Themenblock Datenbeschaffung und -aufbereitung: 04_02Beschaffung_Datenbeschaffung

Grundlagen der Datenbeschaffung

Erarbeitet von
Dr. Ann-Kathrin Selker

Lernziele 1
Inhalt 2
Datenmengen 2
Kuratierte Datensätze 2
Webcrawler 3
Quellen 4
Weiterführendes Material 4
Disclaimer 5

Lernziele

- Du kannst den Datenbeschaffungsprozess erklären
- Du kannst Webcrawler und ihre Einsatzgebiete erklären
- Du kannst Datenquellen nennen

Projekt ai4all „KI für Alle"
HeiCAD –Heine Center for Artificial Intelligence and Data Science

Inhalt

Bevor mit dem Training einer Maschine begonnen werden kann, müssen erst einmal Trainingsdaten gesammelt werden. Doch wo bekommt man die benötigte Menge an Daten überhaupt her?

Datenmengen

Im Zeitalter von Big Data gibt es Daten überall. Es wird geschätzt, dass im Jahr 2024 Daten in der Größenordnung von über 100 Zettabyte generiert werden.

Quelle [1]

Damit du diese Zahl besser einschätzen kannst, gucken wir uns mal kurz im Vergleich die Größen einiger Datentypen an. Dieses Bild hat eine Größe von über 1 MB.

Einblendung Katzenbild

Ein modernes Computerspiel hat eine Größe von mehreren Gigabyte. Externe Festplatten belaufen sich inzwischen auf Größen im unteren Terabyte-Bereich. Um 100 Zettabyte zu erreichen, benötigst du dann ungefähr 100 Milliarden Festplatten. Du könntest etwa 100 Billionen moderne Computerspiele speichern, oder stattdessen etwa 100 Billiarden Fotos digital aufbewahren.

Wahnsinn, oder? Bei dem Großteil dieser Daten handelt es sich um unstrukturierte Daten, zum Beispiel Videos oder Bilder. Alleine Videos sind geschätzt für mehr als die Hälfte des aktuellen Internetverkehrs verantwortlich. Pro Minute werden über 200 Millionen E-Mails versendet und fast 6 Millionen Google-Suchen durchgeführt.

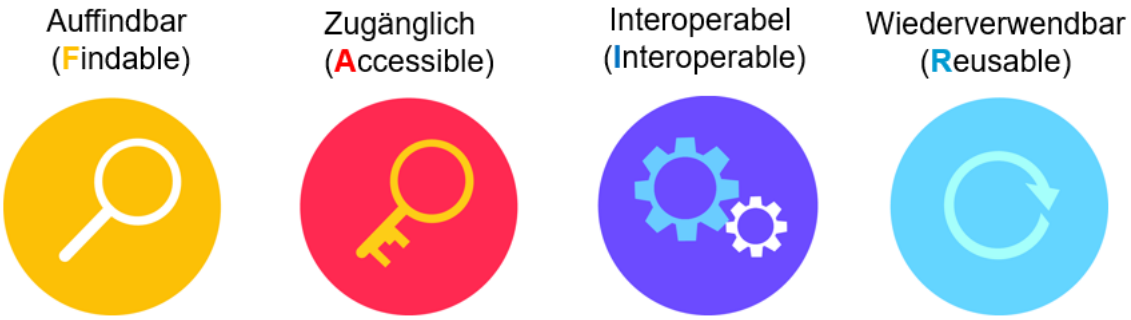
Quelle [2]

Kuratierte Datensätze

Bei diesen Datenmengen sollte es doch wohl nicht so schwer sein, passende Daten zu beschaffen, oder? Genau darin liegt aber häufig das Problem. Du musst dich erst einmal durch die Flut an Internetdaten wühlen, um an die Daten zu kommen, die dich für deinen Anwendungsfall eigentlich interessieren.

Glücklicherweise gibt es viele Stellen, an denen Daten offen und kuratiert zur Verfügung gestellt werden. Viele Regierungen und öffentliche Einrichtungen bieten von ihnen gesammelte Datensätze an. In Deutschland kannst du dir zum Beispiel einmal die Seite des Statistischen Bundesamtes ansehen. Für Forschungsdaten gibt es je nach Disziplin viele Forschungsdatenrepositorien. Diese Quellen haben den Vorteil, dass die Daten bereits mit Metadaten versehen wurden. Unter Metadaten verstehen wir strukturierte Informationen über unsere Daten bzw. Datenpunkte. So können bei einem Textdokument Autor*in, Titel

und Erscheinungsdatum, aber auch Angaben wie Dateityp oder Lizenzen als Metadaten vorliegen. Diese Daten sind auffindbar, zugänglich, interoperabel und wiederverwendbar – FAIR eben.



Einblendung FAIR-Prinzip (**Quelle [3]**)

Um einfacher an die Datensätze zu kommen, bieten viele Webseiten Schnittstellen an, sogenannte APIs (Application Programming Interfaces). Diese Schnittstellen erlauben es Anwendungen, Anfragen an diese Webseiten zu schicken, ohne dass du als Mensch vor dem Computer sitzen und die Anfragen erstellen musst. Wie immer musst du bei solchen Anfragen auf die Nutzungsbedingungen achten. So gibt es bestimmte Richtlinien, wie häufig und wie viele Daten auf einmal angefragt werden dürfen, welche Inhalte gespeichert werden dürfen usw.

Webcrawler

So viele Daten es auch schon als Datensätze gibt, für manche Anwendungsfälle ist leider doch nicht das Richtige dabei. Vor allem bei datenschutzrelevanten Daten führt häufig leider nichts daran vorbei, Datensätze zeitaufwändig selber zu erstellen. Aber auch für deine tolle neue Forschungsidee gibt es vielleicht noch niemanden, der entsprechende Datensätze zusammengestellt hat. Da hilft dann manchmal nur, dich selber durch das Netz zu wühlen. Immerhin musst du das nicht manuell machen, sondern kannst sogenannte Webcrawler verwenden. Bei Webcrawlern handelt es sich um Bots, also Computerprogramme, die automatisiert laufen. Sie durchsuchen Webseiten und speichern alle spezifizierten Informationen, die sie finden können. Dieser Vorgang nennt sich Web Scraping. Da die Auswertung der Seiten automatisch erfolgt, gibt es natürlich keine Garantie für die Vollständigkeit, Aktualität und Korrektheit der gesammelten Daten. Webcrawler sorgen auch für einen überdurchschnittlich hohen Verkehr bei Webseiten und müssen deshalb bedacht eingesetzt werden. Außerdem gibt es auch rechtliche Probleme zu bedenken, zum Beispiel in Bezug auf Urheberrecht und Datenschutz bei den gescrapten Inhalten.

Viel Erfolg bei deiner Suche nach den richtigen Daten! In diesem Video hast du schon einmal einige gute Anlaufstellen für Daten kennengelernt.

Quellen

Quelle [1] *Data growth worldwide 2010-2025*. (2023, 16. November). Statista. https://www.statista.com/statistics/871513/worldwide-data-created/

Quelle [2] *Global Internet Phenomena Report* (2023). Sandvine. https://www.sandvine.com/phenomena

Quelle [3] Paulina Halina Sieminska. CC-BY-SA 4.0. https://forschungsdaten.info/themen/veroeffentlichen-und-archivieren/faire-daten/

Weiterführendes Material

Du brauchst Daten? Schau doch zum Beispiel einmal hier vorbei:

- Forschungsdatenrepositorien
  - https://www.re3data.org/ (Register für Forschungsdatenrepositorien)
  - https://www.fdm.hhu.de/fdm-tools/repositorien
  - https://www.gesis.org
  - https://www.radar-service.eu/radar/de/home
  - UC Irvine Machine Learning Repository (https://archive.ics.uci.edu).
  - https://libguides.osl.state.or.us/data/repositories
  - https://datasetsearch.research.google.com
  - https://datadryad.org/stash
- Regierungen und öffentliche Einrichtungen
  - https://www.destatis.de/
  - https://www.deutsche-digitale-bibliothek.de
- Webdaten:
  - https://www.commoncrawl.org
  - https://www.archive.org
  - https://think.cs.vt.edu/corgis/
  - https://dataeuropa.gitlab.io/data-provider-manual/
  - https://www.lib.ncsu.edu/formats/teaching-and-learning-datasets
  - https://www.kaggle.com
- Open Data
  - https://de.wikipedia.org/wiki/Open_Data (Tabelle mit Links zu Projekten, die Daten offenlegen)

Diese Liste soll dir nur einen groben Überblick über frei verfügbare Datensätze geben. Erster Ansprechpartner für Daten ist normalerweise dein*e Projektleiter*in.

Disclaimer

Transkript zu dem Video „04 Datenbeschaffung und -aufbereitung: Grundlagen der Datenbeschaffung“, Ann-Kathrin Selker.