### 1.1 Recap last Lecture

- Regex für Extraktion + Säubern
  - man muss nur ungefähr wissen wonach suchen
  - generalisierte Form = Muster
- Literale = Zeichen steht für tatsächliches Zeichen (buchstabentreu)
- Meta-Zeichen = Zeichen mit spezieller Bedeutung
- Fragen zu RegEx oder Übung?

### 1.2 Outline

- heute letzte Sitzung zu Kommandozeile
- zweiteilige Sitzung mit wenig Technischem
  - existierende Daten, eigene Daten
- interessannte Datasets für die Sozialwissenschaften
  - es gibt allerdings nicht viele
  - zumeist eigene Daten präparieren
- eigene Textdaten nutzen unabhängig von
  - Formaten
  - historischem Kontext (digital native)
- Überlegen, was für eine Analyse in MiniProject

### 2.2 Data Sources
### 2.3 What Data Sources are there?
### 2.4 Interesting Publishers

- Ressourcen gelistet auf ZHBLuzern
  - Zugang tlw. über ezproxy
- Nexis vielleicht spannendste Quelle für Analyse soziale Probleme
- Wieso Literatur? –> Zeitgeist
  - genderspezifische Sprache, Verweise Natur/Umwelt
- Constellate
  - kurze Demo von Constellate
  - brandneue Platform
  - einfache Zusammenstellung von JSTOR Artikeln
  - sehr gute Metadaten
  - auch gut für schnelle Recherchen ohne Download

### 2.5 Dataset Search

- moderne Wissenschaft veröffentlicht nicht nur Papers, sondern auch Daten
  - computergestützte Textanalyse ist aber immer noch Nische
- Suchmaschinen für Datensätze

  - allerlei Datensätze, primär aus Wissenschaft
- UZH hat Institut Computerlinguistik
  - verschiedene Korpora
  - Credit Suisse PDF Bulletin Corpus

### 2.6 Some great historical Corpora

- sehr wenige standardisierte Datasets
- nicht wie bei Survey-Forschung, numerischer Daten aus Politik und Ökonomie

### 2.7 Online Computational Text Analysis

- Datenanalysen online durchführen
- Absicherung über andere Quellen
- Impresso: Complete re-digitization of the NZZ (together with the Zurich Central Library and Swiss National Library)

### 2.8 Search Techniques

- Quotes für Wörter die zusammen gehören
- Boolean Search
  - OR / AND

### 2.9 Data is property :no_entry_sign:

- Zugang zu Daten nicht immer einfach
  - open data unterschiedlich unterstützt
- Datenbereitstellung oftmals Teil von Geschäftsmodell
  - dann restriktiv
- oftmals ist Verwendung nicht geregelt
  - nutzt Graubereich

### 3.9 Preparing your own Data
### 3.10 A world for humans ..... and a jungle of file formats.

- extrem viele File-Typen
- mühsam, aber es gibt einfache Tools für Umwandlung

### 3.11 Common Conversions

- PDF ist Publikationsstandard
  - neue (digital) vs. alte (scans)
  - Kriterium: Suche möglich?
- anschliessend Schritte zur Umwandlung der wichtigsten Formate
- Keine Konzepte lernen, wie bei RegEx
  - nur welches Tool, für welche Umwandlung

  - mehr oder weniger copy-paste
- **Pause (etwas früher)**

**3.12 Conversion of DOCX**
**4.12 convert docx to txt**

- pandoc ist ein fast-alles Könner für Dokumentkonversion
  - kann auch html konvertieren: pandoc slides/KED2023_01.html -t plain
- zusätzliche Installation
- Nexis = News-Datenbank
  - freier Zugang ezproxy
  - kennen ezproxy alle?

**4.13 Conversion of native PDF**
**5.13 convert native pdf to txt**

- pdftotext: Name ist Programm
  - Outputfilename kann nicht spezifiziert werden
- dieselben Parteiprogramme, die wir schon analysiert haben
- Layout kann Extraktion erschweren
  - Spalten/Tabelle
- Häufigkeitsanalysen von Wörter sind robust, Struktur egal

**5.14 Optical Character Recognition (OCR)**

- tatsächlicher Buchstabe, nicht nur Bild davon
- Zwischenschritt Verbesserung Kontrast, B/W
- technisch Deep-Learning, nicht weiter von Bedeutung
- früher teure Programme, heute sogar iPhone
  - für viele Dokumente jedoch nicht geeignet

**5.15 Conversion of digitalized PDF**
**6.15 convert scanned pdf to tiff, control quality with parameters**
**7.15 run OCR for German ("eng" for English, "fra" for French etc.)**

- Zwei Schritte: Bildumwandlung + OCR
- tesseract funktioniert für viele Bildformate
  - nicht direkt für PDF
- Beispiel: Kassenbon fotografieren & mit Regex parsen
  - Wirtschaftswissenschaften: indexierter Warenkorb

**7.16 Configure ImageMagick properly**
**8.16 disable security policy for Windows**
**9.16 increase memory limits**
**9.17 LifeHack: Make a PDF searchable**
**10.17 output searchable pdf instead of txt**

- Output als PDF statt Text
- für Suchen/Zitate rauskopieren
- convert hier mit Kompression, da PDFs zu gross werden ansonsten

**10.18 Scraping PDF from Websites**
**11.18 get a single file**
**12.18 get all linked pdf from a single webpage**
**13.18 –accept FORMAT_OF_YOUR_INTEREST**
**14.18 –directory-prefix YOUR_OUTPUT_DIRECTORY**

- bis hierher: Wie Daten in txt Format bringen
- jetzt Download automatisieren
  - Vorteil: schneller systematischer Download & Dokumentation von Quellen
- Haupt-URL angeben, wo PDF verlinkt sind
- Scraping von Blogs möglich über Python
  - nicht Teil von Seminar
- nicht auf alle Argumente eingehen

**14.19 Interesting Resources**
**14.20 Basics of Batch Processing**
**15.20 loop over all txt files**

- Batch Processing = gleiche Operation durchführen für alle Files
- Erklären von Loop/Schleife und Variable
  - Wildcard zur Selektion > Liste von Files > Variable
- for-loop wichtiges Programmierkonzept
- Tabulator fürs Einrücken

**15.21 Perform OCR for many PDF**

- sehr ähnlich wie vorher, nur für jedes einzelne File jetzt

**15.22 Preprocessing → RegEx**

- Aufbereitung unterschiedlich aufwendig
- für schnelle Analyse nicht notwendig
- nun alles da für Mini-Project, ausser wenn Lösung in Python

**16.22 Questions?**
**16.23 In-class: Exercises I**
**16.24 In-class: Exercises II**
**16.25 In-class: Exercises III**
**16.26 Resources**
**16.27 Have a nice Easter break!**