CEDUS WORKSHOP: TAG 2

EINFÜHRUNG IN DATA SCIENCE

HEUTIGE AGENDA

- Python: fortgeschrittene Konzepte
- Was ist Data Science?
- die Data-Science-Pipeline
- Case Study: Kickstarter Datensatz

"We are drowning in information but starved for knowledge."

John Naisbitt - Megatrends

MOTIVATION: DATA SCIENCE

- Wir produzieren *riesige* Mengen Daten
  - IBM schätzt 2.5 Quintillionen Byte/Tag
  - Umgerechnet rund 500 Milliarden Bücher, 400 Seiten, DIN A4
- Daten sind größtenteils roh, Wissen muss gefördert werden

Forbes: How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read

DEFINITION: DATA SCIENCE

Data Science (von englisch data „Daten“ und science „Wissenschaft“) bezeichnet generell die Extraktion von Wissen aus Daten.

https://de.wikipedia.org/wiki/Data_Science

- Interdisziplinäres Feld:
  Informatik ↔ Statistik ↔Domänenwissen
- Prozess der aus Daten Wissen generiert
  - Data-Science-Pipeline

DATA SCIENCE - PIPELINE

- **O.S.E.M.N**
  - **O**btain Data
  - **S**crub / Clean Data
  - **E**xplore Data
  - **M**odel / Visualize Data
  - I**N**terpret our Data

Medium: A beginner's Guide to the Data Science Pipeline

BEISPIEL: SPIEGELMINING

Link zum Video

OBTAIN DATA

- **(Semi-)automatisiertes** Sammeln von Daten
- Oftmals auch: Data Mining
- Herausforderungen/Fragestellungen:
  - Datenformate sollten **maschinenlesbar** sein
  - Woher kommen meine Daten?
    - APIs
    - Crawling
    - Open Data
  - In welcher Qualität sammel ich Daten?

DATENFORMATE

- Legen fest wie Daten formatiert sind
- Aus dem Alltag bekannt: .jpg, .pptx, .gif, .docx
- Maschinenlesbare Datenformate
  - CSV (**C**omma **S**eparated **V**alues)
  - JSON (**J**ava **S**cript **O**bject **N**otation)
  - Markup-Formate (X**ML**, HT**ML**, YA**ML**)

API

- **A**pplication **P**rogram **I**nterface
- Schnittstelle, "Datenzapfhahn"
- spezielle URL -> maschinenlesbare Antwort (oft .json)

Beispiel: Wetterdaten openweathermap.org

OPEN DATA

- Datenquellen die offen zur Verfügung stehen
- Open Knowledge Foundation - okfn.de/projekte
- Nachteil: statisch, manuell eingepflegt

Beispiel: open.nrw - Behindertenparkplätze Düsseldorf

CRAWLING

- Alternative wenn weder API noch Open Data verfügbar
- Individuallösung bauen für Daten einer konkreten Quelle bauen
- Nachteil: Hoher Wartungsbedarf; keine Funktionsgarantie
- Webseite ändert Layout
  - Crawler muss umgebaut werden

BEISPIEL: SPIEGELMINING

Link zum Videoausschnitt

(WEB)CRAWLING-TOOLS FÜR PYTHON

- Crawler lassen sich mit Python umsetzen
- Hilfreiche Bibliotheken:
  - `requests` - Ermöglicht Aufrufen/Zwischenspeichern von URL-Inhalten
  - `BeautifulSoup4` - HTML-Parser, extrahiert Informationen aus Webdokumenten

MINIDEMO: RSS-FEED CRAWLEN

SCRUB / CLEAN DATA

- Daten sind fehlerhaft/unvollständig/schlecht formatiert
- Beispiele:
  - Sensorik misst falsche Werte (Luftfeuchtigkeit: -4%)
  - Unterschiedliche Formate 05.09.2018 vs. 09/05/2018 vs. 2018/09/05
  - Fehlende Sammelzeiträum (vgl. Spiegelmining)
- Auch: Data Munging / Data Wrangling
- Ziel: Qualität der (Roh-)daten erhöhen!

EXPLORE DATA

- Um Wissen zu extrahieren brauche ich einen Überblick
- Typische Arbeitschritte:
  - Datendichte reduzieren (Filtern)
  - Überblick durch Visualisierung verschaffen (Plotting)
  - Ausreißer finden
- Durch Exploration ergeben sich Untersuchungspunkte

BEISPIEL: SPIEGELMINING

Link zum Videoausschnitt

MODEL / VISUALIZE DATA

- Daten müssen modelliert werden
- Typische Aufgaben:
  - Features/Variablen kombinieren
  - Variablen korrelieren
  - Visualisierungsform anpassen
  - Modelle/Algorithmen anwenden (bspw. Clustermodelle)
- Ziel: präzisere Interpretation

BEISPIEL: SPIEGELMINING

Link zum Videoausschnitt

INTERPRET DATA

- Daten in Wissen übersetzen
- Handlung ableiten
- Kritischster Schritt: Daten können falsch interpretiert werden

Die Seite spuriouscorrelations lässt dich absurde Statistiken zusammenkorrelieren

BEISPIEL: SPIEGELMINING

Link zum Videoausschnitt

DATA SCIENCE MIT PYTHON

NÜTZLICHE BIBLIOTHEKEN:

- `pandas` - Daten einlesen, verarbeiten, filtern
- `matplotlib` - Daten visualisieren
- `scipy`- Lineare Algebra / Optimierung
- `scikit-learn`- Bibliothek fürs Machine Learning
- `nltk` - Bibliothek fürs Natural Language Processing

KAGGLE - DATASCIENCE-PLATTFORM

- "Your Home for Data Science"
- Bietet.. :
  - Lernmaterialen
  - Notebooks zum Nachlesen & Nachbauen
  - von der Community zur Verfügung gestellte Datensätze

WALKTHROUGH: KAGGLE

Kaggle - "Your Home for Data Science"

CASE-STUDY: KICKSTARTER-DATENSATZ

FRAGEN? FEEDBACK?