http://bit.ly/kleineuebung

1. Erstellt eine kleine Grafik der Zutatenanteile von Bier
2. Warum kann der Computer einige Zutaten nicht verwenden?

CAS Datenjournalismus

1_Einstieg ins Programmieren

http://bit.ly/Tag2_CAS_Datenjournalismus

Programm

- Github
- Command Line (Terminal)
- Computer sprechen lassen, PDFs cracken et al.
- Workflow
- Jupyter Notebook (strings, integers, floats)
- Feedback

Feedback gestern

- Spreadsheets, zu schnell

Wechseln wir auf Github

[MAZ Github Repo]

git clone & git pull

- Auf github in eine Repo gehen
- “clone & download” Knopf suchen, kopieren
- Nun auf der command line: git clone XXXXXURLXXXXXX
- In den Folder navigieren und git pull XXXXXURLXXXXXX

Die eigene Repo

- Start a Project
- Beschreibung: öffentlich
- .gitignore, ls -a
- Lizenz
- Mit Readme

Die eigene Repo

Einmaliges Setup:

- `git config --global user.name "Peter Müller"`, derselbe Nutzername, den wir auf Github benutzt haben.
- `git config --global user.email deine@email.com`, dieselbe Email, die ihr auf Github gebraucht habt.
- `git config --global user.password "your password"`
- Mit `git config --list` könnt ihr alles überprüfen.

Die eigene Repo

Nun clonen wir die eigene Repo

Die eigene Repo

- Ergänzen wir das Readme file
- Abspeichern
- `git add .`
- `git commit -m "nachricht"`
- `git push`

Übung 1

- Kreiert im Text-Editor Atom lokal ein neues ein neues File.
- Nennt das File "01Übung.md".
- Schreibt ein paar Sätze rein.
- Dann pusht es auf Github.
- Nun geht ihr auf Slack.
- Und teilt eure Repo mit allen.
- Genauso werdet ihr Hausaufgaben lösen.

Say

- say hello
- man say
- q
- say -v ?
- say -v Anna hello, wie geht es dir
- say -f file.txt
- say -v Anna -f file.txt
- say -v Anna -f file.txt -r 120

PDFCrack

- PDF runterladen
- brew install pdfcrack
- pdfcrack -f topsecret.pdf
- Dokumentation

So many more things...

- Eine Liste mit viel mehr Terminal packages und Befehlen

Crontab, Automatisierung

- `export EDITOR=nano`
- crontab -e
- Und wir geben die fünf * Zeichen ein, und dann `say "hello you"`
- Der Computer wird nun jede Minute "hello you" sagen.
- crontab.guru

Übung 2

1. Baut einen Crontab, der alle 2 Stunden "Hello You" sagt
2. Baut einen Crontab, der jeden Tag um 11.30 Uhr "Hello You" sagt
3. Baut einen Crontab, der jeden Montag um 9.30 "Aufwachen, die Woche hat begonnen" in einer deutschen Stimme sagt.
4. Kreiert ein File names “02Übung.md” und pusht es auf Github und teilt das File via Slack

Kaffeepause

“Big Data” und die Commandline

- Gehen wir auf die Website des Nationalfonds: http://p3.snf.ch
- Und sucht die Daten der Fonds
- Ladet das P3_GrantExport.csv File auf euere Gerät runter

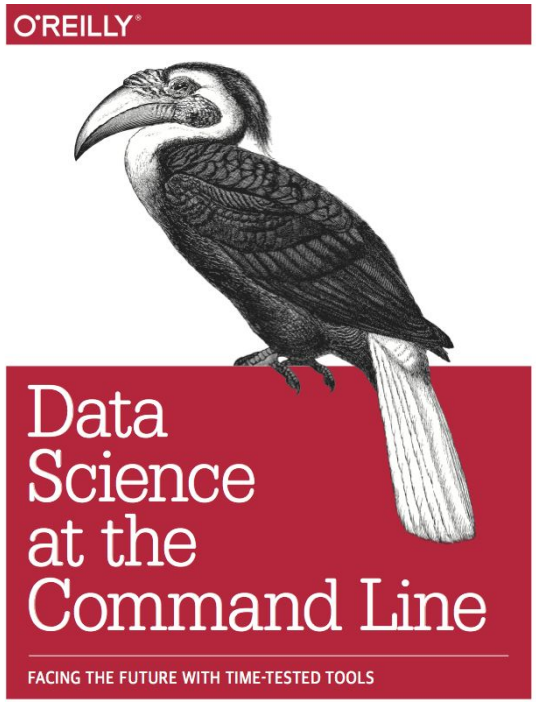
Counting

- `wc P3_GrantExport.csv`
- `wc -l P3_GrantExport.csv`
- `wc -w P3_GrantExport.csv`
- `wc -m P3P3_GrantExport.csv`
- `man wc`

grep & pipe

```
- grep "Geschichte" P3_GrantExport.csv
- grep "Geschichte" P3_GrantExport.csv
  > geschichte.csv
- grep "Geschichte" P3_GrantExport.csv
  | wc -l
```

Data Sience from the Commandline



Jeroen Janssens

Übung 3

1. In wievielen Projekte wurde die ETH erwähnt?
2. In wievielen die Stadt Basel?
3. Wieviele nannten Psychologie?
4. Und wieviele die Zoologie?
5. Legt die Ergebnisse in ein File unter dem Namen “03Übung.md” und lädt das File wieder Github.

Mittagspause

Zusatzpackages wget

- `brew install wget`
- `wget --version`
- `Man wget`
- `wget https://www.balthasar-glaettli.ch/`
- `wget --user-agent=Firefox https://www.facebook.com/glaettli.ch`
- `wget --recursive --no-clobber --page-requisites --html-extension --convert-links --restrict-file-names=windows --no-parent https://www.balthasar-glaettli.ch`

Übung 4

1. Wähle eine Website Deiner Wahl aus, und lade deren Frontseite mit Wget auf den Desktop runter. Speichere die Zeile, die Du dafür verwendet hast, in einem Text-Dokument ab.
2. Weisst Du noch, wie wir Crontab entwickelt haben? Entwickle ein Crontab, der alle 15 Minuten eine Website besucht und deren Inhalt auf Deinem Gerät abspeichert. Lege auch diese Zeile in das Textdokument, nenen es “04Übung.md” ab und lade das File in Deinen Github-Konto.

Wget und Datennamen

- `wget -O glaettlibeispiel.htm https://www.balthasar-glaettli.ch/`
- `date`
- `wget -O=glättli$(date +%Y-%m-%d%H:%M:%S).htm https://www.balthasar-glaettli.ch`

CSV Kit - Installation, Formatierung

- `In virtual env: pip install csvkit`
- Kein manual, aber ein hier mehr
  https://csvkit.readthedocs.io/en/1.0.3/
- `csvlook P3_GrantExport.csv`
- `(neues Fenster) csvcut -n P3_GrantExport.csv`
- `csvformat -d ";" P3_GrantExport.csv > data.csv`

CSV-Kit Auswahl, Übersicht

- `csvcut -n data.csv`
- `csvcut -c 5,8,9,10,17 data.csv > data_selected.csv`
- `csvstat data_selected.csv`

Kaffeepause

Work Flow

- Nichts in der Side-Bar nichts, was ich nicht täglich verwende.
- Mit den verschiedenen Screens zu arbeiten, wo ich den ersten Screen immer frei behalte.
- Um die Screen-Ansicht anzuzeigen, drei Finger und nach oben swipen
- Von einem Screen zum anderen zu wechseln, mit drei Fingern rechts und links swipen
- Mit zwei Fingern neue Terminals aufmachen. Sie lege ich meist auf ein eigenes Fenster.
- “cd /”, die oberste Stufe zu erreichen.
- Kann vorkommen, dass der Computer hängt: Activity Monitor griffbereit

Weitere Dienste und Organisation

- Evernote, um Code-Stücke, die ich dauernd verwende, abzulegen.
- Ein Ordner auf dem Desktop mit allen datengetriebenen Projekten
- Ein zweiter Ordner, in dem ich alle Ordner habe, die ich mit Github synchron halte.
- Grosse Datensammlungen auf Dropbox ab. Oder mit Github lhs. Das ist ein kostenpflichtiger Dienst, um grössere Datenmengen abzulegen.
- Kleinere Datensammlungen bei Google Spreadsheets.
- Und dann eine kleine Server-Farm bei Digital Ocean.

Jupyter Notebook

Print, Input, Strings, Integers, Floats, Variabeln.

Hausaufgaben

Nun wollen wir mit unsere Commandline Skills mit einem richtig grossen Datensatz üben. Wir arbeitn mit den Google Copyright Infringement Meldungen. Das sind über 3 Milliarden Meldungen, die in den letzten Jahren bei Google eingegangen sind. Ihr findet sie hier. Ganze 3,2 GB. Und hier könnt ihr die Daten runterladen.

Hausaufgaben

- Nachdem ihr die Daten herunter geladen habt, sucht den Befehl, um die Daten zu entzippen.
- Lesen wir das README
- Schauen wir uns die Kopfzeile der "Domains" Daten an, damit ihr wisst, was sich in den drei Datensammlungen überhaupt befindet.
- Suchen wir alle ".ch" Domains in den Datensatz. (Bei 3 Milliarden linien kann das eine Weile dauern.)
- Speichern wir das ab als ch_domains.csv
- Und nun zählen wir die Linien in dieser Kopfzeile
- Rufen wir csvstat auf, um die häufigsten Schweizer Domains zu zählen.

Lektüre

- Why Learning to Code Is So Frustrating
- Python has brought computer programming to a vast new audience
- Data Journalism in 2017 (Google Report)

Feedbacks online

URL:	http://www.maz.ch/feedback
Kursnummer:	J 99343
Kurstitel:	Themen finden und datenjournalistische Recherchen planen
Dozierende:	Barnaby Skinner, Dominique Strebel

Formular für Dozenten:
http://www.maz.ch/dozfeedback