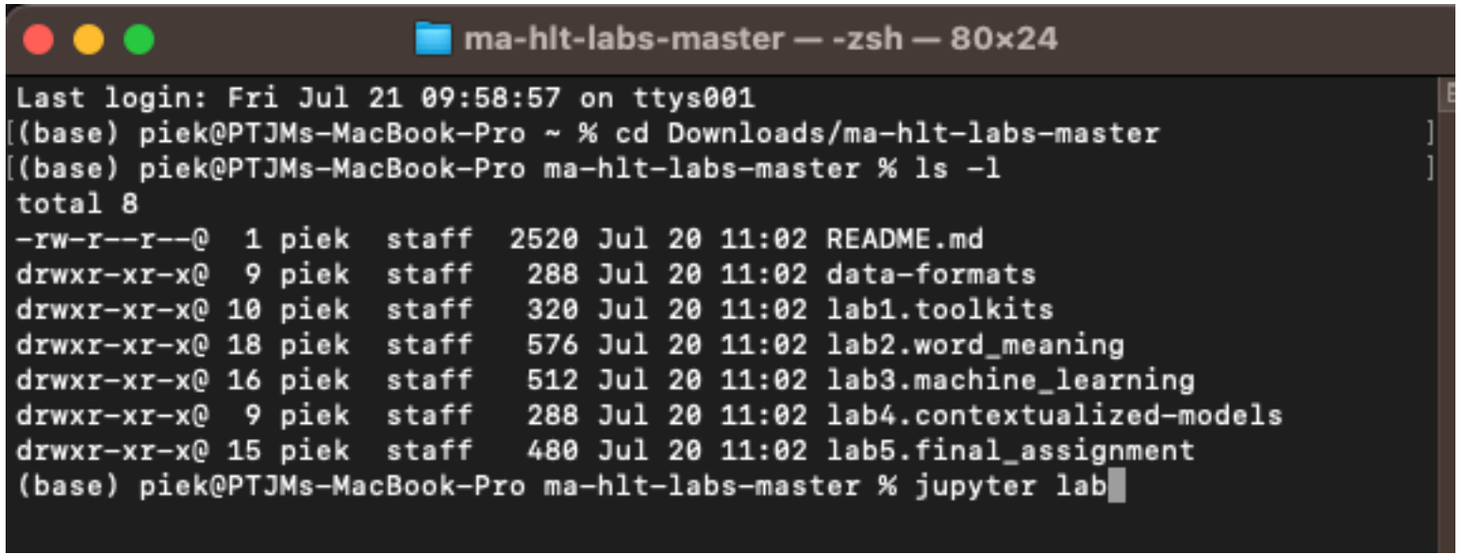
```
-rw-r--r--  1 piek  staff  13 Jul 21 10:13 stuffed_file.txt
(base) piek@PTJMs-MacBook-Pro test % cd ..
(base) piek@PTJMs-MacBook-Pro Downloads % rmdir test
rmdir: test: Directory not empty
(base) piek@PTJMs-MacBook-Pro Downloads % cd test
(base) piek@PTJMs-MacBook-Pro test % rm stuffed_file.txt
(base) piek@PTJMs-MacBook-Pro test % ls -l
total 0
(base) piek@PTJMs-MacBook-Pro test % cd ..
(base) piek@PTJMs-MacBook-Pro Downloads % rmdir test
(base) piek@PTJMs-MacBook-Pro Downloads % ls -l
total 0
drwxr-xr-x@  10 piek  staff   320 Jul 20 11:02 ma-hlt-labs-master
drwxr-xr-x  238 piek  staff  7616 Jul 21 09:46 old
(base) piek@PTJMs-MacBook-Pro Downloads %
```

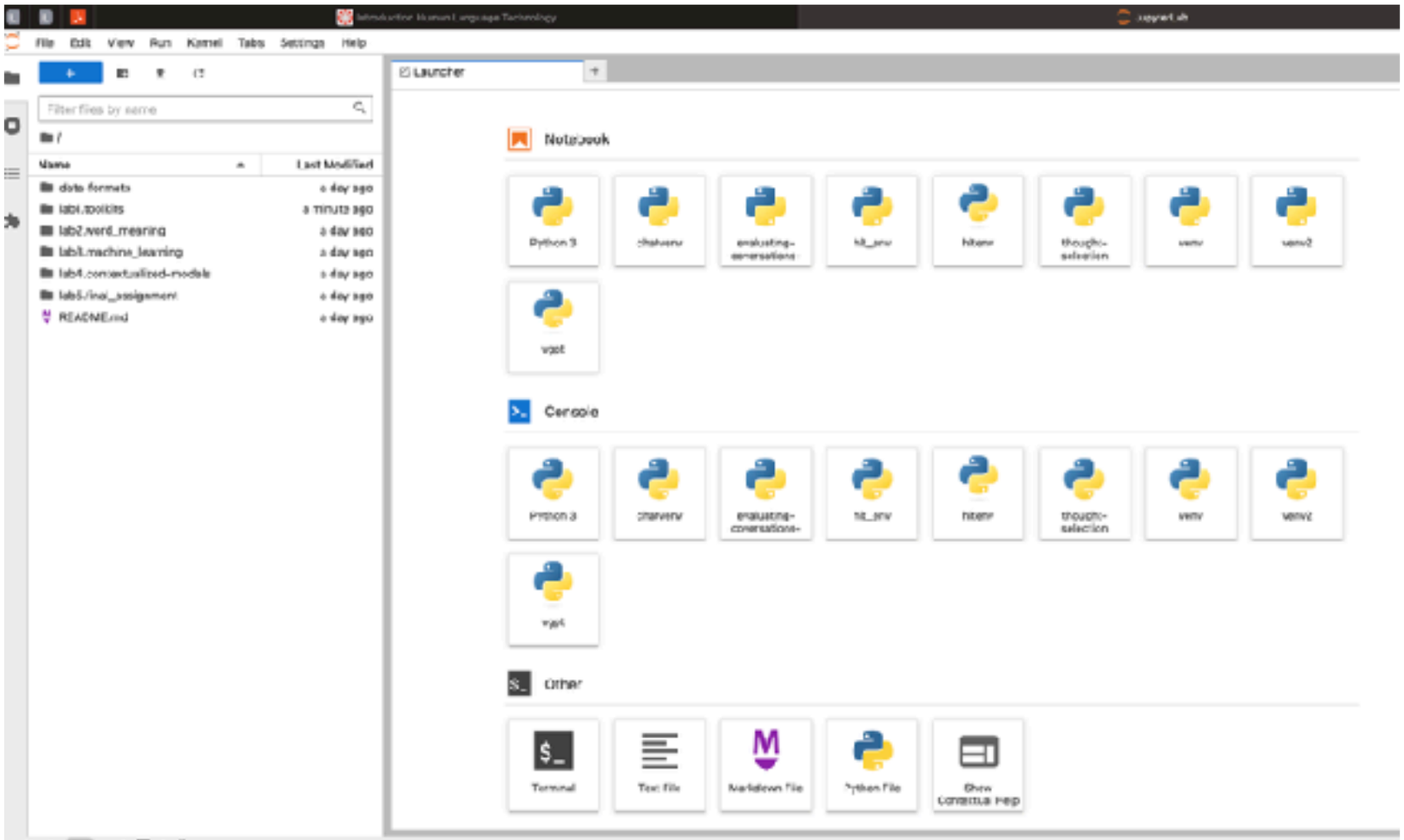
Preparations: Jupyter notebooks

- If you have successfully installed Anaconda, you should be able to launch Jupyter lab from the command line:

1. open the command line or a terminal
2. navigate to the directory where the notebooks are (see the navigation instructions)
3. Type "jupyter lab" after the prompt and press enter as shown here.
4. Your default browser should open the page as shown below with the lab session folders to the left.
5. Navigate to the folder where you stored the notebooks for Lab 1 and open the notebook with the introduction.

- Documentation for JupyterLab can be found here: https://jupyterlab.readthedocs.io/en/stable/





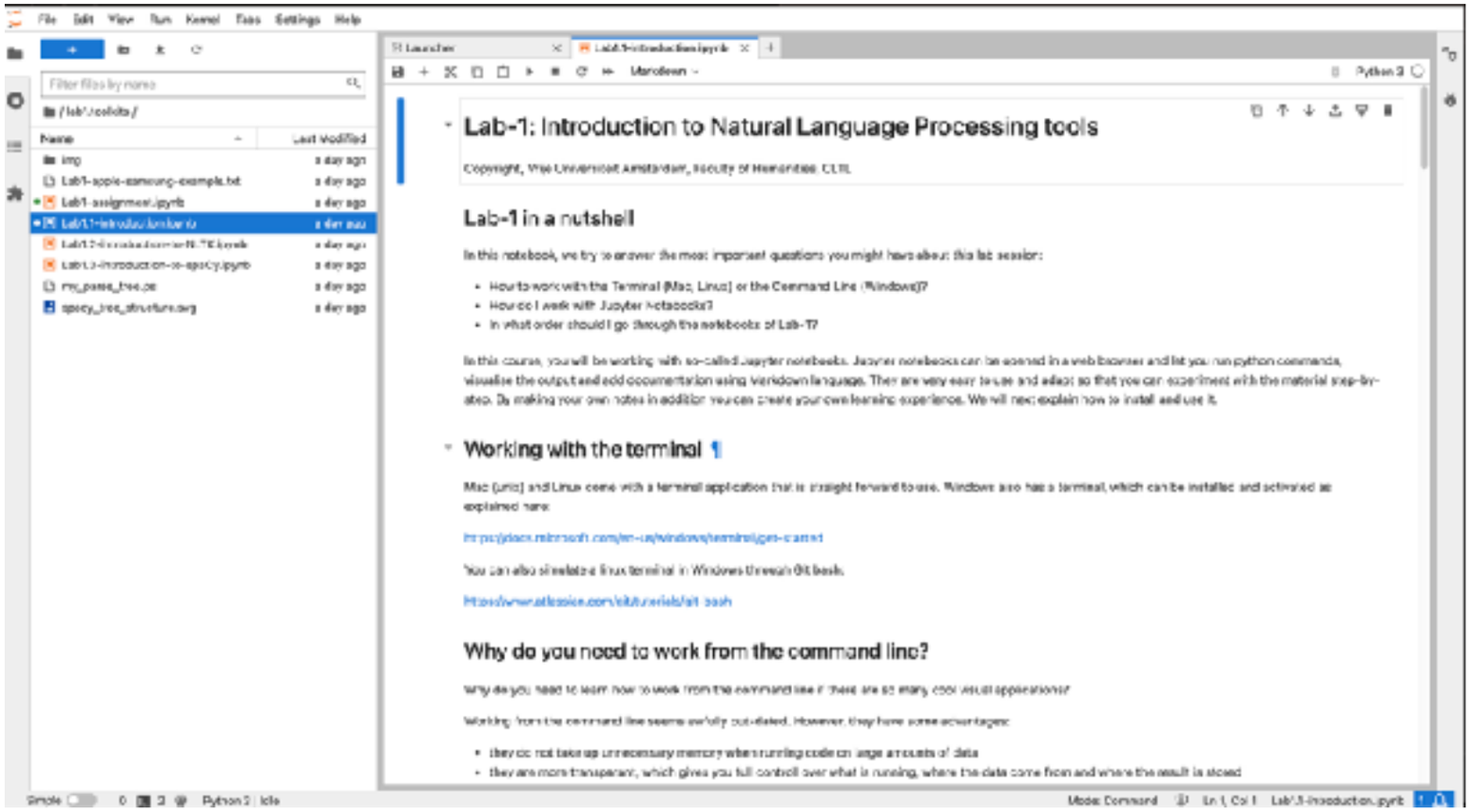
Preparations: Jupyter notebooks

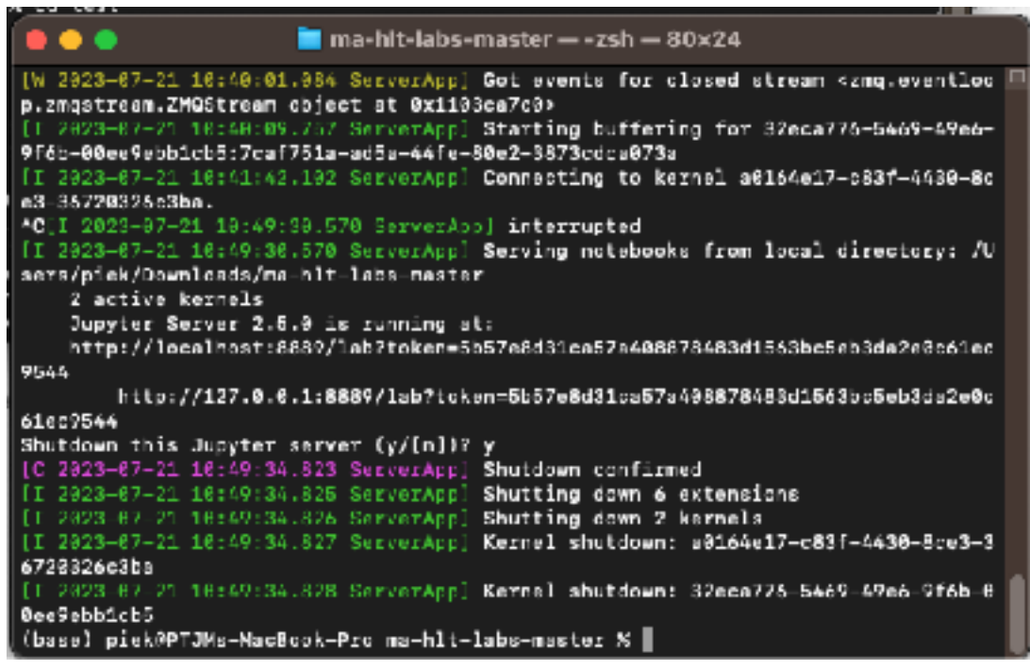
- Notebooks contain instructions and so-called 'code blocks'. The instructions are paragraphs of text that explain the concepts we are going to use. The 'code blocks' contain Python code.

- Some tips:

  - Cells in a notebook contain code or text (Markdown). If you run a cell, it will either run the code or render the text from the Markdown.
  - There are five ways to run a cell:
    - Click the 'play' button next to the 'stop' and 'refresh' button in the toolbar.
    - Alt + Enter runs the current cell and creates a new cell.
    - Ctrl + Enter runs the current cell without creating a new cell.
    - Shift + Enter runs the current cell and moves to the next one.
    - Use the menu and select *Kernel -> Restart Kernal and Run All Cells*
  - The instructions are written in Markdown:
  - https://github.com/adam-p/markdown-here/wiki/Markdown-Cheatsheet
  - https://www.dataquest.io/blog/jupyter-notebook-tutorial/

- To stop Jupyter you can close the tab in the browser but you also need to stop Jupyter in the terminal:
  - Go to the terminal and press CTRL-c This will interrupt the program and ask you to confirm: Shutdown this Jupyter server (y/[n])?
  - If confirmed it terminates
  - You can also press CTRL-c twice to kill it directly.





Preparations

Text editors

- We will work with text that is stored in files.
- Typically, we do NOT use Word, HTML or PDF as these contain a lot more than just the text or they are in binary format.
- Our code needs the pure text as input to work.
- For inspecting the text, use a “plain” text editor (not Word):
    - Windows Notepad++: https://notepad-plus-plus.org/
    - Mac/Linux:
        - Atom: https://github.com/atom/atom/releases/tag/v1.60.0
    - Mac:
        - Bbedit: https://www.barebones.com/products/bbedit/