1.1 Recap last Lecture

- Einstieg in Shell
  - Verzeichnisbaum, Erstellen von Files/Ordner
  - Piping für komplexere Operationen
- Übungen ok? technische Fragen?
- letztes Mal inhaltliche Zumutung, heute erste inhaltlich interessante Analysen
- ähnliches Tempo, dafür mehr Zeit zum Üben

1.2 Get around in your filesystem :evergreen_tree:
1.3 Outline

- Frequenzanalysen = Schweizer Taschenmesser
  - äusserst effektiv
- Ziel: mehr Übungszeit
- Syntax nicht merken, Wichtiges werdet ihr schlussendlich erinnern

1.4 When politics changes, language changes.

- Positionierung Parteien im politischen Raum über Zeit
- Gleiche Parteien, neue Ziele. Also doch nicht so gleich!
- Wie erkenne ich semantische Veränderungen?
  - hier: Abstimmungsparolen von Parteien ausgewertet
  - Welche Ziele/Ideologien stehen dahinter? –> Texte fundamental
- Wenn Politik ändert, ändert sich Sprache
  - oder gerade umgekehrtes zeitliches Verhältnis
  - in Politik werden Narrative erprobt

1.5 Processing a Text Collection

- Start sehr oft Kommandozeile (z.B. Datenextraktion), dann Auswertung in Python
- txt-files erste Stufe bei Datensatzerstellung
- Daten existieren viele, Datensätze eher wenige
- bei Datensatz
  - Python praktischer
  - Dokument in Zelle in tsv/csv-file
- vorerst arbeiten wir nur mit txt files

2.5 Counting Things
2.6 Frequency Analysis

- Häufigkeit indiziert Form von Relevanz
- in Häufigkeitsanalyse sind Worte kontextlos
  - BoW = Sack mit Wörtern

  - Approach schmerzt aus sozialwissenschaftlicher Perspektive
  - Verlust Ambiguitäten = Nachteil // radikale Vereinfachung (einfaches Zählen) = grösster Vorteil
- theoetische Übersicht von Approaches später im Seminar
  - Kontrolle, was dahinter steht
- ähnlich wie Google Ngram, aber eigene Daten

2.7 Key Figures of Texts

- zuerst Charakterisierung Datenquelle, nicht nur Inhalt
- Zahlen für einzelne Dokumente und aggregiert auf Sammlung

2.8 Word Occurrences
3.8 common egrep options:
4.8 -i search case-insensitive
5.8 -r search recursively in all subfolders
6.8 –colour highlight matches
7.8 –context 2 show 2 lines above/below match

- options
  - ignore case
  - recursive / specific files
- Dateinamen als Filter benutzen
  - Quelle/Jahr
  - egrep -ir " daten" `*svp*.txt`
- wc als Alternative
- zeige in Kurs-Repo
  - egrep -irc –colour –context 3 "data" lectures/md | sort

cd /home/alex/KED2023/materials/data/swiss_party_programmes

egrep -irc "ökologisch" .

7.9 Word Frequencies
8.9 piping steps to get word frequencies
9.9 explanation of individual steps:

- Zweck: Häufigkeiten aller Wörter
- kein direkter Befehl -> Kombinieren von Befehlen (modular)
- Befehle erklären
  - Zusammenfassen gleicher Zeilen mit uniq
- Newline Character
- Aggregation extrem flexibel
  - anderer Text, alle Texte (*)
- Frage an Klasse: häufigstes Wort SVP?
  - Schweiz, Bürger etc.: national, männlich

- `cat materials/data/swiss_party_programmes/txt/svp_programmes/*txt | tr " " "\n" | sort | uniq -c | sort -h`

9.10 Word Frequencies

- Korpus = Textsammlung
- absolut nur, wenn grösserer Output (z.B. mehr Flyers) mitgemessen werden soll

9.11 Convert Stats into Dataset
10.11 convert word frequencies into tsv-file
11.11 additional step: replace a sequence of spaces with a tabulator

- -s alle Leerschläge durch Tabulator ersetzen
- relative frequency in Excel

11.12 In-class: Matching and counting

Pause

12.12 Preprocessing
12.13 Common Preprocessing

- Preprocessing für bessere Resultate
- Regex nächste Woche

12.14 Lowercasing

- Grossschreibung Satzanfang

12.15 Removing and Replacing Symbols

- Es gibt Zeichenklassen für Buchstaben, Zahlen und Interpunktion
- löscht alle Einzelzeichen in Text (keine Sequenzen)
- Interpunktion wird sehr oft entfernt, da sowieso Kontext verloren geht in BoW

12.16 Standard Preprocessing
13.16 lowercase, no punctuation, no digits

- Kleinschreibung , keine Interpunktion, keine Zahlen
- standardmässige Repräsentation in BoW (hier noch mit Reihenfolge)

13.17 Join Lines

- harte Zeilenumbrüche entfernen
- squeeze repeated newline and replace with a single whitespace

13.18 Trim Lines
13.19 Splitting Files
14.19 splits file at every delimiter into a stand-alone file
14.20 Check Differences between Files
15.20 show differences side-by-side and only differing lines
15.21 Where there is a shell,there is a way. :thumbsup:

- Zusammenfassung
  - Nach Filesystem, nun auch Bearbeiten, Zählen
- Shell = flexibles + mächtiges Werkzeug durch Kombinieren von mehreren Commands
- Stackoverflow liefert Antworten auf ein Problem

15.22 Organizing Code

- Version Managment Software
  - ähnlich Änderungsmodus in Word
- Nutzen
  - für moderne Software-Entwicklung nicht wegzudenken
  - neuerdings für Tracking wissenschaftlicher Arbeiten
- Repository = Ablage

16.22 Questions?
16.23 In-class: Getting ready
16.24 In-class: Analyzing Swiss Party Programmes I
16.25 In-class: Analyzing Swiss Party Programmes II
16.26 Additional Resources