

AirBnB Webscraping – Christian Kälin

Zur Extraktion der Daten von AirBnB habe ich die Chrome Extension „ScrapeThings“ verwendet. Die Daten wurden in ein einzelnes CSV File zusammengefügt, mit den Juni Einträgen zuerst und die Oktober Daten nachfolgend, getrennt durch eine leere Zeile. Folgende Gründe waren ausschlaggebend für die Wahl dieses Tools:

1. Genauigkeit und Richtigkeit der Daten:

Bereits erste Tests mit ChatGPT zeigten, dass die mit diesem Tool gescrapten Daten nicht immer vollständig genau oder korrekt waren. Dies liegt in der Natur von LLM-Modellen, deren Funktionsweise für Nutzer intransparent ist. Dadurch lässt sich nicht nachvollziehen, nach welchen Kriterien oder Schritten ChatGPT beim Scraping vorgegangen ist.

2. Benutzerfreundlichkeit und Verifikation:

Durch die Verwendung der Chrome Extension „ScrapeThings“ kann klar nachvollzogen werden, welche Bereiche des User-Interfaces und welche Variablen konkret extrahiert werden. Zudem visualisiert das Tool direkt auf der Webseite, welche Items bereits gecrawlt wurden. Dies erhöht die Transparenz und erleichtert die Überprüfung der Datengenauigkeit.