Laboratorio di IU
Manipolazione di Dati

https://tinyurl.com/y2jpjupw

9 marzo 2021

Rachele Sprugnoli – rachele.sprugnoli@unicatt.it

UNIVERSITÀ CATTOLICA del Sacro Cuore

PRIMI PASSI

Come aprire la linea di comando?

- Mac & Linux: terminale (shell, bash)
  - Mac: Applications → Utilities → Terminal
- Windows: Cygwin terminal 

Dove mi trovo?

- Digitate `pwd`

Cosa c'è qui?

- Digitate `ls`

PRIMI PASSI

Manuale completo:

- `man` comando (si esce con q)

  `man pwd`

Aiuto sintetico:

- comando `--help`

  `pwd --help`

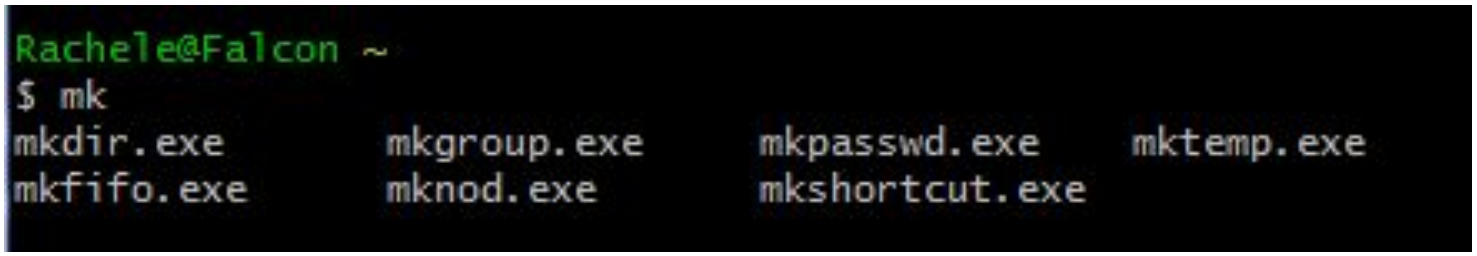
```
C:\Users\Rachele>pwd --help
Usage: pwd [OPTION]...
Print the full filename of the current working directory.

  -L, --logical   use PWD from environment, even if it contains symlinks
  -P, --physical  avoid all symlinks
      --help     display this help and exit
      --version  output version information and exit

If no option is specified, -P is assumed.

NOTE: your shell may have its own version of pwd, which usually supersedes
the version described here.  Please refer to your shell's documentation
for details about the options it supports.

GNU coreutils online help: <http://www.gnu.org/software/coreutils/>
Report pwd translation bugs to <http://translationproject.org/team/>
Full documentation at: <http://www.gnu.org/software/coreutils/pwd>
or available locally via: info '(coreutils) pwd invocation'
```

PRIMI PASSI

Spostarsi

- `cd` (Change Directory) `nome_cartella`
- `cd` spazio e poi trascinare l'icona della cartella scaricata
- cd .. per tornare indietro di una cartella

Usare il tasto **tab** (⇆) per il completamento automatico e ridurre gli errori: clicco 2 volte per vedere tutti i comandi



TESTI

Concatenare file di testo

- `cat` *(conCATenare)* `verso1.txt verso2.txt`

Come salvare il risultato in un nuovo file: `>`

- `cat verso1.txt verso2.txt > nomefile.txt`
  - non vale solo per `cat`!

Come aggiungere un testo a un altro: `>>`

- `cat verso2.txt >> verso1.txt`
  - aprite i file: cosa è successo?
  - attenti alle sovrascritture!

STATISTICHE DI BASE

Usiamo il file *canto1.txt*:

- `wc` (Word Count)

  `wc -l canto1.txt` (conta le righe)
  `wc -w canto1.txt` (conta le sequenze separate da spazio)
  `wc -m canto1.txt` (conta i caratteri)
  `wc -c canto1.txt` (conta i byte)

```
Rachele@Falcon /cygdrive/c/Users/Rachele/Desktop/Presentazioni_2019/Pavia
$ wc -l canto1.txt
180 canto1.txt

Rachele@Falcon /cygdrive/c/Users/Rachele/Desktop/Presentazioni_2019/Pavia
$ wc -w canto1.txt
952 canto1.txt

Rachele@Falcon /cygdrive/c/Users/Rachele/Desktop/Presentazioni_2019/Pavia
$ wc -m canto1.txt
5209 canto1.txt

Rachele@Falcon /cygdrive/c/Users/Rachele/Desktop/Presentazioni_2019/Pavia
$ wc -c canto1.txt
5289 canto1.txt
```

STATISTICHE DI BASE

Usiamo il file *canto1.txt*:

- `wc` (Word Count)

  `wc -l canto1.txt` (conta le righe)
  `wc -w canto1.txt` (conta le parole - sequenze separate da spazio)
  `wc -m canto1.txt` (conta i caratteri)
  `wc -c canto1.txt` (conta i byte)

Cosa succede con questo comando?

- `wc -l *.txt`

RICERCHE

Il comando fondamentale

- `grep` (General Regular Expression Print)

  SINTASSI: `grep (-opzione) “ricerca” file`

  ```
  grep 'Dio' canto1.txt canto2.txt
  ```

  - restituisce le righe, può esserci più di una corrispondenza per riga → attenzione alle maiuscole!
  - altre opzioni utili: -i (case insensitive) --color (colorazione)

  ```
  grep -o 'Dio' canto1.txt canto2.txt
  ```

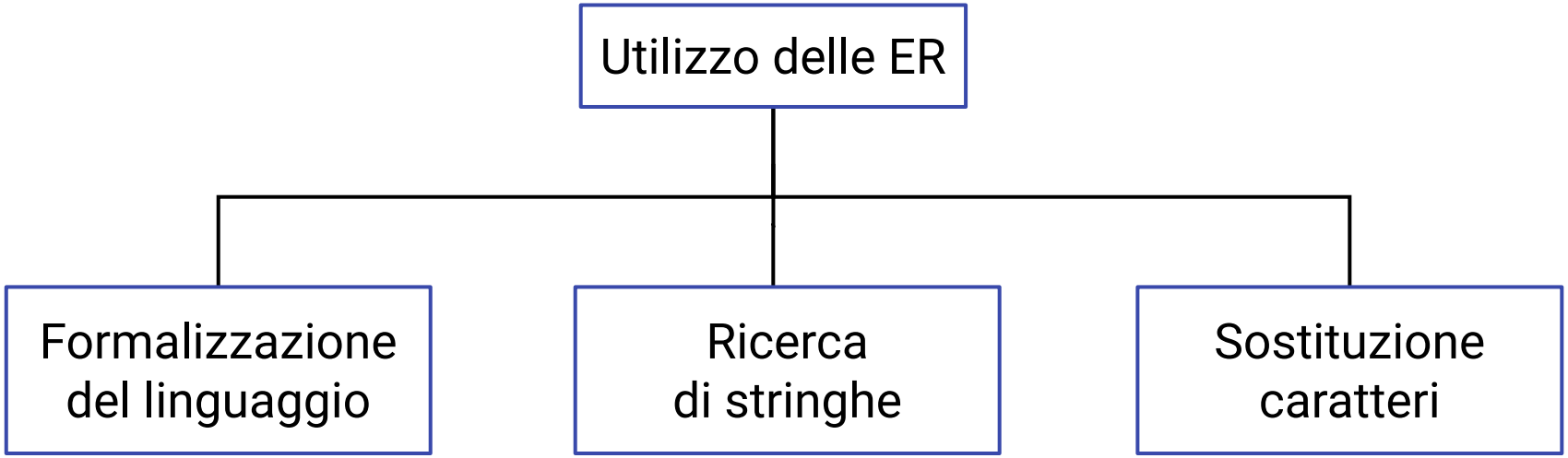
  - per avere solo le corrispondenze: -o

  ```
  grep 'Dio' canto1.txt canto2.txt | wc -l
  ```

  - la pipe | concatena comandi

ESPRESSIONI REGOLARI

- Linguaggio formale per definire stringhe di testo



- Link utili:
  - Tutorial interattivo: **https://regexone.com/**
  - Piattaforma di test online: **https://regexr.com/**
  - Per Sublime Text: **https://tinyurl.com/r5dfj7uh**

ESPRESSIONI REGOLARI

Pattern	Significato	Matching
[A-Z]	una lettera maiuscola	**C**hatbot
[a-z]	una lettera minuscola	**c**hatbot
[0-9]	una cifra numerica	Capitolo **1**

Pattern	Significato	Matching
gatto|cane	o “gatto” o cane”	il **cane** è nella cuccia
[cC]ane|[gG]atto	iniziale maiuscola o minuscola	**Cane**, **cane**, **gatto**, **Gatto**

ESPRESSIONI REGOLARI

Pattern	Significato	Matching
p**.**zza	qualsiasi carattere	**pazza pozza pizza**
pi**?**azza	0 o 1 occorrenza	**piazza pazza**
no*****	0 o più occorrenze	**n no nooooooo**
no**+**	1 o più occorrenze	**no nooooooo**
[0-9]**{**0,3**}**	numero di occorrenze	**0 39 111**
piazza**$**	fine della stringa	Tutti in **piazza**
^Piazza	inizio della stringa	**Piazza** pulita
piazza\.	carattere di escape	Tutti in **piazza.**

ESPRESSIONI REGOLARI

- `grep` con opzione `-E`

  `grep -E ‘^Ahi’ canto1.txt`
  - Tutte le occorrenze di “Ahi” ad inizio riga

  `grep -E -o '^[A-Z][a-z]*' canto1.txt`
  - le prime parole di una riga che iniziano per maiuscola

  `grep -E -o '^[A-Z][a-z]+' canto1.txt`
  - le prime parole di una riga che iniziano per maiuscola e che sono formate da almeno 2 lettere

  `grep -E " pel|per " canto1.txt`
  - le occorrenze di “per” o “pel”

  `grep -E -o "[[:punct:]]" canto1.txt`
  - tutta la punteggiatura

FILE TSV/CSV

Selezionare/filtrare una o più colonne di dati
Apriamo il file `BellumGallicum.txt (con Sublime Text) e BellumGallicum.csv (con LibreOffice Calc)`

- Comando `cut`

  `cut -f 3 BellumGallicum.txt`
  - opzione -f per specificare il numero della colonna da selezionare, la numerazione parte da 1

  `cut -f 3,1 BellumGallicum.txt`
  - seleziono sia la terza che la prima colonna

  `cut -f 2 -d ‘,’ BellumGallicum.csv`
  - se le colonne non sono divise da \tab, specifico il delimitatore di campo con l’opzione `-d`

USIAMO PIÙ COMANDI

Provate ad eseguire i seguenti comandi: qual è il risultato?
Cosa fanno sort, uniq, uniq -c?

```
cut -f 3 BellumGallicum.txt | sort

cut -f 3 BellumGallicum.txt | sort | uniq

cut -f 3 BellumGallicum.txt | sort | uniq -c

cut -f 3 BellumGallicum.txt | grep "NOUN" | wc -l

cut -f 1,3,5 BellumGallicum.txt | grep "NOUN" | grep
"Plur"
```

ESPRESSIONI REGOLARI SU SUBLIME TEXT

- Apriamo *Letters.txt* con Sublime Text: si tratta dell'indice del libro "The Letters of Anne Gilchrist and Walt Whitman"

  → trasformiamo l'indice in un foglio di calcolo con 4 colonne: mittente, destinatario, luogo del mittente, data

  → da dato non strutturato (testo) a dato strutturato (tabella)

  → possibile input di strumenti di network analysis

- Ctrl+h (Windows) o alt+cmd+F (Mac) per aprire la finestra di Replace
- Selezionare le opzioni 1,2 e 4 sulla sinistra

ESPRESSIONI REGOLARI SU SUBLIME TEXT

Eseguire i seguenti passi uno alla volta

1. eliminare gli spazi all'inizio di ogni riga: ^( )* --> nulla
2. togliere il numero di pagina: ( )*[0-9]+\n --> nulla
3. mettere informazioni sulla stessa riga: \n_ --> _
4. togliere i numeri romani iniziali: ^[A-Z]+\. --> nulla
5. aggiungere tabulazioni come separatori di colonne: _ --> \t
6. togliere virgola prima degli anni: , ([0-9]{4}) --> \1
7. togliere doppi tab: \t\t --> \t
8. dividere mittente e destinatario: TO --> \t

Alla fine copiare e incollare su un foglio di calcolo e aggiungere nomi alle colonne

MODIFICARE TESTI

- Comando `tr` (TRanslation)

- `tr carattere1 carattere2 < testo1.txt > testo2.txt`
  - si possono usare le classi `[:punct:], [:lower:], ...`
  - opzione `-d` (delete): cancella i caratteri scelti
  - attenzione: < input; > output

```
tr -d “[[:punct:]]” < canto1.txt > canto1_nopunct.txt

tr “[[:upper:]]” “[[:lower:]]” < canto1_nopunct.txt > canto1_minuscole.txt

tr " " "\n" < canto1_minuscole.txt > canto1_parole.txt

sort canto1-parole.txt | uniq -c | sort
```

GRAZIE!

Email: rachele.sprugnoli@unicatt.it
Twitter: @RSprugnoli