Korpusbearbeitung Sommersemester 2021

Einführung

Florian Fink

15. April 2021

Organisatorisches

- Vorlesung Donnerstag 14-16 Uhr (ct)
- Vorlesung über Zoom: `https://lmu-munich.zoom.us/j/8366632112?pwd=cWc3ck5ML0t1c0VnUTZ2Zit2aUpFdz09`
- Homepage des Kurses `cis-kb21.github.io`
- Folien, Übungsaufgaben und Videos auf der Hompepage
- Bearbeitung der Übungsaufgaben ist freiwillig
- Besprechung der Übungsaufgaben in der Vorlesung
- Termin Klausur: 15.07.2021 (voraussichtlich)
- Bei Fragen Email an `kb21@cis.lmu.de`

Überblick

- Shell und Shell-Skripte
- Unix-Werkzeuge
- awk und sed
- Kodierungen
- Dateiformate
- verschiedene Korpora
- POS-Tagging (Tree-Tagger)
- ...

Unix-Shells

- (interaktive) Kommandozeileninterpreter
- verschiedene Shells mit unterschiedlicher Syntax(**?**):
  - `sh` die ursprüngliche *Bourne shell*
  - `bash` die *Bourne-again shell*
  - `zsh` die *z-shell*
  - `fish` die *friendly interactive shell*
  - `dash` die *Debian Almquist shell*
  - ...
- Unixoide Shells (insbesondere die `bash`) verfügbar für OSX und Windows (wsl1/2)

Unix-Umgebung

- die Unix-Umgebung besteht aus einer Vielzahl kleiner, vielseitiger Programme(**?**)
- Programme können flexibel kombiniert werden um komplexere Aufgaben zu bewältigen
- Programme für verschiedene Aufgaben(**?**):
  - Dateiverwaltung
  - Textverarbeitung
  - Datenverarbeitung
  - Benutzerverwaltung
  - Netzwerkverwaltung
  - ...

Interaktive Kommandozeilenumgebung

- die Shell bietet eine interaktive Umgebung um Befehle auszuführen
- Eingabezeilen werden an Leerzeichen in Token aufgetrennt
- einzelne Token (Befehle) werden ausgeführt
- es stehen verschiedene Tastaturkürzel für die interaktive Eingabe zur Verfügung

Tastaturkürzel

- `CTRL+k` schneidet Text vom Cursor bis zum Zeilenende aus (kill)
- `CTRL+u` schneidet Text vom Cursor bis zum Zeilenanfang aus
- `CTRL+y` fügt ausgeschnittenen Text am Cursor ein (yank)
- `CTRL+a` setzt den Kursor an den Zeilenanfang
- `CTRL+f` / `RIGHT` bewegt den Cursor nach rechts
- `CTRL+b` / `LEFT` bewegt den Cursor nach links
- `CTRL+p` / `UP` geht einen Schritt rückwärts in der Befehlsgeschichte
- `CTRL+n` / `DOWN` geht einen Schritt vorwärts in der Befehlsgeschichte
- `CTRL+x CTRL+e` öffnet einen Editor um einen Befehl zu editieren
- ...

Laufzeitumgebung

- beim Starten einer Shell-Sitzung werden verschiedene Variablen in der Laufzeitumgebung gesetzt
- Ausgabe der Laufzeitumgebung mit `env`
- Programme und Shell-Skripte erben die Laufzeitumgebung
- wichtige Variablen:
  - `PATH` Liste von Verzeichnissen, in denen nach Programmen gesucht wird (separiert durch `:`)
  - `HOME` Pfad des Benutzerverzeichnis
  - `EDITOR` Standardeditor
  - `USER` Benutzername
  - `SHELL` Standard-Shell
  - `LANG` Spracheinstellung

Shell-Skripte

- interaktive Befehle können auch in *Shell-Skripten* zusammengefasst und ausgeführt werden
- Shell-Skripte werden zeilenweise gelesen und abgearbeitet
- vor allem geeigntet für kurze Hilfsprogramme
- vor allem geeignet zur einfachen Stringverarbeitung; numerische Anwendungen sind nur sehr eingeschränkt möglich
- Shells bieten auch Möglichkeiten Verzweigungen und Schleifen zu verwenden (`if`, `case`, `for`...)
- Listen und assoziative Listen sind vorhanden (seit `bash 4.0`) aber sehr arkane Syntax