Booking.com Webcrawler - Bedienungsanleitung

Raphael Prinz

(Geospatial Data-Mining, Seminar im WS21, raphael.prinz@online.uni-graz.at)

Inhalt

1 ALLGEMEINES

Dieser Webcrawler dient dazu automatisch die Koordinaten, Namen, Preise und Adressen von Hotels aus der Webseite Booking.com zu extrahieren. Dazu nutzt der Webcrawler ein durch Python gesteuertes Mozilla – Firefox Fenster. Die Koordinaten der Hotels sind nicht direkt in der Webseite enthalten, aber die Webseite sendet beim Öffnen der Kartenansicht einen GET-Request an eine Datenbank, die Antwort der Datenbank ist ein JSON-File, dass die Koordinaten enthält. Deshalb geht der Webcrawler selbstständig durch jede Seite und sammelt die JSON-Files. Um das zu tun, müssen die Knöpfe, die der Webcrawler benötigt um sich durch die Seiten durchzuklicken als CSS-Suche vorgegeben werden. Booking.com ändert jedoch regelmäßig die Namen dieser Website-Elemente, deshalb kann man in diesem Programm die Namen der Elemente, falls nötig abändern.

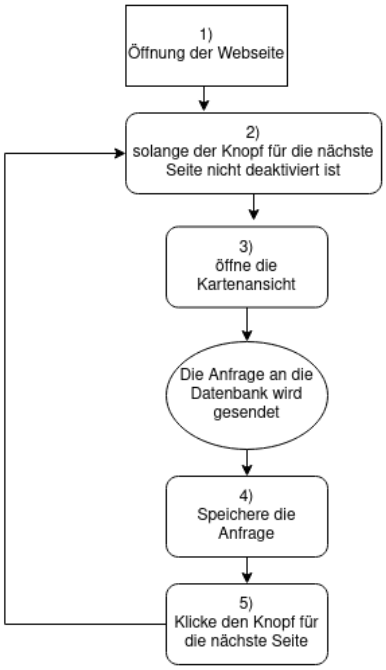

Figure 1: Programmablauf des Webcrawler (Quelle: eigene Darstellung, 2022)

2 PROGRAMMSTART

Das Programm kann mit einem Doppelklick auf „autoCrawler.exe“ gestartet werden. Die Datei „geckodriver“ ist für das Steuern von Firefox notwendig und muss im gleichen Ordner sein wie „autoCrawler.exe“

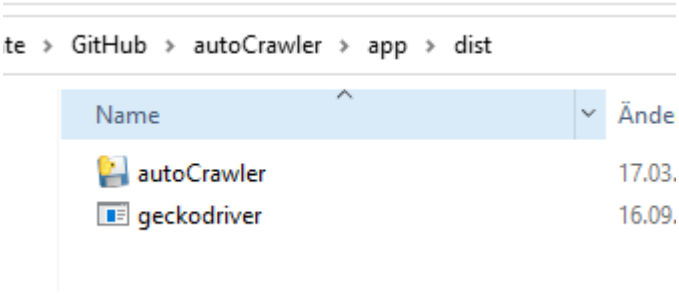

Figure 2: "autoCrawler.exe" Datei (Quelle: eigene Darstellung 2022)

Nach Start des Programmes werden zwei weitere Ordner erzeugt, in denen Zwischenergebnisse gespeichert werden, als Nutzer muss man jedoch nicht mit diesen interagieren.

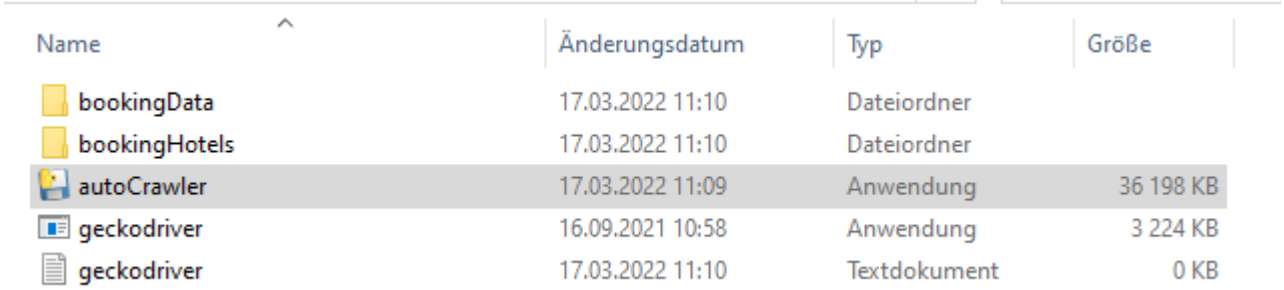

Figure 3: Neu erstellten Ordner "bookingData" und "bookingHotels" (Quelle: eigene Darstellung 2022)

Nach Öffnung des Programmes wird Firefox automatisch geöffnet, zwei Tabs werden aufgerufen, eines ist die Bedienoberfläche des Webcrawlers, das andere die Booking.com Webseite.

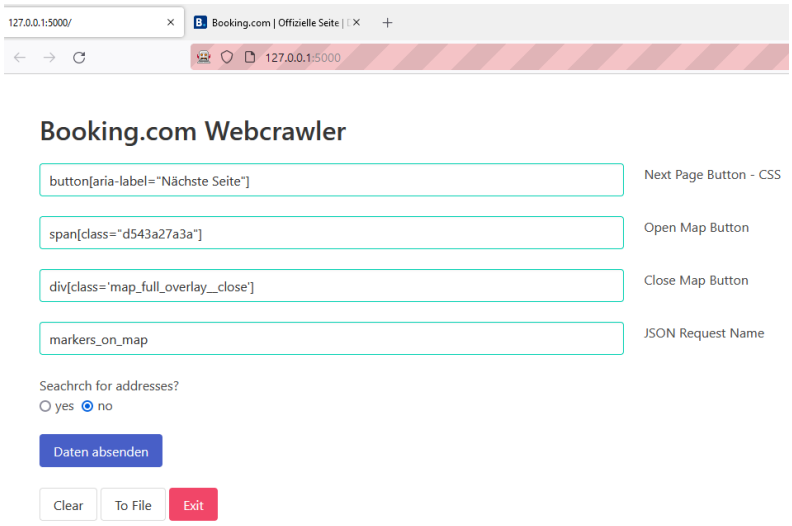

Figure 4: Bedienoberfläche (Quelle: eigene Darstellung 2022)

3 BOOKING.COM – SUCHEN

In Booking.com muss nach dem gewünschten Gebiet gesucht werden, entsprechende Filter und Suchparameter können beliebig eingestellt werden. Es ist wichtig, die Cookies zu akzeptieren, da der Webcrawler ansonsten nicht funktioniert.

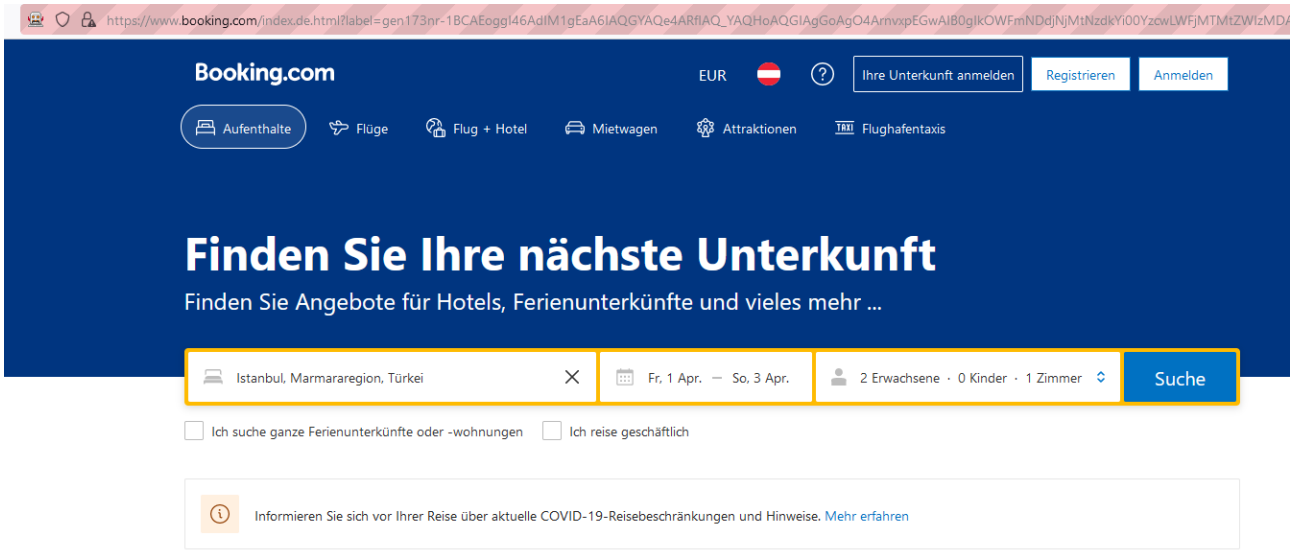

Figure 5: Titelseite Booking.com (Quelle: eigene Darstellung 2022)

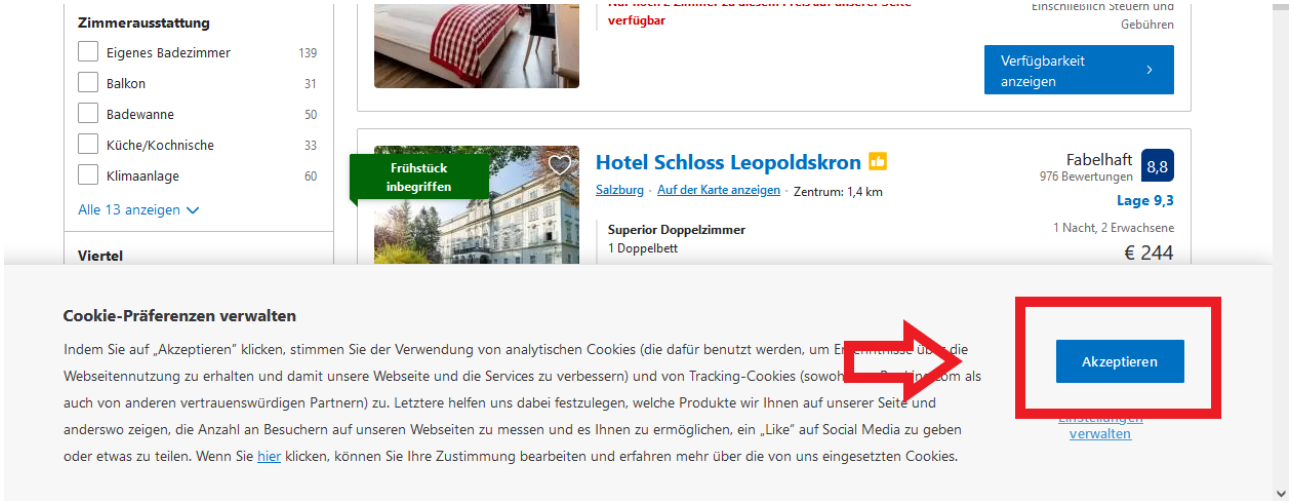

Figure 6: Akzeptieren der Cookies (Quelle: eigene Darstellung, 2022)

Die Booking.com Webseite sollte ähnlich wie in der Ansicht von Abbildung 7. sein, ohne diese Ansicht funktioniert der Webcrawler nicht.

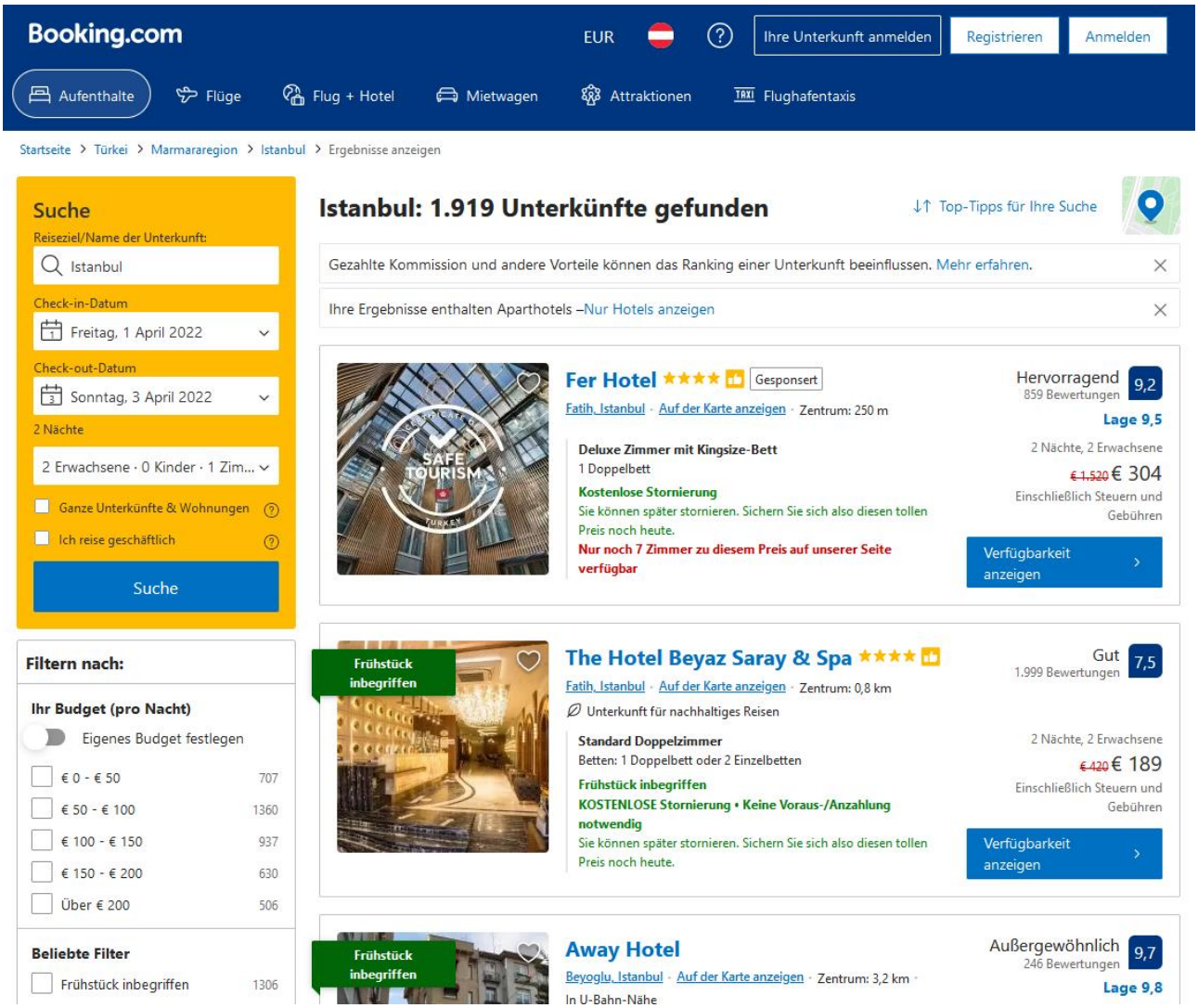

Figure 7: Hotels in der Anzeige von Booking.com, die Webseite sollte so aussehen, bevor der Webcrawler gestartet wird.

(Quelle: eigene Darstellung 2022)

4 BEDIENELEMENTE

Der Webcrawler beinhaltet 4 Textfelder und 6 Knöpfe, in den Textfeldern werden die CSS/HTML-Tags aus Booking.com angegeben, die der Webcrawler zum Navigieren, durch jede Seite benötigt. Dies beinhaltet: (1) Der Knopf für die nächste Seite von Hotels. (2) Der Knopf, der die Kartenansicht öffnet. (3) Der Knopf, der die Kartenansicht schließt. (4) Der Name des JSON-Files, in dem die Koordinaten sind.

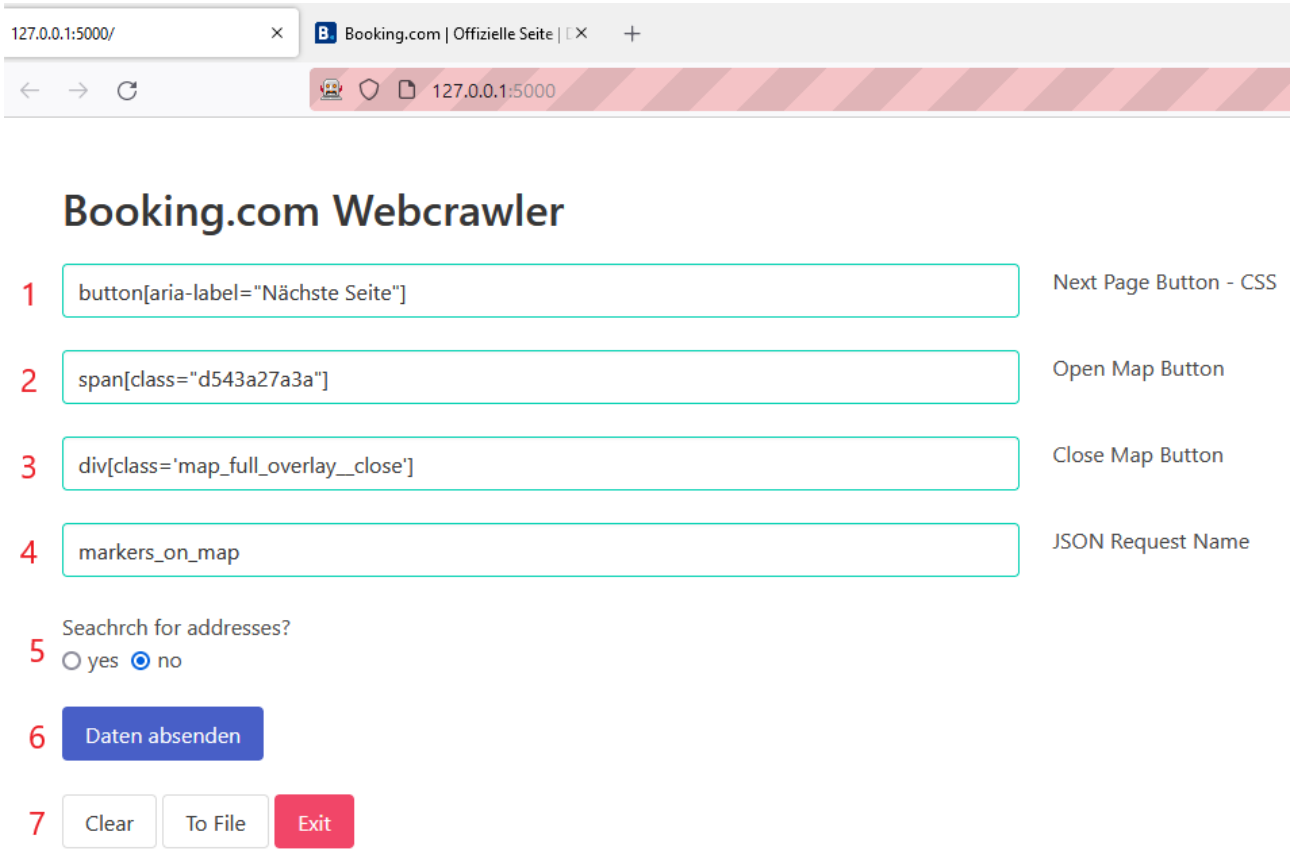

Figure 8: Bedienelemente des Webcrawler (Quelle: eigene Darstellung, 2022)

(5) **Die Suche nach Adressen ist standardmäßig deaktiviert**, da sich dadurch die Laufzeit signifikant erhöht, durch Klick auf „yes“ wird auch nach Adressen gesucht.

(6) Durch Anklicken von Daten absenden, wird der Webcrawler gestartet.

(7) „Clear“ löscht alle bisher gefundenen JSON-Dateien und kann z. B. bei einem Neustart verwendet werden. „To File“ konvertiert die gesammelten Dateien in ein GeoJSON und ein CSV-File mit „ ; “als Trennzeichen. Beide Files werden nach dem Klick als Download bereitgestellt. Durch „Exit“ wird der Browser und der Webcrawler beendet.

5 ÄNDERUNG DER VOREINSTELLUNGEN

Die Attributnamen der HTML-Elemente, die der Webcrawler zur Navigation benötigt (Textfeld 1–4), werden laufend geändert, wenn der Webcrawler ein Element nicht finden kann, muss der Identifikator in dem Textfeld geändert werden.

Die Auswahl der Objekte erfolgt über den folgenden Syntax:

Elementname[Attributname="Attributwert"]

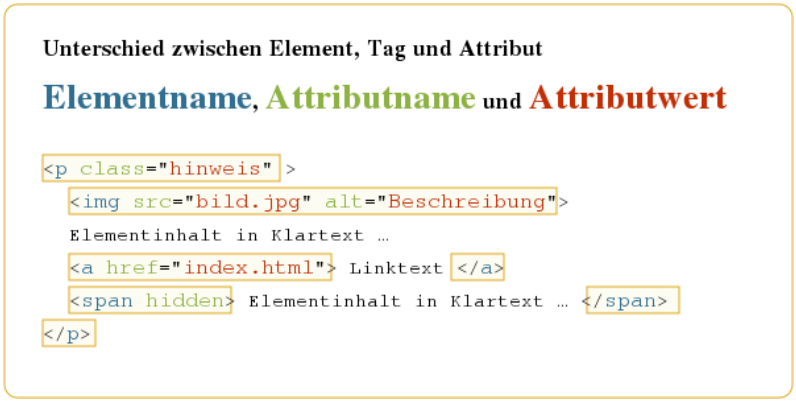

Figure 9: HTML-Elemente (Quelle: https://wiki.selfhtml.org)

Die Werte können aus mithilfe der Entwicklertools für Firefox gesucht werden. Dafür im Anwendungsmenü auf in „Weitere Werkzeuge“ das Feld „Werkzeuge für Webentwickler“ auswählen. Ansonsten kann auch die Tastenkombination „Ctrl“ `+ „Shift“ + „i“ verwendet werden. Objekte können anschließend mit dem Inspektor Knopf ausgewählt werden.

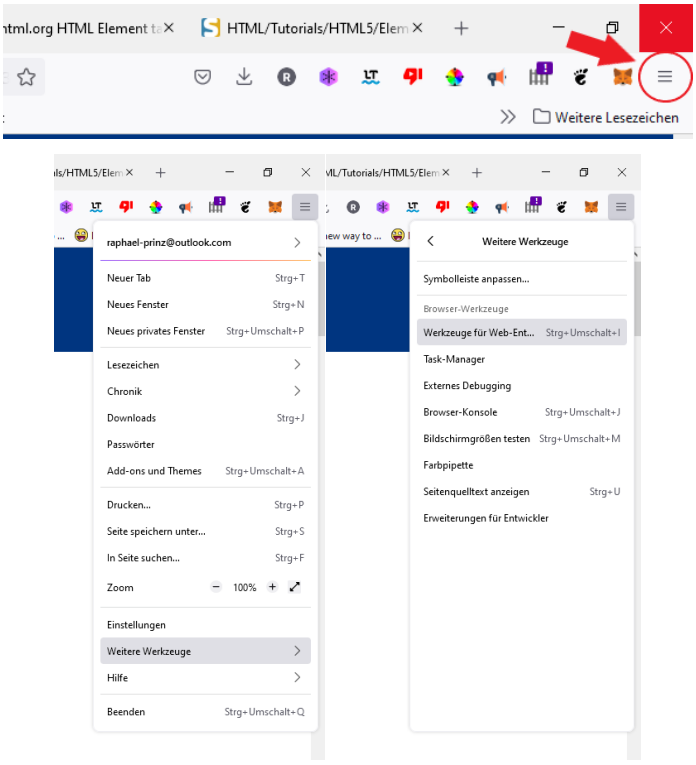

Figure 10: Öffnen der Werkzeuge für Webentwickler in Firefox (Quelle: eigene Darstellung 2022)

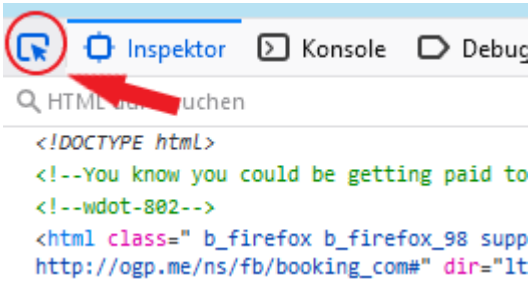

Figure 11: Inspektor zum Finden der HTML-Tags (Quelle: eigene Darstellung 2022)

5.1 Next Page Button

Im Feld “Next Page Button” muss ein Identifikator für den Knopf angegeben werden, mit dem die nächste Seite in Booking.com aufgerufen wird. Der Knopf befindet sich ganz unten in der Webseite. Als Default Parameter wird 'button[aria-label="Nächste Seite"]‘ verwendet.

Figure 12: Knopf für die nächste Seite in Booking.com (Quelle: eigene Darstellung, 2022)

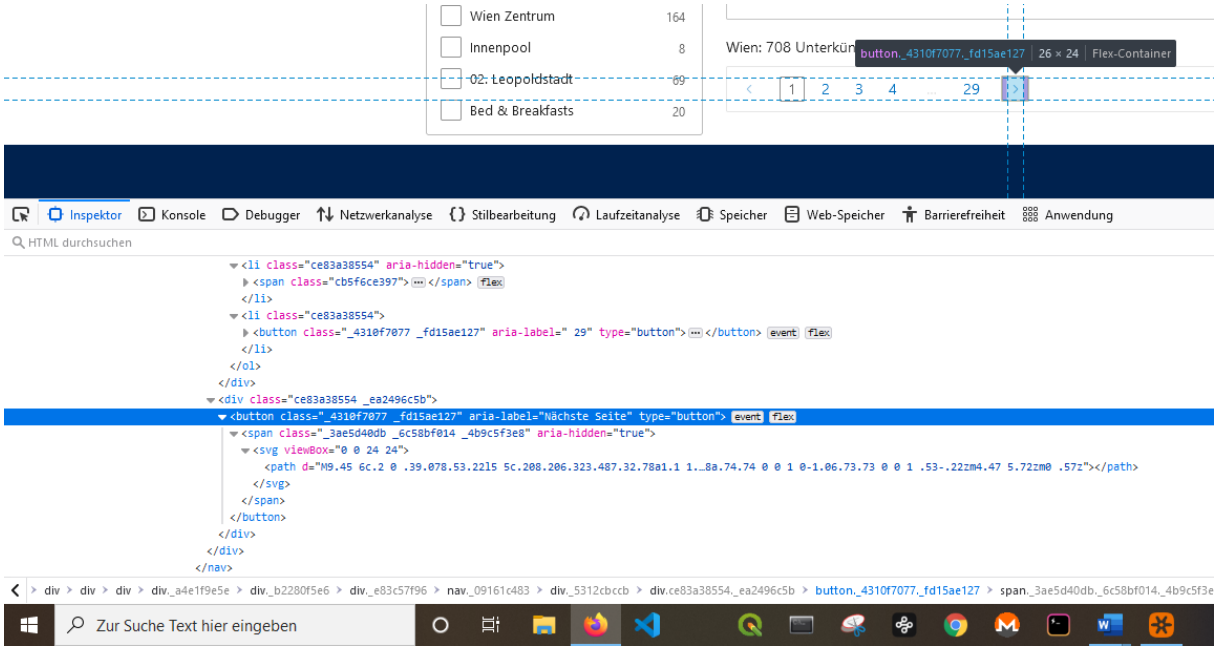

Figure 13: Ansicht des HTML-Codes für den "nächste Seite Knopf" (Quelle: eigene Darstellung 2022)

5.2 Open Map Button

Der “Open Map Button” öffnet die Kartenansicht, er befindet sich rechts oben in der Webseite. Beobachtet wurde, dass diese am häufigsten verändert wird. Anstatt dem Tag „button“ wird auch „span“ oder „div“ verwendet. Der Typ des Tags ist nicht wichtig, solange er die Karte öffnet. Als Standard wird 'span[class="d543a27a3a"]' häufig wird bei der Webseite auch der Tag div[data-testid="map-trigger"] verwendet.

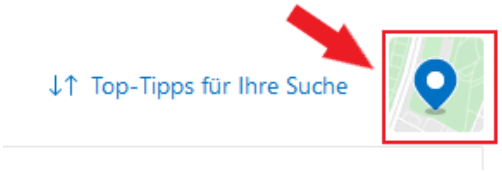

Figure 14: Knopf für das Öffnen der Karte (Quelle: eigene Darstellung, 2022)

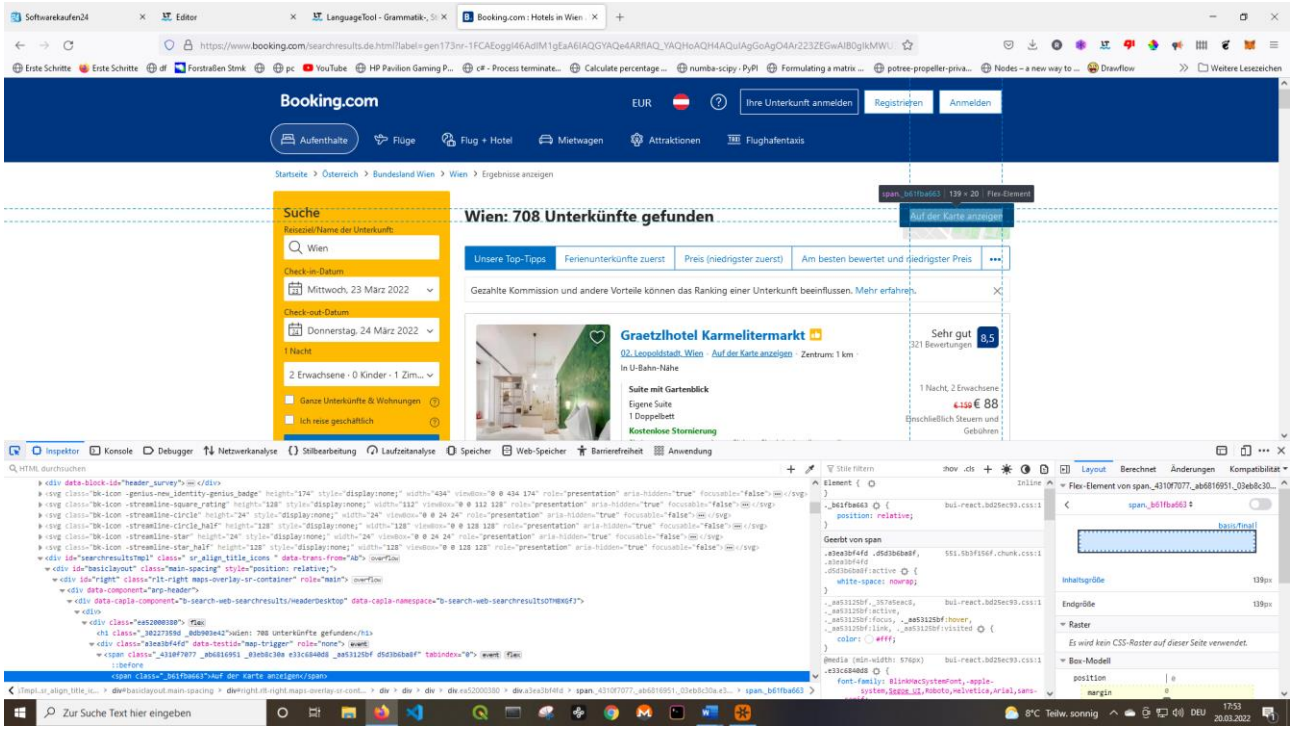

Figure 15: Inspektor für den Karten-Knopf (Quelle: eigene Darstellung 2022)

5.3 Close Map Button

Der Knopf, um die Kartenansicht nach dem Öffnen wieder zu schließen, standardmäßig wird der Identifikator ‚div[class=“map_full_overlay__close“] ‘ verwendet.

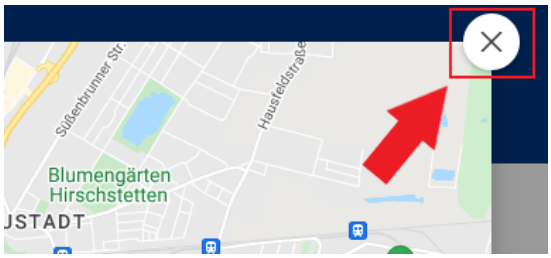

Figure 16: Knopf zum Schließen der Kartenansicht (Quelle: eigene Darstellung, 2022)

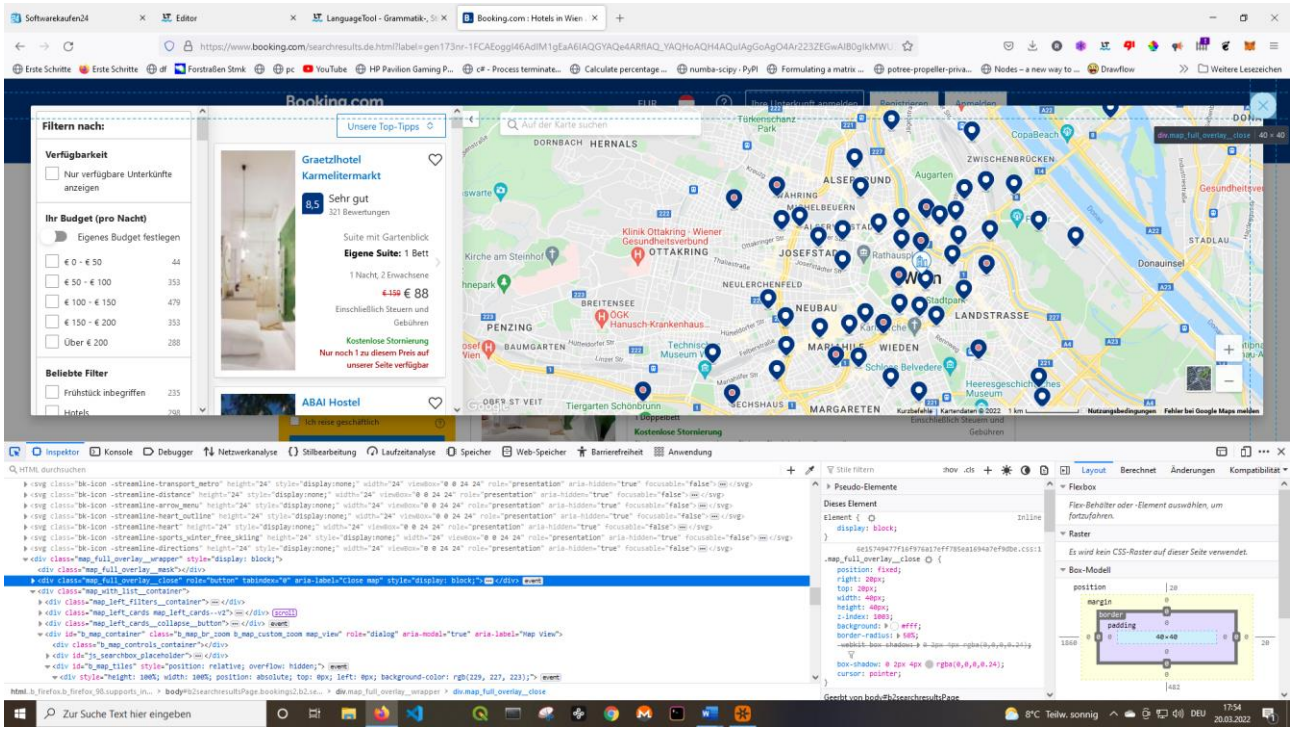

Figure 17: Knopf zum Schließen der Kartenansicht im Inspektor (Quelle: eigene Darstellung 2022)

5.4 JSON Request Name

Dies bezeichnet den Namen des JSON-Files, das von der Webseite angefordert wird. Der Name kann über die Netzwerkanalyse der Firefox – Entwicklertools gefunden werden. Dazu muss allerdings in jedem Request, der das JSON Format hat, nach den Koordinaten gesucht werden. Als Standardwert wird „markers_on_map“ verwendet. Der vollständige Name des Requests muss nicht angegeben werden, nur ein ausreichend langer Teil, um ihn eindeutig zu identifizieren.

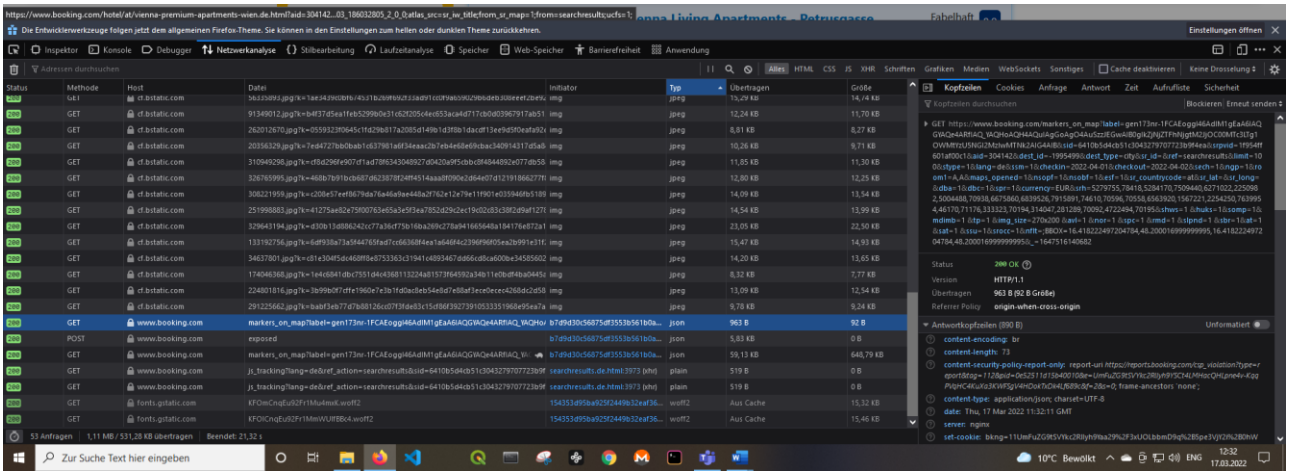

Figure 18: Suchen des JSON-Requests mithilfe der Netzwerkanalyse (Quelle: eigene Darstellung, 2022)

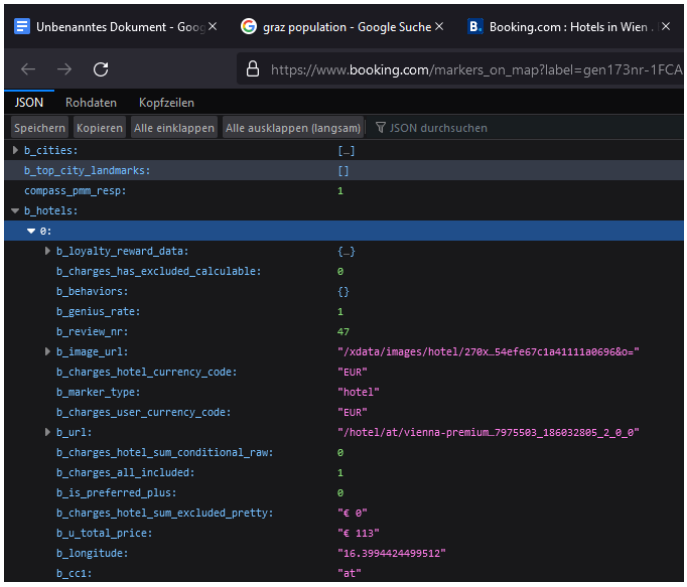

Figure 19: Ansicht der Rohdaten in Firefox (Quelle: eigene Darstellung, 2022)

6 PROGRAMM AUSGABEN

Nach Klick auf den Knopf „To File“ werden die gesammelten Daten in ein CSV und ein GeoJSON konvertiert, beide können anschließend als Zip-Datei heruntergeladen werden.

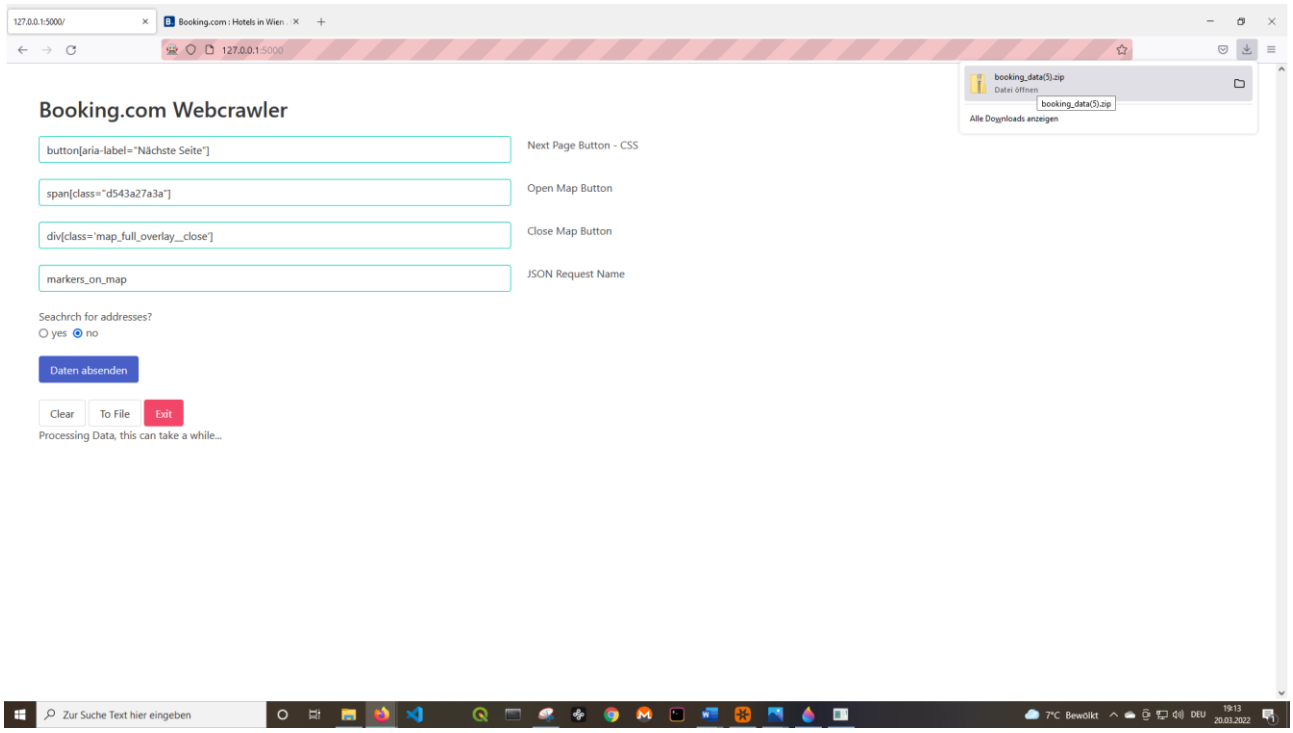

Figure 20: Herunterladen der verarbeiteten Dateien (Quelle: eigene Darstellung 2022)

7 KONTAKT

Name: Raphael Prinz

Email: raphael.prinz1993@gmail.com | raphael.prinz@edu.uni-graz.at

Telefon: 0677 6370 5851