Programmiertechniken in der Computerlinguistik I

Programming for Linguists

University of Zurich, September 2023
Martin Volk, Simon Clematide

Exercise 0: Getting acquainted with Unix commands

Try the following Unix commands

Command	**Meaning**	**Erklärung / Explanation**
ls -l	listing	DE: Auflisten der Dateien in einem Verzeichnis EN: List the files in a folder (also called: directory)
pwd	print working directory	DE: Zeige den aktuellen Pfad und das aktuelle Verzeichnis EN: Show the current path and the current directory
cd	change directory	DE: Wechsle in ein anderes Verzeichnis EN: Change to a different directory
mkdir	make directory	DE: Erzeuge ein Unterverzeichnis im aktuellen Verzeichnis EN: Create a new directory in the current directory
wc	word count	DE: Zeige die Grösse einer Datei: Anzahl Zeilen, Wörter und Zeichen EN: Show the size of a file: number of lines, words, characters
mv	move	DE: Verschiebe eine Datei aus einem Verzeichnis in ein anderes Verzeichnis; Ändere den Namen einer Datei EN: Move a file from one directory to another; Change the name of a file
rm	remove	DE: Lösche eine Datei. **Vorsicht!** Die Datei ist dann unwiederbringlich verloren. EN: Delete a file. **Attention**! This file is gone and cannot be recovered.
cp	copy	DE: Kopiere eine Datei. EN: Copy a file.
touch		DE: Weise einer Datei das aktuelle Datum zu. EN: Assign the current date (day and time) to a file.
chmod	change modalities	DE: Verändere die Zugriffsrechte auf eine Datei für Benutzer, Gruppe, Andere (user, group, others). EN: Change the access rights of a file for user, group and others.
sort		DE: Sortiere die Zeilen einer Datei (alphabetisch). EN: Sort the lines in a file (alphabetically).
uniq –c		DE: Fasse nebeneinander liegende identische Zeilen

		einer Datei zusammen (und zähle die Häufigkeit jeder Zeile). EN: Merge adjacent lines in a file if they are identical (and count their frequencies).
grep		DE: Finde einen Suchbegriff in einer Zeile. EN: Find a search term in a line.
tr	transliterate	DE: Ersetze einen String mit einem anderen String. EN: Substitute a string with another string.

1. Use an editor to create a small test file with a few lines of text. Try the above commands with your test file.

2. In OLAT the SAC books "Die Alpen" from 1930 to 1939 are available for you in 3-column format (word form \t PoS tag \t Lemma). Download them to your computer.

2.1. Sort the lines in the 1935 volume alphabetically. Calculate the number of different words (~ lines) in the file. What are the 10 most frequent words?

2.2. Calculate the number of different words across all 10 volumes. What are the 10 most frequent words?

2.3. Investigate the 10 volumes with the help of **grep**. Example questions:

- How often does the word "*Freund*" (= the word form) appear in the 10 volumes? How often does the word "*Kamerad*" occur?
- How often does the lemma "*Freund*" or the lemma "*Kamerad*" (= the base forms) appear in the 10 volumes?
- In which volume does the lemma "*Freund*" occur most often?
- Calculate the number of different words in one file (in all files). What are the 10 most frequent words?
- How many verbs occur in the 10 volumes?
- How many German adjectives and how many French adjectives occur in the 10 volumes? Store all the different German adjectives from the 10 volumes without frequency information in one file (i.e. store each adjective only once).

Goals:

- Getting to know the handling of an input window (terminal)
- Getting to know the most important UNIX commands
- Learn how to connect commands (Pipe |) and how to redirect the output to a file (>)

Submission:

- This task is only for your own first steps into the course. You do not have to submit anything. ☺ If you get stuck, please ask the tutors.