TD3/9: Work with text manipulation tools in Linux

Exercise 1: Grep and awk on tabular data

1. Display the list of files and folders at the root using ls -l
2. In a pipeline (using |), append a **grep** instruction to only display informations of *bin*
3. Append an **awk** instruction to only display the size of *bin*
4. Now rather extract the month, day and year of creation of the folder *bin*
5. Now rearrange the instruction to get the following output format : 2020-Oct-26 (from original data : Oct 26 2020)

Exercise 2: Grep with Regex, and sed on unstructured data

1. Run the following command :
   curl https ://en.wikipedia.org/wiki/List_of_cyberattacks > cyberattacks.txt
2. Use grep to extract all the lines that contain the keyword *"meta"*
3. Now only extract *"meta"* and the first following word. You might use grep options to enable the use of **regex (Regular Expressions)** [1]
4. Only extract the follwing word (but not the keyword *"meta"*)
5. Let's now try more interesting (yet complex) patterns. You might use vim to open the file and look for useful patterns. Let's extract the introduction
   — We could ask grep to catch the paragraph corresponding to a sentence that is only present in the introduction. Try to run the following command *cat cyberattacks.txt | grep -P 'A cyberattack is'*
   — This does not work since the source code is here different from what is visible on the web page. Now try the following : *cat cyberattacks.txt | grep -P 'A cyberattack is any type'*
   — It is now working, but what if the text evolves over time ? Try the following instead : *cat cyberattacks.txt | grep -A1 'mw-content-text' | grep -v 'mw-content-text'*. This is based on the text above that seems to be more stable.
6. Your turn
   — Extract the tab title
   — Make a list of cyber attacks based on section titles

1. https ://regexr.com/2tr5t