

Modèles de Linguistique Computationnelle

CM 2 : Bash, un langage utilitaire

M1 Langue et Informatique

Crédits: Gaël Lejeune, Karén Fort, Iana Atanassova, Djame Seddah, Eleni Kogkitsidou, Olga Seminck

Alice Millour prenom.nom@sorbonne-universite.fr

Yoann Dupont prenom.nom@sorbonne-universite.fr

2020-2021

Sorbonne-Université

Plan

Bash

Expressions régulières

Pourquoi utiliser bash ?

Bash = un langage intégré à linux qui permet (entre autres) :

- de naviguer dans le système de fichiers
- de visualiser / modifier les fichiers

⇒ utile pour manipuler les fichiers / ressources linguistiques utilisées en modélisation linguistique

Commande bash

```
$ commande -option1 -option2 -option3 argument1 argument2
```

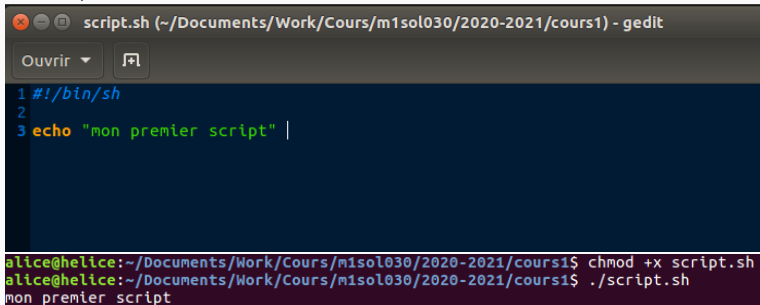
ligne de commande vs. script

deux modes d'utilisation :

- ligne de commande (directement dans le **terminal**) :

```
alice@helice:~/Documents/Work/Cours/misol030/2020-2021/cours1$ echo "Hello world !"
Hello world !
```

- script (dans un fichier qui est ensuite **exécuté dans le terminal**) :



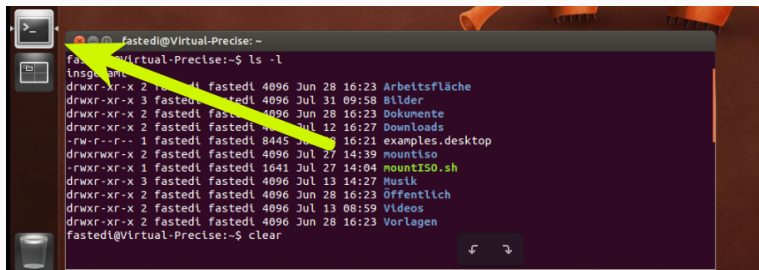
The screenshot shows a terminal window titled "script.sh (~/Documents/Work/Cours/misol030/2020-2021/cours1) - gedit". The window contains a text editor with the following content:

```
1 #!/bin/sh
2
3 echo "mon premier script" |
```

Below the editor, the terminal shows the following commands and output:

```
alice@helice:~/Documents/Work/Cours/misol030/2020-2021/cours1$ chmod +x script.sh
alice@helice:~/Documents/Work/Cours/misol030/2020-2021/cours1$ ./script.sh
mon premier script
```

Le terminal



```
fastedi@Virtual-Precise: ~  
fastedi@Virtual-Precise:~$ ls -l  
insgesamt 40  
drwxr-xr-x 2 fastedi fastedi 4096 Jun 28 16:23 Arbeitsfläche  
drwxr-xr-x 3 fastedi fastedi 4096 Jul 31 09:58 Bilder  
drwxr-xr-x 2 fastedi fastedi 4096 Jun 28 16:23 Dokumente  
drwxr-xr-x 2 fastedi fastedi 4096 Jul 12 16:27 Downloads  
-rw-r--r-- 1 fastedi fastedi 8445 Jun 28 16:21 examples.desktop  
drwxrwxr-x 2 fastedi fastedi 4096 Jul 27 14:39 mountiso  
-rw-r-xr-x 1 fastedi fastedi 1641 Jul 27 14:04 mountISO.sh  
drwxr-xr-x 3 fastedi fastedi 4096 Jul 13 14:27 Musik  
drwxr-xr-x 2 fastedi fastedi 4096 Jun 28 16:23 Öffentlich  
drwxr-xr-x 2 fastedi fastedi 4096 Jul 13 08:59 Videos  
drwxr-xr-x 2 fastedi fastedi 4096 Jun 28 16:23 Vorlagen  
fastedi@Virtual-Precise:~$ clear
```

Le terminal *interprète* les commandes Bash

Le terminal - la base

se déplacer dans l'arborescence des fichiers avec la commande `cd`

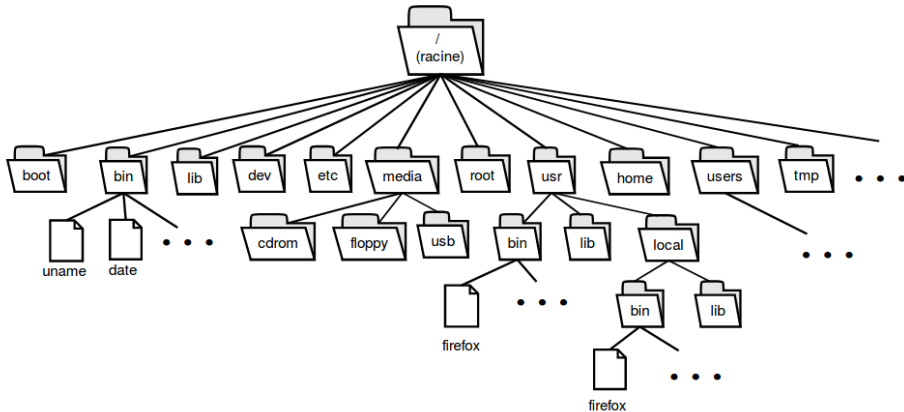


Figure 1 – arborescence des fichiers dans linux

Le terminal - la base

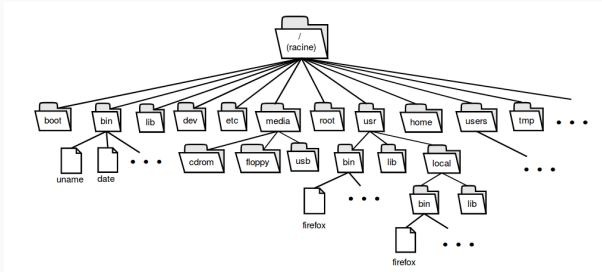


Figure 2 – arborescence des fichiers dans linux

1. en utilisant les chemins absolus (par exemple `/usr/local/bin`)
2. en utilisant les chemins relatifs (rappel : `..` désigne le dossier parent du dossier courant)

ex : comment se rendre dans le dossier `lib` depuis le dossier `usb` ?

Le terminal - la base

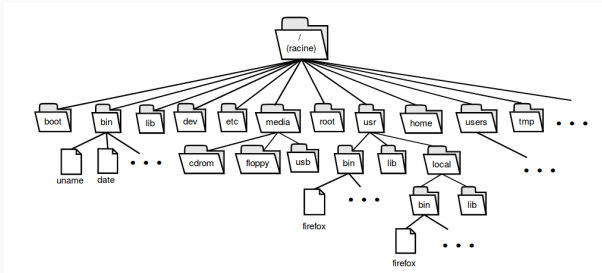


Figure 3 – arborescence des fichiers dans linux

1. \$ cd : déplace vers le dossier "home" de l'utilisateur courant
2. \$ cd - : déplace vers la position précédente

```
alice@helice:~$ cd Documents/Work
alice@helice:~/Documents/Work$ cd
alice@helice:~$ cd -
/home/alice/Documents/Work
alice@helice:~/Documents/Work$
```

Exemple concret d'utilisation avancée (1)

```
alice@helice:~/Documents/Work/Cours/misolo30/2020-2021/cours2$ tail corpus_francais.conll -n100
2 Yves Yves N NPP g=m|n=s|s=p 0 root - -
3 Gras Gras N NPP g=m|n=s|s=p 2 mod - -
4 " " PUNCT PUNCT s=w 2 punct - - -
5 " " PUNCT PUNCT s=w 6 punct - - -
6 Histoire histoire N NC g=f|n=s|s=c 2 mod - -
7 de de P P - 6 dep - - -
8 la le D DET g=f|n=s|s=def 9 det - - -
9 Guerre guerre N NC g=f|n=s|s=c 7 obj.p - - -
10 d' de P P - 9 dep - - -
11 Indochine Indochine N NPP s=p 10 obj.p - - -
12 " " PUNCT PUNCT s=w 6 punct - - -
13 " " PUNCT PUNCT s=w 2 punct - - -
14 Plon Plon N NPP s=p 2 mod - - -
15 " " PUNCT PUNCT s=w 2 punct - - -
16 Paris Paris N NPP g=m|n=s|s=p 2 mod - -
17 " " PUNCT PUNCT s=w 2 punct - - -
18 1979 1979 N NC s=card 2 mod - -
19 " " PUNCT PUNCT s=s 2 punct - - -
20 " " PUNCT PUNCT s=s 2 punct - - -
21 " " PUNCT PUNCT s=w|sentid=frwiki.50.1000.00982 2 punct - -
22 Jules Jules N NPP g=f|n=s|s=p 0 root - -
23 Roy Roy N NPP g=m|n=s|s=p 2 mod - -
24 " " PUNCT PUNCT s=w 2 punct - - -
25 La le D DET g=f|n=s|s=def 6 det - - -
26 bataille bataille N NC g=f|n=s|s=c 2 mod - -
27 de de P P - 6 dep - - -
28 Dien Dien N NPP s=p 7 obj.p - - -
29 Bien Bien N NPP s=p 8 mod - - -
30 Phu Phu N NPP s=p 8 mod - - -
31 " " PUNCT PUNCT s=w 2 punct - - -
32 Julliard Julliard N NPP s=p 2 mod - -
33 " " PUNCT PUNCT s=w 2 punct - - -
34 1963 1963 N NC s=card 2 mod - -
35 " " PUNCT PUNCT s=w 2 punct - - -
36 Albin Albin N NPP s=p 2 mod - - -
37 Michel Michel N NPP g=m|n=s|s=p 16 mod - -
38 " " PUNCT PUNCT s=w 2 punct - - -
39 1989 1989 N NC s=card 2 mod - -
40 " " PUNCT PUNCT s=s 2 punct - - -
41 " " PUNCT PUNCT s=w|sentid=frwiki.50.1000.00984 2 punct - -
```

extraction des adverbes d'un corpus au format Conll :

```
$ cat corpus_francais.conll | grep ADV | cut -d' ' -f3 | sort | uniq |
tr '_' ' ' > liste_adverbes.txt
```

Exemple concret d'utilisation avancée (2)

```
alice@helice:~/Documents/Work/Cours/misolo30/2020-2021/cours2$ tail corpus_francais.conll -n100
2 Yves Yves N NPP g=m|n=s|s=p 0 root - -
3 Gras Gras N NPP g=m|n=s|s=p 2 mod - -
4 " " PUNCT PUNCT s=w 2 punct - - -
5 " " PUNCT PUNCT s=w 6 punct - - -
6 Histoire histoire N NC g=f|n=s|s=c 2 mod - -
7 de de P P - 6 dep - - -
8 la le D DET g=f|n=s|s=def 9 det - - -
9 Guerre guerre N NC g=f|n=s|s=c 7 obj.p - - -
10 d' de P P - 9 dep - - -
11 Indochine Indochine N NPP s=p 10 obj.p - - -
12 " " PUNCT PUNCT s=w 6 punct - - -
13 " " PUNCT PUNCT s=w 2 punct - - -
14 Plon Plon N NPP s=p 2 mod - - -
15 " " PUNCT PUNCT s=w 2 punct - - -
16 Paris Paris N NPP g=m|n=s|s=p 2 mod - -
17 " " PUNCT PUNCT s=w 2 punct - - -
18 1979 1979 N NC s=card 2 mod - -
19 " " PUNCT PUNCT s=s 2 punct - - -
20 " " PUNCT PUNCT s=w|sentid=frwiki.50.1000.00982 2 punct - -
21 Jules Jules N NPP g=f|n=s|s=p 0 root - -
22 Roy Roy N NPP g=m|n=s|s=p 2 mod - -
23 " " PUNCT PUNCT s=w 2 punct - - -
24 La le D DET g=f|n=s|s=def 6 det - - -
25 bataille bataille N NC g=f|n=s|s=c 2 mod - -
26 de de P P - 6 dep - - -
27 Dien Dien N NPP s=p 7 obj.p - - -
28 Bien Bien N NPP s=p 8 mod - - -
29 Phu Phu N NPP s=p 8 mod - - -
30 " " PUNCT PUNCT s=w 2 punct - - -
31 Julliard Julliard N NPP s=p 2 mod - -
32 " " PUNCT PUNCT s=w 2 punct - - -
33 1963 1963 N NC s=card 2 mod - -
34 " " PUNCT PUNCT s=w 2 punct - - -
35 Albin Albin N NPP s=p 2 mod - - -
36 Michel Michel N NPP g=m|n=s|s=p 16 mod - -
37 " " PUNCT PUNCT s=w 2 punct - - -
38 1989 1989 N NC s=card 2 mod - -
39 " " PUNCT PUNCT s=s 2 punct - - -
40 " " PUNCT PUNCT s=w|sentid=frwiki.50.1000.00984 2 punct - -
```

Décompte de chaque catégorie grammaticale à partir du corpus :

```
$ cat corpus_francais.conll | cut -d' ' -f4 | sort | uniq -c
```

Plan

Bash

Expressions régulières

Expressions régulières

expression régulière (*regex*) = *chaîne de caractères, qui décrit, selon une syntaxe précise, un **ensemble de chaînes de caractères possibles***

Les regex permettent notamment de sélectionner des **motifs** :

- les adresses e-mail : `^[a-zA-Z-]+@[a-zA-Z-]+[a-zA-Z]{2,6}$`
- les mots commençant par une majuscule : `[A-Z][a-z]+`
- les mots d'au moins 5 lettres : `[a-zA-Z]6`
- etc.

https://fr.wikipedia.org/wiki/Expression_r%C3%A9gul%C3%A8re

Expressions régulières

Les expressions régulières répondent à une certaine **syntaxe** qu'il faut apprendre...

Tutoriel de prise en main : <https://www.lucaswillems.com/fr/articles/25/tutoriel-pour-maitriser-les-expressions-regulieres>

Manipulation de **bash** avec les **expressions régulières** :

- manipuler les fichiers *dont le nom correspond à un motif*, par exemple tous les fichiers '.txt'
- extraire tous les mots *contenant un motif*, par exemple démanger, mangerai, démangerait etc.
- extraire toutes les phrases *contenant plus de n mots*
- etc.