Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Bachelorthesis

Vor- und Zuname
Timo Kock [redacted]

Titel:

„Web Scraping – Tools und Techniken zur Beschaffung von Daten aus dem Internet“

Abgabedatum:
28.02.2020

Betreuender Professor: Prof. Dr. Wolfgang Swoboda

Zweiter Prüfender: Prof. Dr. Steffen Burkhardt

Fakultät Design, Medien und Information

Department Information

Studiengang:

Medien und Information

Zusammenfassung:

Das weltweit erzeugte Datenaufkommen steigt rasant an. Daten sind ein wichtiges Unterscheidungsmerkmal für Gesellschaft, Firmen und Unternehmen. Um die enorm große Menge an Daten kontrollieren und auswerten zu können, bedarf es an “Web Scraping”. Unter Web Scraping versteht man gemeinhin das automatisierte Auslesen von Information aus Websites. Im Verlauf dieser Arbeit wird das Thema “Web Scraping" erläutert und anhand von Beispielen der Datenerhebungsprozess von unterschiedlichen “Web Scraping Tools” getestet und dokumentiert. Die erhobenen Daten werden anschließend miteinander verglichen, wodurch erkenntlich wird, welches der Tools die besten Ergebnisse erzielt hat.

I. Inhaltsverzeichnis

II. Abbildungsverzeichnis

III. Tabellenverzeichnis

1. Einleitung

Ein Unternehmer, welcher ein neues Start-up plant, ein CEO eines Fortune 500 Unternehmens, ein Aktienanalyst, ein Vermarkter sowie ein Journalist haben alle etwas gemeinsam, sie alle leiten ihre Strategien und Erkenntnisse aus Daten ab - Daten sind das neue Unterscheidungsmerkmal (vgl. Patel 2018). Sie sind der Kern der Marktforschung und aus ihnen werden Geschäftsstrategien entwickelt. Besonders mittelständische Unternehmen mit einem geringen Marketingbudget können durch die Analyse von Daten effizient herausfinden, welche Inhalte für sie relevant sind (vgl. Beckmann 2019). Unabhängig davon, ob sie ein neues Projekt starten oder eine neue Strategie für ein bestehendes Unternehmen entwickeln möchten, müssen sie stets auf eine große Datenmenge zugreifen und diese analysieren (vgl. Patel 2018).

Aus einer Studie, die der amerikanische Festplattenhersteller Seagate und das IT-Marktbeobachtungshaus IDC im November 2018 veröffentlicht haben, geht hervor, dass die Wachstumsrate der weltweit erzeugten und ausgetauschten Daten enorm ansteigt. Demnach wird das weltweite Datenaufkommen bis zum Jahr 2025 auf 175 Zettabytes anwachsen (vgl. Reinsel/ Gantz/ Rydning 2018, statista 2018). Das sind acht Mal so viel wie im Jahr 2017, denn da lag das gesamte Datenvolumen bei 23 Zettabytes. Würden all diese Daten aus den 175 Zettabytes auf herkömmliche DVD’s abgespeichert werden, dann würde der Stapel davon 23 Mal der Entfernung zwischen dem Mond und der Erde entsprechen (vgl. Kroker 2018, Statista 2018).

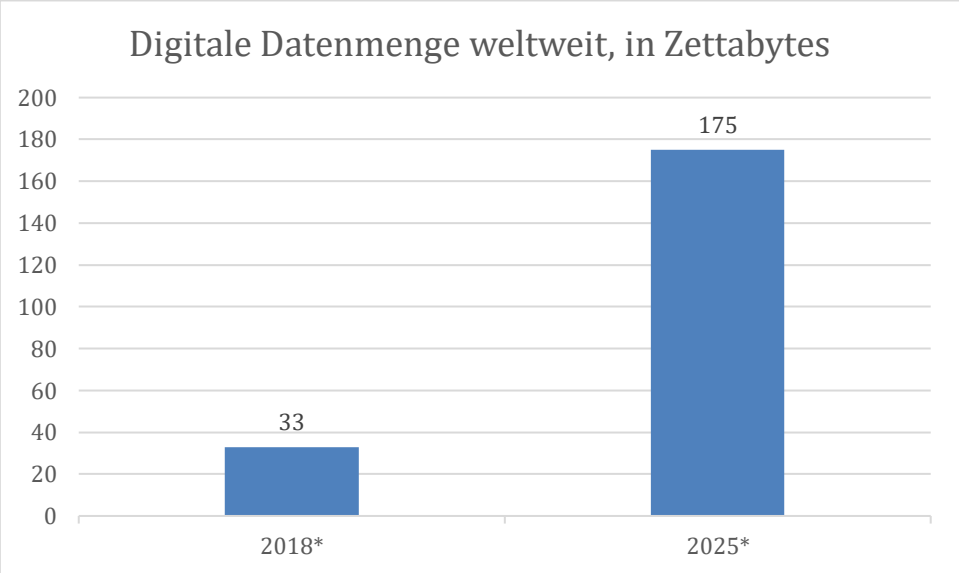

Abbildung 1: Digitale Datenmenge weltweit, in Zettabytes, Statista, 2018

Die produzierten Daten werden nicht fortlaufend nur durch Verbraucher erzeugt, vor allem das Datenvolumen in Unternehmen steigt enorm an. Eine weitere Studie von Seagate und IDC sagt aus, dass die in Unternehmen gelagerten Bytes im Jahr 2025 ca. 80 Prozent der Gesamtmenge aller erzeugten Daten ausmachen werden (vgl. Kroker 2018). Im Vergleich dazu, lagen im Jahr 2018 private Nutzer und Unternehmen mit dem Volumen der produzierten Daten noch relativ ausgeglichen. Zudem sollen bis 2025 knapp 50 Prozent der weltweit gelagerten Daten in “Public Cloud” Umgebungen zu finden sein (vgl. Reinsel/ Gantz/ Rydning 2018). Außerdem wird der Anteil von Echtzeit-Daten auf 30 Prozent des Gesamtvolumens ansteigen. Das bedeutet, dass der Mensch dann im Schnitt alle 18 Sekunden in irgendeiner Form mit Daten interagieren wird, ob beruflich oder im privaten Umfeld (vgl. Kroker 2018).

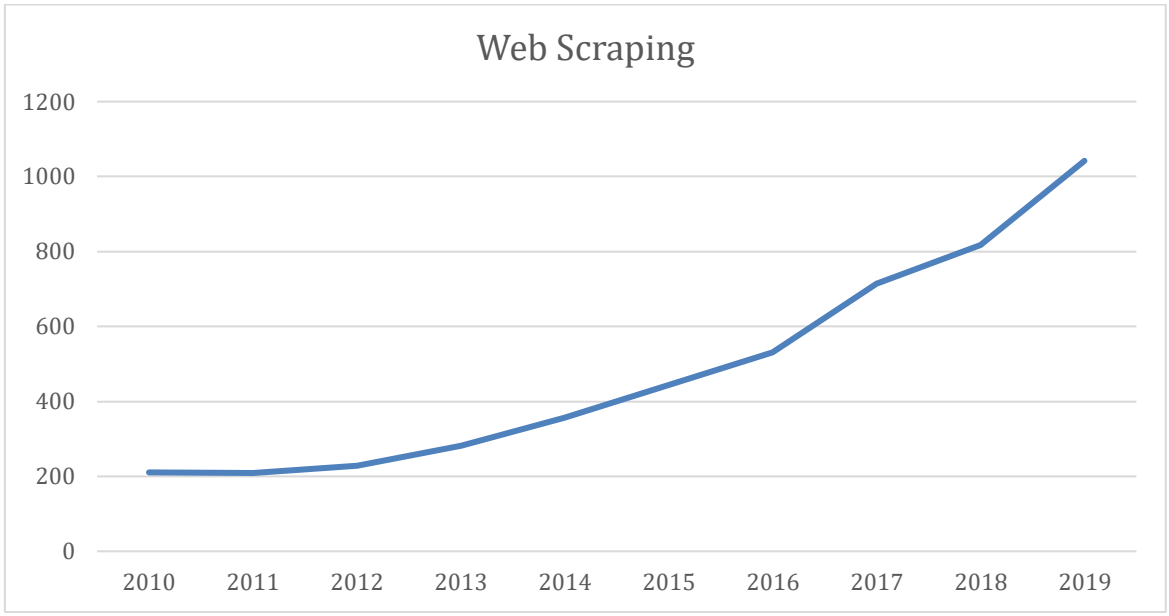

Abbildung 2: Web Scraping, trends.google.de, 2020

Laut einer Statistik auf “Google Trends”, haben sich auf der Suchmaschine Google in den letzten zehn Jahren weltweit die Suchanfragen nach dem Begriff: “Web Scraping” mehr als verfünffacht. Daraus lässt sich erschließen, dass dieses Thema immer mehr an Bedeutung gewinnen wird (trends.google.de 2020).

Unter Web Scraping versteht man gemeinhin das automatisierte Auslesen von Information aus Websites (vgl. Lohrey 2019).

Web Scraping ist auch unter einigen anderen Bezeichnungen bekannt, je nachdem, wie ein Unternehmen oder eine Firma es nennen möchte. Mal wird es “Screen Scraping” genannt, dann wieder “Web Data Extraction”, oder auch gerne “Web Harvesting” (vgl. Stuck 2019). Wie es auch genannt wird, “Web Scraping” ist eine Technik, welche dazu verwendet wird, um große Mengen an Daten aus dem Internet zu extrahieren (vgl. Stuck 2019).

Diese Daten können nicht nur aus Bildern und Texten bestehen, sondern auch aus “URL’s” (Uniform Resource Locator), dynamischen Inhalten, wie aus “CMS” (Content Management System), PHP (Personal Homepage) sowie Javascript (vgl. Lohrey 2019). Diese Daten werden von den verschiedensten Websites und aus Datenträgern extrahiert und lokal gespeichert, um sie meist sofort nutzen oder analysieren zu können (vgl. Stuck 2019). Die Daten werden anschließend in lokalen Dateisystemen oder Datenbanktabellen gespeichert, je nach der Struktur der extrahierten Daten, kann dies entschieden werden (vgl. Lohrey 2019).

Seitdem das Internet im Hinblick auf die Datenqualität und -größe gewachsen ist, haben verschiedene Interessengruppe wie Forscher, Online-Unternehmer, Programmierer und weitere Datennutzer nach Tools gesucht, um Daten von unterschiedlich großen Websites zu extrahieren (vgl. semalt.com 2018). Im Verlauf dieser Arbeit wird das Thema “Web Scraping Tools” behandelt. Es wird geklärt, worauf beim "Webscrapen” geachtet werden muss und anschließend wird anhand von Beispielen der Bearbeitungsprozess von drei unterschiedlichen Tools dokumentiert und en daraus resultierenden Ergebnissen miteinander verglichen.

2. Methodik und Vorgehensweise

Das Hauptaugenmerk dieser Arbeit ist es, Web Scraping Tools / Software zu testen, um anschließend Vergleiche ziehen zu können. Anhand von drei Anwendungsbeispielen werden mit drei unterschiedlichen Programmen bzw. Tools, Daten aus Websites extrahiert. Somit werden von allen drei ausgewählten “Web Scraping Tools”, jeweils dieselben Datensätze überprüft. Nach dem Datenerhebungsprozess werden die Daten miteinander verglichen, wodurch erkenntlich werden soll, welches Programm am effizientesten gearbeitet hat bzw. welches der Programme die besten Ergebnisse erzielt hat.

In dem ersten Anwendungsbeispiel soll eine Auflistung von Veranstaltungen jeglicher Art erstellt werden, welche in Hamburg im Jahr 2020 stattfinden. Es geht also um einen Kulturkalender, in dem aufgelistet ist, wann welche Veranstaltung an welchem Ort stattfindet und um welche Art von Veranstaltung es sich handelt. Die Arbeit mit den Tools wird bei diesem Beispiel dokumentarisch festgehalten.

In dem zweiten Anwendungsbeispiel soll aus einem Online Shop für “In-Game-Equipment" für ein Computerspiel, Daten extrahiert werden. Die gesuchten Faktoren sind die “Upload-Zeiten” und die Seitenaufrufe (Klicks) des jeweiligen Produktes, um anschließend urteilen zu können, um welche Uhrzeiten es am profitabelsten wäre, seine Produkte online zu stellen.

In dem dritten Anwendungsbeispiel werden aus einem Online Shop für Spielkonsolen und Zubehör, Daten erhoben. Es soll eine Auflistung von allen Produkten zwischen 1-10€ entstehen. Durch solch eine Liste können interessante Angebote (Schnäppchen) entdeckt werden.

Der anschließende Vergleich der Tools und Ergebnisse wird tabellarisch festgehalten. Der schriftliche Teil dieser Arbeit wird durch die erhobenen Daten in Form von Excel Tabellen als zusätzliche Leistung ergänzt, welche dem Anhang beiliegen.

3. Web Scraping Tools

Websites, auf denen Informationslisten angezeigt werden, führen dies im Allgemeinen durch, indem sie eine Datenbank abfragen und die Daten benutzerfreundlich anzeigen. Ein Web Scraping Tool kehrt diesen Prozess um, indem es unstrukturierte Websites in eine organisierte Datenbank zurückverwandelt (vgl. heliumscraper.com, Abrufdatum: 11.02.2020).

Je nach Anforderung kann eine “Web Scraping Software” mehrere Webseiten nacheinander automatisch laden und extrahiert die darin vorhandenen Daten (vgl. Stuck 2019). Es gibt Tools die speziell für eine bestimmte Website entwickelt wurden. Die meisten dieser Tools wurden jedoch dazu entwickelt, die basierend auf einer Reihe von Parametern konfiguriert werden können, um mit jeder Website zu arbeiten (vgl. Stuck 2019). Mit einem Klick auf eine Schaltfläche können die auf einer Website verfügbaren Daten auf einfache Weise in einer Datei auf dem Computer in folgenden Formaten gespeichert werden: .xls, .csv, .sql, .xml (vgl. Stuck 2019).

Abbildung 3: What is Web Scrapig? Hirinfotech.com, 2019

3.1 Anwendungsbereiche

Heutzutage kommen “Web Scraping Tools” in den verschiedensten Anwendungsbereichen zum Einsatz, wie z.B. bei Stimmungsanalysen im Social Media Bereich. Die Haltbarkeit von Social-Media-Posts ist sehr gering, aber im Ganzen betrachtet können sie jedoch wertvolle und durchaus interessante Trends aufzeigen (vgl. hirinfotech.com 2019). Bei den meisten Social Media-Plattformen wird ein API (application programming interface, deutsch: “Anwendungsprogrammier-schnittstelle”) benutzt, wodurch Drittanbieter mit ihren Tools auf diese Daten Zugriff erhalten. Diese Daten sind jedoch nicht immer ausreichend. Wenn in solchen Fällen “Web Scraping” zum Einsatz kommt, erhalten die Nutzer Zugriff auf Echtzeitinformationen, wie z.B. Tendenzen, Themen uvm (vgl. hirinfotech.com 2019, wikipedia 2002).

Um die **Reiseplanfunktion** in Google Maps zu optimieren, beschafft sich Google durch Web Scraping relevante Daten aus Fahr- und Flugplänen (de.ryte.com Abrufdatum: 11.01.2020).

Im Bereich des **E-Commerce** ist das Web Scraping ein äußerst wichtiger Faktor.(hirinfotech.com 2019) Da die meisten E-Commerce-Verkäufer ihre Produkte häufig auf mehreren Marktplätzen bzw. Plattformen anbieten, können sie mithilfe von Web Scraping die Preise auf den unterschiedlichen Plattformen überwachen und dann gezielt auf dem Markt verkaufen, welcher den höchsten Gewinn generiert. (hirinfotech.com 2019)

Außerdem müssen die **Preisstrategien der Konkurrenten** verfolgt werden. (Patel 2018) Diesen Vorgang manuell zu bewerkstelligen ist keine praktikable Option. Zusätzlich entstehen durch das ständige Wechseln der Produktpreise regelmäßig neue Informationen, wodurch es quasi unmöglich wird, die Preisentwicklungen manuell zu erfassen und zu verfolgen (vgl. Patel 2018). Durch das Web Scraping wird der Prozess der Preisermittlung der Konkurrenz automatisiert und hält den Anwender über die neuen Preisstrategien der Mitbewerber auf dem laufenden. Dieser Prozess kann endlos wiederholt werden, wodurch gewehrleistet wird, dass die Informationen auf einem aktuellen Stand sind (vgl. Patel 2018).

Hersteller mittels “Web Scraping” überprüfen, ob Einzelhändler den **Mindestpreis** ihrer Produkte einhalten oder nicht. Sie können die Daten einfach und effektiv überwachen, da

diese sehr schnell durch “Web Scraping” generiert werden, welches mit einem geringen Zeitaufwand ausgeführt werden kann (vgl. Patel 2018).

Ein Immobilieninvestor möchte für seine nächsten **Anlagemöglichkeiten** über vielversprechende Stadteile informiert werden, um mögliche Investitionen zu überprüfen. Es gibt viele Wege, um an die nötigen Daten zu gelangen, jedoch bekommt der Anwender durch das “Web Scraping” von Reisebüroseiten und Reiseportalen sowie von Wohnungsvermittlungsseiten, wertvolle Daten. Informationen, wie z.B. die am besten bewerteten Gebiete von Kunden, die von Käufern gewünschten Annehmlichkeiten sowie beliebtesten Standorte. Diese Daten können relevante Informationen für die nächste Investition darstellen (vgl. hirinfotech.com 2019).

Einige Websites, wie beispielsweise Reiseportale nutzen die **Wetterdaten** von großen Meteo-Seiten, um ihre eigene Funktionalität zu erhöhen (vgl. de.ryte.com Abrufdatum: 11.01.2020).

Um sich weiterentwickeln und verbessern zu können, benötigen **Modelle für maschinelles Lernen** gewisse Rohdaten (vgl. hirinfotech.com 2019). Durch Web Scraping Tools lassen sich innerhalb von relativ kurzer Zeit eine große Anzahl von Daten (Text, Bilder usw.) sammeln. Heutige technologische Wunderwerke wie fahrerlose Autos, Raumfahrt, Bild- und Spracherkennung werden durch maschinelles Lernen angetrieben. Diese Modelle benötigen eine große Menge an Daten, um präzise und zuverlässig arbeiten zu können (vgl. hirinfotech.com 2019).

Für das **Abrufen von Bildern und Produktbeschreibungen** bietet sich ebenfalls der Web Scraping Prozess an (vgl. Patel 2018). Es wäre äußerst mühsam die Bilder und Produktbeschreibungen von verschiedenen Herstellern manuell abrufen zu müssen. Durch das “Web Scraping” geht dies zeiteffizient und nutzerfreundlich. Der gesamte Prozess kann automatisiert werden und liefert in kürzester Zeit die gewünschten Daten (vgl. Patel 2018).

Für viele Unternehmen und Firmen ist die **Überwachung der Verbraucherstimmung** notwendig. Damit die Konsumentenstimmung analysiert werden kann, müssen Kundenfeedbacks und die Bewertungen verschiedener Unternehmen überprüft werden (vgl. Patel 2018). Auch hier wäre das manuelle Abrufen der Bewertungen auf verschiedenen Websites zu aufwändig und zeitintensiv. Mit Web Scraping können alle Bewertungen in

einer Tabelle veranschaulicht und anhand von Stichwörtern miteinander verglichen werden (vgl. Patel 2018).

Nachrichten sind **im Bereich der Finanzen und Versicherungen** die größte Quelle für Erkenntnisse (vgl. Patel 2018). Jedoch ist es nicht möglich jede Zeitung geschweige denn jeden Artikel manuell zu lesen. Durch die Anwendung des “Web Scraping” ist der Nutzer in der Lage, wertvolle Inputs aus verschiedenen Nachrichten, Überschriften usw. zu extrahieren (vgl. Patel 2018).

Im Bereich der Marktdatenaggregation wird “Web Scraping” verwendet, da sehr viele Marktdaten im Internet auf diversen Websites vorhanden sind (vgl. Patel 2018). Die einzelnen Websites und Suchergebnisse zu analysieren, ist jedoch zu zeitaufwändig. Im Hinblick auf die Aktienanalyse, dient “Web Scraping” in diesem Fall dazu, relevante Informationen von verschiedenen Websites zu sammeln (vgl. Patel 2018).

Damit Analysten ihren Kunden ein mögliches Investment in ein Unternehmen empfehlen können, brauchen sie dafür die **Jahresabschlussbilanz** des Unternehmens. Jedoch ist es schwierig, manuell die Abschlüsse von mehreren Unternehmen über mehrere Jahre hinweg zu generieren. Durch Web Scraping Tools werden Abschlüsse von verschiedenen Standorten und für verschiedene Zeiträume zur weiteren Analyse extrahiert und auf Grundlage dessen dann Investitionsentscheidungen getroffen (vgl. Patel 2018).

3.2 Hinweise zur Anwendung

Vor jedem “Web Scraping Projekt” ist es zu empfehlen, sich die Datei “Robots.txt” der Website anzuschauen, von welcher Daten extrahiert werden sollen. (Stuck 2019) Das Dokument zeigt einige Regeln auf, welche definieren, wie Bots mit der Website verkehren dürfen (vgl. yandex.com, Abrufdatum: 14.12.2019). Sollten jedoch diese vorgeschriebenen Regeln nicht beachtet werden, so befindet sich der Anwender in einer so genannten rechtlichen Grauzone. (Stuck 2019) Die “Robots.txt” Datei kann mit einer sehr einfachen Methode aufgerufen werden. Dazu muss in den Webbrowser hinter die Internetadresse “/robots.txt” hinzugefügt werden. Wenn z.B. Daten aus der Website www.haw-hamburg.de extrahiert werden sollten, müsste vorher in den Browser folgendes eingegeben werden:

www.haw-hamburg.de/robots.txt. In dieser Datei ist eine Liste mit Befehlen und Datennamen zu sehen. Alle Daten die hinter dem Befehl “disallow:” stehen, dürfen nicht extrahiert werden, zum Beispiel:

```
User-agent:

*Disallow:/typo3/
Disallow:/typo3_src/
Disallow:/tslib/
Disallow:/t3lib/
```

Bei sehr hoher Belastung werden einige Webserver zum Opfer von Ausfallzeiten (vgl. Stuck 2019). Bots fügen einem Server einer Website mehr Interaktionslast hinzu als gewohnt. Sollte diese Last einmal einen kritischen Punkt überschreiten, führt das dazu, dass der Server langsamer wird oder sogar abstürzen kann und somit die Benutzererfahrung der Website zerstören. Es ist also demnach ratsam, wenn die Daten während der sogenannten Leerlaufzeiten (nachts oder zu Zeiten mit geringem Traffic) durchsucht werden (vgl. Stuck 2019). Dadurch wird das Risiko verringert in Web-Traffic und Serverausfallzeiten verstrickt zu sein. Ebenso sollte darauf geachtet werden, dass die Richtlinien eingehalten werden und verantwortungsvoll mit den Daten umgegangen wird, wie beispielsweise die Achtung der urheberrechtlich geschützten Daten (vgl. Stuck 2019).

3.3 Rechtliche Probleme beim Web Scraping

Betreiber von Websites wollen sich meistens gegen das Auslesen von ihren Daten und gegen deren Vergleich mit anderen schützen (vgl. Solmecke 2013). Grund dafür ist, dass ihre Seiten durch Werbeeinnahmen finanziert werden, wodurch es auch im Sinne des Betreibers ist, dass seine Seite auch wirklich von einem Menschen angeklickt wird, der die gewünschten Informationen sucht (vgl. Solmecke 2013).

“Betreiber einer Datenbank können sich grundsätzlich auf: iSd § 87a Abs. 1 Satz 1 aus dem Urhebergesetz stützen. Denn Datenbanken sind nach den §§ 87b Abs. 1 Satz 2 UrhG gegen die öffentliche Wiedergabe eines wesentlichen Teils geschützt. Doch wann extrahiert und veröffentlicht die Vergleichsseite einen Teil?” (Solmecke 2013)

Gerichte haben sich schon mehrfach damit beschäftigt, ob sich ein Betreiber mit zusätzlichen Mitteln, wie beispielsweise in den AGB’s, dagegen wehren oder schützen kann,

dass von seiner eigenen Webseite Daten entnommen werden.(Solmecke 2013). Ob solch ein Verhalten oder Prozess möglicherweise wettbewerbswidrig ist (iSd § 4 Nr. 10 UWG), da den Betreibern durch fehlende Besucher Werbeeinahmen entgehen (vgl. Solmecke 2013)

“Bis 2009 waren die Klagen gegen dieses automatische Extrahieren von Daten noch meist erfolgreich. Jedoch hat sich das in jüngerer Vergangenheit geändert. Mittlerweile wird häufig in die Richtung entschieden, dass wenn weder wesentliche Teile der Datenbank kopiert werden noch es zur technischen Überlastung der “gescrapten” Seite kommt, automatisiertes Sammeln von Daten zulässig ist, solange die Seite rechtlich und technisch frei zugänglich ist.” (Solmecke 2013)

4.webscraper.io

Das “Web Scraping Tool” “webscraper.io” ist mit mehr als 250.000 Benutzern ein häufig genutztes Tool. (webscraper.io, Abrufdatum: 22.01.2020) Ziel des Anbieters ist es, Datenextraktion aus dem Web für alle einfach und zugänglich zu machen. (webscraper.io) Die Extraktion von Webdaten muss so einfach wie nur möglich konfigurierbar sein. Webscraper.io ist eine Browsererweiterung von Google Chrome (Add-On). Dieses “Tool” lässt sich einfach konfigurieren, indem mit der Maus auf die Elemente (Selektoren) gezeigt und anschließend angeklickt wird. Eine Codierung ist nicht erforderlich. Das Extrahieren von Daten aus dynamischen Websites und Websites mit mehreren Navigationsebenen, stellt kein Problem dar. Außerdem kann der Nutzer auf allen Ebenen einer Website navigieren (Kategorien und Unterkategorien, Seitennummerierung, Produktseiten). Heutzutage basieren Websites häufig auf “JavaScript-Frameworks”, welche die Verwendung von Oberflächen vereinfachen, für “Web Scraping Tools” sind diese meist jedoch weniger einfach zugänglich. Webscraper.io löst dieses Problem durch:

- die vollständige JavaScript Ausführung
- das Warten auf Ajax-Anfragen (asynchrone Datenübertragung zwischen Browser und Server) (Wikipedia.de 2020)
- das Erkennen von Seitennummerierungen
- das Runterscrollen der Seite

Mit diesem Tool können Sitemaps aus verschiedenen Arten von Selektoren erstellt werden, wodurch es möglich ist, die Datenextraktion an verschiedene Standortstrukturen anzupassen. (Webscraper.io 2020) Die gewonnenen Daten lassen sich leicht exportieren und in folgenden Formaten abspeichern:

- CSV
- XLSX
- JSON

Durch zusätzliche Dienste wie “Web Scraper Cloud” kann auf die Daten über “API’s” oder “Webhooks (ein nicht standardisiertes Verfahren zur Kommunikation von Servern)” (Wikipedia 2011) zugegriffen werden oder sie können über “Dropbox” exportiert werden.

(webscraper.io 2020) Das Tool ist über Google Chrome sowie Mozilla Firefox als Add-On verfügbar. (vgl. Chrome 2019, Firefox 2018)

Anwendungsdokumentation von webscraper.io

Schritt 1: Im Webbrowser die Tastenkombination: alt + cmd (oder ctrl) + I (großes i) drücken, um die Konsole zu öffnen. Anschließend das “webscraper.io” Add-on öffnen.

Schritt 2: Um eine neue Datei anzulegen auf “create new sitemap” klicken und dann “create sitemap” auswählen.

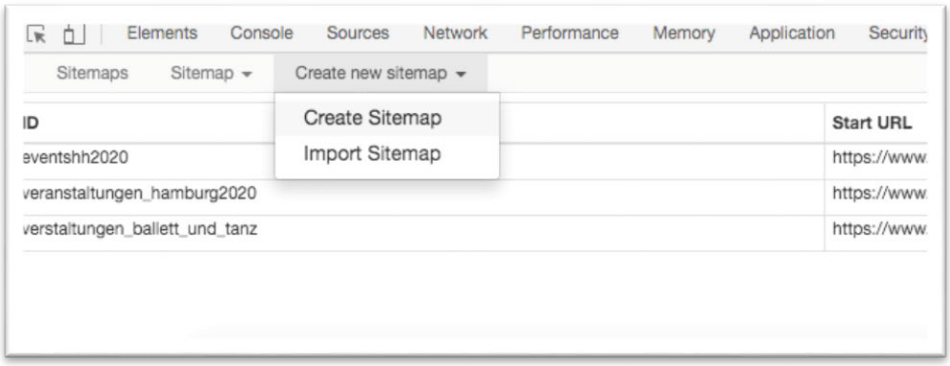

Abbildung 4: Erstellung einer Sitemap im Webscraper, webscraper.io

Schritt 3: In der Spalte “Sitemap name:” den Namen des Projektes eingeben, in diesem Fall: “Kulturkalender_hh_2020”. Als nächstes in die Spalte “Start URL:” die Webadresse angeben, von der das Tool die Suche beginnen soll, in diesem Fall “www.hamburg.de/kultur/”. Anschließend, mit dem “create sitemap” Button bestätigen.

Schritt 4: Als nächstes “add new selector” anklicken. Nach dem Öffnen des neuen Menüs in Spalte “Id:” Name des 1. selectors eingeben. In diesem Fall “Art”, weil es um die Veranstaltungsart geht, welche gesucht sind. Als nächstes in die Spalte “Type:” “Link” auswählen, da es sich in diesem Fall um Links handelt, auf die navigiert werden sollen. Unter der Spalte “Selector” auf “Select” klicken und darunter die “multiple” Funktion aktivieren, um anschließend im Browser die Links zu markieren. Da die “multiple” Funktion aktiviert ist, werden durch Klicken des 2. Links alle weiteren ähnlich relevanten, sich auf der Seite befindenden Links automatisch markiert. Um die Selektion abzuschließen, muss auf

den “done selecting!” Button geklickt werden und dieses anschließend mit “save selector” bestätigen.

Schritt 5: Im Browser auf einen der markierten Links klicken, um auf die nächste Seite zu gelangen. Als nächstes unten in der Konsole auf den vorher erstellten Selektor (in diesem Fall “Art”) klicken um diesem noch weitere Suchbefehle hinzuzufügen.

Schritt 6: Auf “add new selector” klicken, um als nächstes wieder die Spalten: “Id” und “Type” zu füllen. Id = Art_Daten (weil es sich um zusätzliche Veranstaltungstypen geht) und Type = Text (weil die auszuwählenden Daten in diesem Fall reine Textinformationen sind). Anschließend “multiple” aktivieren und unter Selector: den “Select” Button anklicken und wieder die ersten beiden relevanten Daten markieren, damit die restlichen wieder automatisch erfasst werden. Abschließend wieder mit “done selecting!” die Selektion beenden und mit dem “save selector” Button bestätigen.

Schritt 7: Auf “add new selector” klicken, um als nächstes wieder die Spalten: “Id” und “Type” zu füllen. Id = Name (weil es sich um die Veranstaltungsnamen handelt) und Type = Text (weil die auszuwählenden Daten in diesem Fall reine Textinformationen sind). Anschließend “multiple” aktivieren und unter Selector: den “Select” Button anklicken und wieder die ersten beiden relevanten Daten markieren damit die restlichen automatisch erfasst werden. Zuletzt wieder mit “done selecting!” die Selektion abschließen und mit dem “save selector” Button bestätigen.

Schritt 8: Auf “add new selector” klicken, um als nächstes wieder die Spalten: “Id” und “Type” zu füllen. Id = Datum (weil es sich um den Zeitpunkt der jeweiligen Veranstaltungen handelt) und Type = Text (weil die auszuwählenden Daten in diesem Fall reine Textinformationen sind). Anschließend “multiple” aktivieren und unter Selector: den “Select” Button anklicken und wieder die ersten beiden relevanten Daten markieren, damit die restlichen automatisch erfasst werden. Abschließend wieder mit “done selecting!” die Selektion beenden und mit dem “save selector” Button bestätigen.

Schritt 9: Auf “add new selector” klicken, um als nächstes wieder die Spalten: “Id” und “Type” zu füllen. Id = Ort (weil es um die Veranstaltungsorte geht) und Type = Text (weil die auszuwählenden Daten in diesem Fall reine Textinformationen sind). Anschließend “multiple” aktivieren und unter Selector: den “Select” Button anklicken und wieder die ersten

beiden relevanten Daten markieren, damit wieder die restlichen automatisch erfasst werden. Zuletzt wieder mit “done selecting!” die Selektion abschließen und mit dem “save selector” Button bestätigen.

Anmerkung: **Um zu überprüfen ob die ausgewählten Selektoren richtig angeordnet wurden: In der Konsole auf “Sitemap Name des Projekts” klicken, und anschließend “selector graph” auswählen.**

(Konsole -> Sitemap kulturkalender_hh_2020 -> selector graph)

Zu sehen ist der Aufbau der Verknüpfung der einzelnen Selektoren, beziehungsweise ihre Beziehung zueinander.

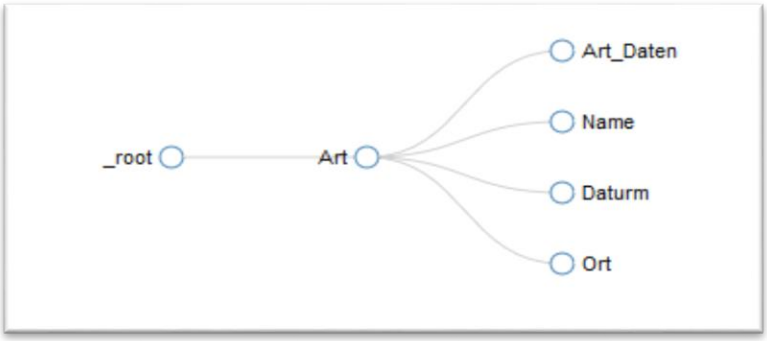

Abbildung 5: Die Verknüpfung der Selektoren, Webscraper.io, 2020

Schritt 10: In der Konsole auf: “Sitemap Name des Projekts” (Hier: “Sitemap kulturkalender_hh_2020) klicken und anschließend auf “Scrape” klicken. Die Werte in den Spalten: [Request Interval (ms)] = 2000 und bei [Page load delay (ms)] = 2000 können unverändert bleiben, oder nach Bedarf angepasst werden. Um die Seite nicht zu belasten, bleiben die Werte in diesem Fall unverändert, denn bei einem niedrigeren Wert, würde das Programm schneller Arbeiten und somit mehr “Traffic” generieren. Mit dem “Start scraping” Button, den “Web Scraping” Prozess beginnen.

Schritt 11: Datenerhebungsprozess (Die Website wird nach den erstellten Selektoren vom “Web Scraping Tool” durchsucht). Das kann einige Zeit dauern. In diesem Fall: 24 Minuten.

Schritt 12: In der Konsole auf: “Sitemap Name des Projekts” gehen und anschließend auf “Export data as CSV” auswählen.

5. Helium Scraper

“Helium Scraper” arbeitet mit einem intuitiven “Point-and-Click-Interface”. (vgl. heliumscraper.com/eng/ Abrufdatum: 20.02.2020). Für eine schnelle Extraktion werden Delegierungsaufgaben automatisch an seperate Browser übertragen. Für eine noch schnellere Extraktion können Bilder oder unerwünschte Webanfragen blockiert werden. Durch die “SQLite-Datenbank” können bis 140 Terabyte an Daten gespeichert werden. (heliumscraper.com/eng/ Abrufdatum: 20.02.2020) Das Programm bietet ein schnelles Verknüpfen und Filtern von Tabellen zum Exportieren oder zum Eingeben von Daten. “Helium Scraper” verfügt über eine JavaScript-Unterstützung (vgl. heliumscraper.com/eng/ Abrufdatum: 22.02.2020). Durch die Funktion der Proxy-Rotation kann es ähnliche Elemente aus einer oder zwei Proben erkennen. Automatisch können Listen und Tabellenzeilen auf Websites ermittelt werden. Das Programm kann über die Befehlszeile oder den Windows Task Scheduler gestartet werden (vgl. heliumscraper.com/eng/ Abrufdatum: 22.02.2020). Die gewonnenen Daten lassen sich leicht exportieren und in folgenden Formaten abspeichern:

- CSV
- XML
- JSON

Anwendungsdokumentation von helium scraper

Schritt 1: Projektname beschließen. In diesem Fall: “Kulturkalender_Hamburg_2020”.

Schritt 2: Im Programm auf der rechten Seite unter “Project Explorer” auf “Globals” klicken, um die Rubrik “Main” aufzuklappen. Dann Doppelklick auf “Main”, um die Befehlskonsole zu öffnen.

Schritt 3: Start “URL” in dem im Helium Scraper eingebetteten Browser eingeben welche dann als Hauptquelle dienen soll, damit das Programm weiß, wo die Suche beginnen soll.

Schritt 4: Auf das “Zauberstab” Symbol auf der linken Seite klicken und im Untermenü auf “load URL” klicken, um die Hauptquelle in die Konsole einzubetten.

Schritt 5: Überprüfen ob die angezeigte “URL” die richtige ist. Falls das der Fall ist, anschließend auf “done” klicken, um zum nächsten Schritt zu gelangen.

Schritt 6: Die Hauptquelle erscheint in der Konsole, wenn alle Schritte bisher erfolgreich ausgeführt worden sind.

Schritt 7: Rechtsklick auf “Browser load” und anschließend “add sibling” auswählen. Dadurch erscheint eine neue Spalte “[..]” unter der Hauptquelle.

Schritt 8: Unter dem eingebetteten Browser befindet sich eine Funktion namens “selection mode”. Diese anklicken, um in den Selektionsmodus zu gelangen.

Schritt 9: Da alle Links durchsucht werden sollen wird als erstes ein Link ausgewählt und anschließend automatisch durch Klicken der Funktion “select similar elements”, werden alle weiteren Punkte ausgewählt. Falls jedoch nur einzelne Punkte bzw. Infotmationen erwünscht sind ist es auch möglich mit der “Strg” bzw. “Cmd” Taste auf der Tastatur gezielt zu markieren.

Schritt 10: Rechtsklick mit der Maus auf den leeren “sibling” links in der Konsole “[..]” und “create selector from sample” auswählen. Somit erscheint in der Konsole “select.” und anschließend hinter den Punkt den Namen des Selektors manuell eingeben. In diesem Beispiel passt “Genre” oder “Rubrik” da es um die verschiedenen Arten von Veranstaltungen geht. Also in diesem Fall “Select.Genre” und mit der Taste “Tab” dem Befehl bestätigen.

Schritt 11: Rechtsklick auf “Select.Genre” und dann auf “Add sibiling” klicken, um erneut einen leeren “sibling [..]” zu erschaffen.

Schritt 12: Mit einem Klick auf den leeren “sibling [..]”, öffnet sich das Untermenü mit einer Übersicht von allen vorhandenen Befehlen. Danach wird “Browser.Navigate” ausgewählt, um somit dem Programm zu befehlen, dass es alle vorher ausgewählten Punkte ausführen soll.

Schritt 13: Jetzt muss zu einem der vorher ausgewählten Links navigiert werden, welcher im fortlaufenden Prozess durchsucht werden soll. Anschließend Rechtsklick auf “browser navigate” und wieder “add sibiling” auswählen.

Schritt 14: Auf den leeren “sibling [..]” klicken und “extract” eingeben und wieder mit der “Tab” Taste bestätigen.

Schritt 15: In der Konsole im “extract” Befehl den Datensatz benennen welcher erhoben werden sollen.

Schritt 16: Das “Selection Mode” Symbol unter dem Browser auswählen, um in den Selektionsmodus zu gelangen. Anschließend müssen die Daten, welche erhoben werden sollen, per Mausklick markiert werden. Danach werden wieder durch das “select similar elements” Symbol, alle ähnlichen beziehungsweise gleichen Datensätze zusätzlich ausgewählt.

Schritt 17: In der Konsole Rechtsklick auf “value”. Anschließend im Untermenü “create Selector from Sample” auswählen. Dadurch erscheint in der Konsole “select.” und dann hinter dem Punkt den Namen der nächsten gewünschten Daten eingeben wie in diesem Fall: “Select.GenreX”, da vorher schon einmal “Select.Genre” benutzt wurde. Zuletzt den Befehl wieder mit der Taste “Tab” bestätigen.

Schritt 18: In der Konsole Rechtsklick auf “genre”. Anschließend wird “add sibling” ausgewählt und dann muss der neue “sibling [..]” benannt werden. In diesem Fall wurde “Name” ausgewählt, weil es um den Namen der zu suchenden Veranstaltung geht.

Schritt 19: Auf der Website einen Namen auswählen und wieder das Symbol “select similar elemets” auswählen und mit Rechtsklick auf “value” in der Konsole gehen und dann “create selector from sample” auswählen und mit “Tab” bestätigen.

Schritt 20: In der Konsole mit Rechtsklick auf “Name” gehen und dann wieder “add sibling” auswählen und anschließend benennen, in diesem Beispiel: “Ort”, weil es um die Veranstaltungsorte geht, die gesucht sind.

Schritt 21: Auf der Website einen Ort auswählen und wieder das untere Symbol “select similar elemets” mit der Maus betätigen, dann wieder Rechtsklick auf “value” und im Anschluss “create selector from sample” auswählen und mit “Tab” bestätigen.

Schritt 22: In der Konsole Rechtsklick auf “Ort” und dann “add sibling” auswählen und benennen. In diesem Beispiel: “Datum”, weil es um die Zeitpunkte, der Veranstaltungen geht.

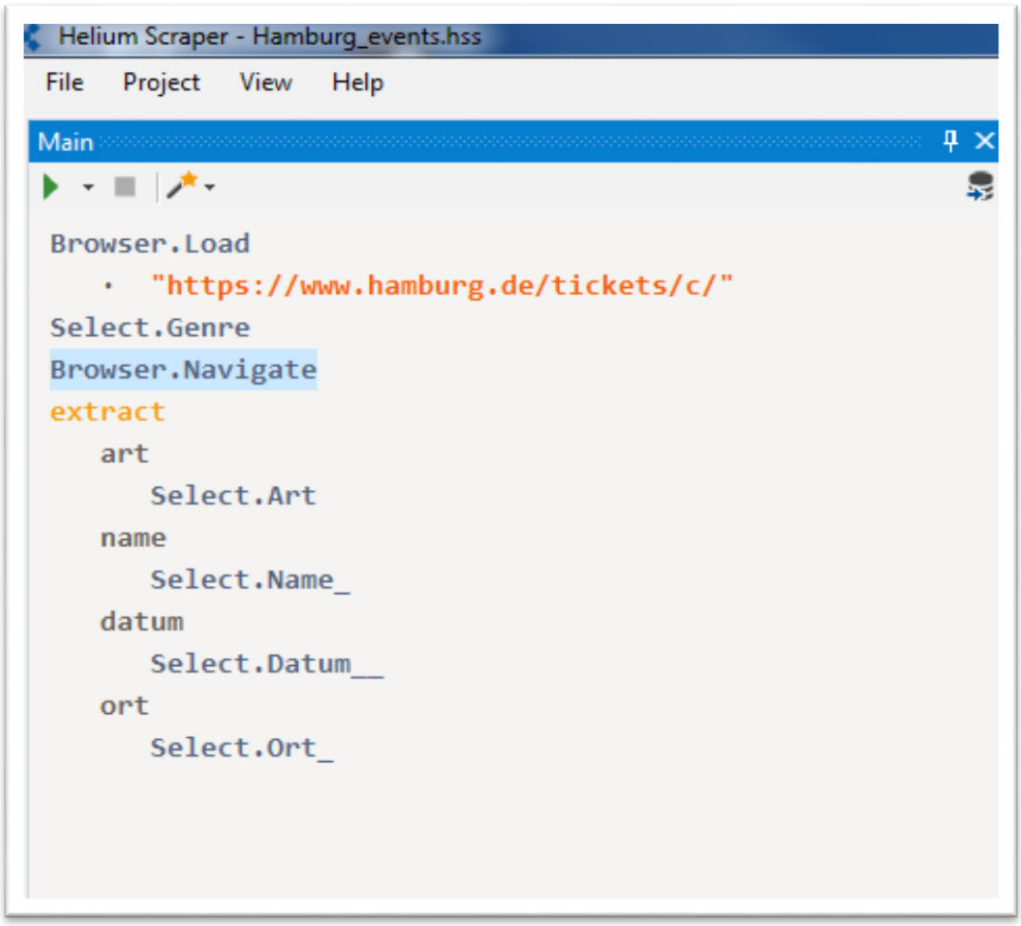

Abbildung 6: Bestimmung von Selektoren in der Konsole vom Helium Scraper

Schritt 23: Auf der Website ein Datum auswählen und wieder das Symbol unten namens “select similar elemets” anklicken und Rechtsklick auf “value” und “create selector from sample” auswählen und mit “Tab” bestätigen.

Schritt 24: Links über der Konsole auf das “Play Symbol” namens (run) klicken, um die aufgeführte Befehlsliste auszuführen. Nun werden die Daten erhoben, dies kann einige Zeit dauern. In diesem Fall 35 Minuten.

Schritt 25: Nach dem Datenerhebungsprozess auf der rechten Seite in der “Project Explorer” Maske, “Data Main” aufklappen und anschließend ein Doppelklick auf “Main” um auf die gewonnenen Daten zugreifen zu können.

Schritt 26: Man erhält eine Liste mit den extrahierten Informationen. Durch das Klicken, auf das "export” Symbol, können die Daten auf dem Computer abgespeichert werden. Es gibt folgende Formate zur Auswahl: “.xlsx”,”.xls” oder “.csv”.

6. Scraper Parsers

Mit der Browser Erweiterung für Google Chrome Namens “Scraper Parsers” können einfach unstrukturierte Daten extrahiert und ohne Code visualisiert werden. (parsers.me Abrufdatum: 12.02.2020) Bei der kostenlosen Version mindestens bis zu 1000 Seiten extrahiert werden. Es gibt keine Begrenzung pro Monat. (parsers.me Abrufdatum: 13.02.2020) Es kann unendlich genutzt werden, wobei bei anderen Tools teilweise nur Testversionen für einen begrenzten Zeitraum zur Verfügung stehen, wie z.B. “Helium Scraper” nur für 10 Tage. (Lohrey Abrufdatum: 23.02.2020). Parsers verwendet maschinelles Lernen. (chrome.google.com Abrufdatum: 23.02.2020) Dadurch analysiert es die Website sobald die Selektoren bestimmt wurden und findet in den meisten Fällen basierend auf der eingegebenen Vorlage, ähnliche bzw. identische relevante Daten. Je nachdem welche Parsers Version genutzt wird, bestimmt sich auch die Geschwindigkeit der Datenextraktion. Die kostenlose Version kann 10 Anfragen gleichzeitig an die Website stellen, wobei die teuerste Version (199 US $ pro Monat) 50 Anfragen gleichzeitig stellen kann, wodurch es dann 5 x schneller arbeitet als die kostenlose Version. (parsers.me/pars/pay, Abrufdatum: 16.02.2020). Kunden bewerten es mit positiv, dass rund um die Uhr ein Chat und Mail Support zur Verfügung steht, der bei Fragen oder Problemen hilft. (chrome.google.com Abrufdatum: 23.02.2020). Außerdem bietet das Tool die Möglichkeit, die Suche zu gewünschten Zeitpunkten automatisch erneut auszuführen. (parsers.me Abrufdatum: 13.02.2020) Ein weiteres interessantes Feature an dem Tool ist, dass einem der Verlauf der extrahierten Daten aller Versionen nach Datum angezeigt werden kann. (parsers.me Abrufdatum: 13.02.2020) Bei der Standard Version welche 49 US $ pro Monat kostet, gibt es sogar einen technischen Support, der dem Kunden hilft, bzw. ihm sogar die Aufgabe abnimmt, falls er kein Erfolg hatte. Der Service dauert 1-2 Tage. (parsers.me Abrufdatum: 13.02.2020)

6.1 Anwendungsdokumentation von Parsers

Schritt 1: Bei dem Tool Parsers muss nicht wie bei den vorherigen Programmen eine Quell URL angegeben werden. Ebenso braucht es keinen Pfad, welcher Programmiert werden muss an dem sich das Tool halten soll. Es muss lediglich eine Seite geöffnet werden, auf

der alle zu suchenden Selektoren auf einmal vorhanden sind. Wenn das der Fall ist, können folgende Schritte eingeleitet werden:

Schritt 2: Im Google Chrome Browser oben rechts das “Parsers” Add-on Symbol anklicken zum Öffnen der Konsole.

Schritt 3: Auf “Label 1” klicken und die leere Spalte benennen. In diesem Fall: “Name”, da es um die Veranstaltungsnamen geht (siehe Abbildung 7).

Schritt 4: Auf die Eingabefeld: “Highlight the field on the site” klicken und anschließend auf die Website navigieren, um den Selektor zu markieren. In diesem Fall wird der Name von der angezeigten Veranstaltung angeklickt, wodurch diese Information automatisch in der Spalte erscheint.

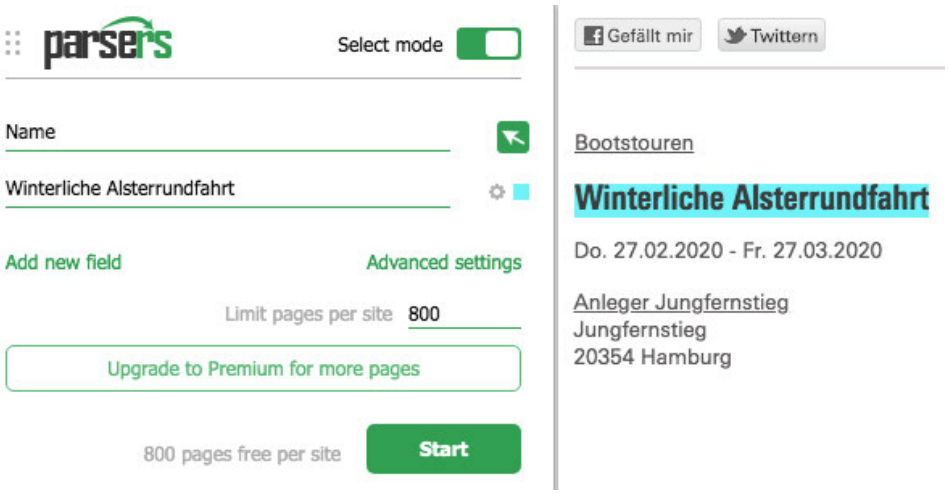

Abbildung 7: Bestimmung eines Selektors in der Konsole vom Parsers Scraper

Schritt 5: Auf “add new field” klicken um das “Label 2” zu eröffnen.

Schritt 6: Auf “Label 2” klicken und die leere Spalte benennen. In diesem Fall: “Art”, da es jetzt um die Art Veranstaltung geht.

Schritt 7: Auf das Eingabefeld: “Highlight the field on the site” klicken und wieder auf die Website navigieren, um den Selektor zu markieren. In diesem Fall wird "Bootstouren” angeklickt, weil eine Veranstaltungsart gesucht ist. Diese Information erscheint wieder automatisch in derselben Spalte.

Schritt 8: Auf “add new field” klicken um das “Label 3” zu eröffnen.

Schritt9: Auf “Label 3” klicken und benennen. In diesem Fall: “Datum”, weil jetzt das Datum der jeweiligen Veranstaltungen gesucht ist.

Schritt 10: Auf das Feld: “Highlight the field on the site” klicken und wieder auf die Website navigieren, um den Selektor zu markieren. In diesem Fall wird "Do. 27.02 - Fr. 27.02. 2020” angeklickt, weil das Datum der Veranstaltung gesucht ist. Diese Information erscheint anschließend wieder automatisch in derselben Spalte.

Schritt 11: Auf “add new field” klicken um das “Label 4” zu eröffnen.

Schritt12: Auf “Label 4” klicken und benennen. In diesem Fall: “Ort”, weil jetzt der jeweilige Veranstaltungsort gesucht ist.

Schritt 13: Auf das Feld: “Highlight the field on the site” klicken und wieder auf die Website navigieren, um den Selektor zu markieren. In diesem Fall wird "Anleger Jungfernstieg” ausgewählt, weil es noch um die Veranstaltungsorte geht (siehe Abbildung 8).

Schritt 14: Nachdem alle Selektoren bestimmt wurden, kann das Web Scraping beginnen.

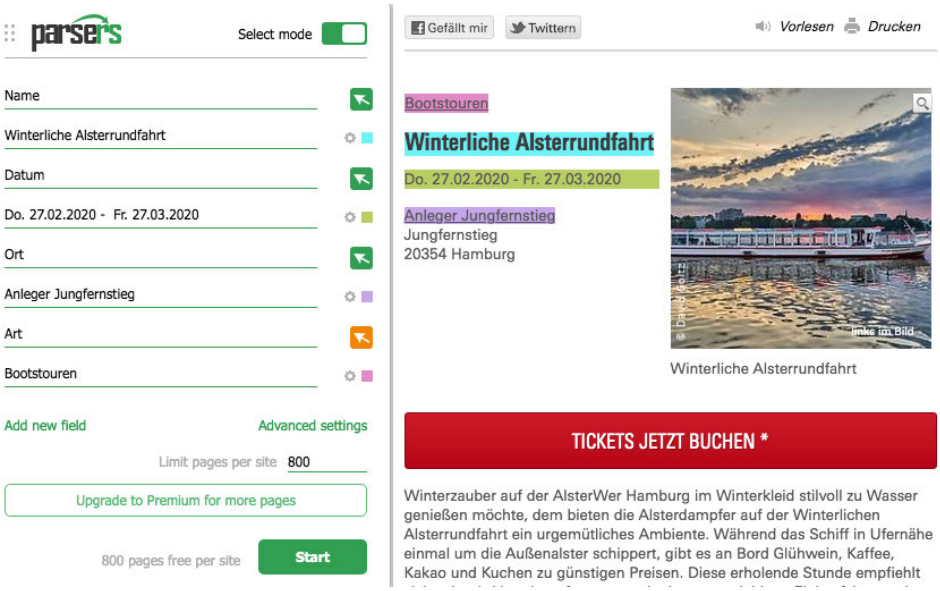

Abbildung 8: Aufbau der Konsole vom Scraper Parsers

Der Datenerhebungsprozess kann mit dem Bestätigen des “Start” Buttons, beginnen. Der Erhebungsp[1]rozess kann einige Zeit dauern. In diesem Fall 1 Stunde 10 Minuten.

7. Vergleich der Tools und der Ergebnisse

Tabelle 1: Allgemeine Daten

Name	Browser	Betriebs-system	Nutzer	Nutzerbewertung	Nutzerfreundlichkeit
Webscraper.io	- Google Chrome - Mozilla Firefox	-mac -windows	Chrome: 305.401 Firefox: 4.232	Chrome **** (676) Firefox **** (8)	- Point & Click Interface - Scraping (dynamische Inhalte) - Output: CSV, XLSX, JSON - Technischer Support - Keine Installation
Helium Scraper	/	-windows	/	/	- Point & Click Interface - Scraping (JavaScript & SQL) - Scraping (dynamische Inhalte) - Output: CSV, XML, SQL,JSON, SQLite - Technischer Support
Scraper Parsers	-Google Chrome	- mac - windows	Chrome: 11.414	Chrome **** (118)	- Point & Click Interface - Scraping (JavaScript & SQL) - Scraping (dynamische Inhalte) - Output: CSV, XML, XLS, XLSX - Technischer Support - Keine Installation

Die durchschnittliche Bewertung im “Mozilla Firefox Add-On Store” für Webscraper.io liegt bei 3,9 von 5 Sternen durch 8 Bewertungen von 4232 Nutzern (vgl. addons.mozilla.org Abrufdatum: 26.02.2020). Vergleichen lässt sich dieser Wert nicht mit den beiden anderen “Tools”, weil diese nicht in diesem Shop angeboten werden. Generell ließ sich keinen Vergleichswert vom “Helium Scraper” zu diesem Aspekt ließ finden, da es sich um ein eigenständiges Programm handelt und somit auch nicht durch Nutzer von “Mozilla Firefox” sowie “Google Chrome” Add-On's bewertet werden können. Die durchschnittliche Bewertung im “Chrome Web Store” für webscraper.io liegt bei 4,1 von 5 Sternen durch 676 Bewertungen von 305.401 Nutzern (vgl. chrome.google.com Abrufdatum: 26.02.2020), wobei durchschnittlich 118 von 11.441 Nutzer das Add-On “Scraper Parsers” mit 4,3 von 5 Sternen einstufen (vgl. chrome.google.com Abrufdatum: 26.02.2020). Die Browser Erweiterungen “Webscraper.io” sowie “Scraper Parsers” laufen auf Windows sowie Mac, weil für die Verwendung, ausschließlich die Installation von den Browsern “Google Chrome”

1

und “Mozilla Firefox” notwendig ist. Helium Scraper hingegen muss auf der Webseite des Anbieters auf den Computer geladen und anschließend installiert werden. Alle drei Programme können dynamische Webseiteninhalte extrahieren. Das “Scraping” von “JavaScript und “SQL” stellt für “Scraper Parsers” sowie “Helium Scraper” keine Probleme dar, wobei bei dem “Webscraper.io” teilweise unvollständige Ergebnisse erzielt werden. Nutzer haben bei allen drei “Tools” die Möglichkeit bei Problemen den jeweiligen technischen “Support Service” zu kontaktieren, um Hilfe zu bekommen.

Tabelle 2: Erhobene Daten aus dem ersten Anwendungsbeispiel

Name	**Leistung: erhobene Daten**	**Dauer: benötigte Zeit zur Erhebung der Daten**	**Geeignet für folgende Anwendungsbereiche:**
Webscraper.io	1933 Zeilen = %3,28	0 Stunden 24 Minuten	- Vermarkter - Datenanalysten - Forscher mit weniger IT-Kenntnissen - E-Commerce
Helium Scraper	1134 Zeilen = %1,92	0 Stunden 35 Minuten	- Vermarkter - Datenanalysten - Forscher mit weniger IT-Kenntnissen - Datenjournalisten - Marktforscher
Scraper Parsers	607 Zeilen = %1,03	1 Stunde 10 Minuten	- Vermarkter - E-Commerce - Entwickler - Manager - Regisseure - Forscher ohne IT-Kenntnisse

Anzahl der Veranstaltungen in Hamburg im Jahr 2020 insgesamt: 59.017 = 100%

Tabelle 3: Erhobene Daten aus dem zweiten Anwendungsbeispiel

Name:	**Leistung: erhobene Daten**	**Dauer: benötigte Zeit zur Erhebung der Daten**
Webscraper.io	1107 > 100%	0 Stunden 15 Minuten
Helium Scraper	667 = 99,1 %	0 Stunden 20 Minuten
Scraper Parser	802 > 100%	1 Stunde 30 Minuten

Gesamtmenge an Produkten (offizielle Inhalte) im Steam Workshop: 673 = 100%

Tabelle 4: Erhobene Daten aus dem dritten Anwendungsbeispiel

Name:	**Leistung: erhobene Daten**	**Dauer: benötigte Zeit zur Erhebung der Daten**
Webscraper.io	12.795 = 39,8%	10 Stunden
Helium Scraper	13.281 = 41,3%	11 Stunden
Scraper Parser	893 = 2,7%	1 Stunde 15 Minuten

Gesamtmenge an Produkten (1-10€) im Konsolenkost Web Shop: 32147

8. Fazit

Das Hauptaugenmerk dieser Bachelorarbeit lag beim Ergründen der Themen “Web Scraping” und den dazugehörigen “Web Scraping Tools”. Aufgrund der Tatsache, dass die Wachstumsrate der weltweit erzeugten Datenmenge rasant ansteigt, sind Programme zur Erfassung und Analyse von Daten unverzichtbar. (vgl. Reinsel/ Gantz/ Rydning 2018, statista 2018). Wenn ein Unternehmer mit der Konkurrenz mithalten möchte, muss er stets auf dem neusten Informationsstand sein. (vgl. Beckmann 2019)

“Marktdaten unterliegen dynamischen Prozessen und sollten deshalb regelmäßig erfasst werden. Wichtig für die Abschätzung zukünftiger Entwicklungen ist die gründliche Auswertung der Daten.” (Mcgrip 2004)

Je häufiger sich mit Themen wie “Open Data” und “Big Data” befasst wird, desto deutlicher wird es welche bedeutsame Rolle “Web Scraping Tools” in Zukunft spielen werden.

Anhand von drei Anwendungsbeispielen wurden in dieser Arbeit verschiedene Daten erhoben. Bei dem ersten Anwendungsbeispiel wurde die Webseite www.hamburg.de durchsucht. Das Ziel war es, alle, bzw. so viele Veranstaltungen wie möglich, welche im Jahr 2020 in Hamburg stattfinden, zu erfassen. Es wurde sich für Add-On’s “Webscraper.io” und “Scraper Parsers” sowie für das Programm “Helium Scraper” entschieden. Die Nutzeroberfläche von “Scraper Parsers” ist sehr einfach und selbsterklärend, jedoch benötigen “Webscraper.io” und “Heliumscraper” einiges an Einarbeitungszeit. Im Gegenzug kann gesagt werden, dass die Arbeit mit “Parsers” zwar die einfachste ist, jedoch können mit den anderen beiden Softwares detailliertere Befehle erteilt werden. Ein Beispiel dafür sind die Ergebnisse aus dem dritten Anwendungsbeispiel. Dabei wurde deutlich, dass “Scraper Parsers” es nicht geschafft hat nur die Produkte zwischen 1 und 10€ zu finden. Stattdessen wurden alle Produkte erfasst, welche gefunden wurden. Dadurch, dass bei der Kostenlosen Version von “Parsers” keine Benutzerdefinierten Einstellungen möglich sind

wie im Vergleich zu “Helium” und “Webscraper”. Es ist jedoch sehr erstaunlich wie der “Scraper Parsers” mithilfe von maschinellem Lernen, ohne jegliche Form der Programmierung häufig zu guten Ergebnissen kommt. Besonders bei dem zweiten Anwendungsbeispiel, sind gute Ergebnisse erzielt wurden. Es ist möglich mit Hilfe der extrahierten Daten zu sagen, wann die besten “Upload-Zeiten” im “Steam Workshop” für Skins (Aussehen für Waffen und Spieler in Videospielen) und Sticker (Spraymotiv im Spiel) für das Spiel “Counter Strike Global Offensive”, sind.

Geduld ist nötig, bei der Arbeit mit “Web Scraping Tools”. Der Datenerhebungsprozess kann je nach Größe des Umfangs des Projektes sehr lange dauern. (siehe Tabelle 4, die benötigte Zeit zur Erhebung von Daten. Es ist von Vorteil verschiedene Tools beherrschen zu können, da je nach Projekt bzw. Arbeitsauftrag das Programm, die Software oder das Tool ausgesucht werden sollte. Durch die Arbeit mit kostenloser Software können zufriedenstellende bis gute Ergebnisse erzielt werden, jedoch ist es aus Sicht eines Unternehmens empfehlenswerter mit Vollversionen zu arbeiten. Auch der Suche nach dem richtigen Tool, ist es von Vorteil, die limitierten Testversionen, wie z.B. Von “Helium Scraper” zu testen. Es ist möglich die Vollversion zehn Tage lang zu nutzen, wodurch dem Anwender alle Funktionen des Programms zur Verfügung stehen, mehr als bei kostenlosen Testversionen.

IIII. Literaturverzeichnis

Beckmann, Dirk (2019): Daten sind der Kern des Marketings. URL: www.artundweise.de/magazin/daten-sind-der-kern-des-marketings (Datum der Recherche: 07.01.2020)

Chrome Web Store (2019): Kundenbewertung von Parsers. URL: https://chrome.google.com/webstore/detail/scraper-parsers-free-web/mhfjedhbggbodliofccpefegbmaoohin/reviews (Datum der Recherche: 15.02.2020)

Chrome Web Store (2019): Scraper Parsers - Free Web Scraping. URL: https://chrome.google.com/webstore/detail/scraper-parsers-free-web/mhfjedhbggbodliofccpefegbmaoohin?hl=de (Datum der Recherche: 15.02.2020)

Chrome Web Store (2019): Web Scraper. URL: https://chrome.google.com/webstore/detail/web-scraper/jnhgnonknehpejjnehehllkliplmbmhn?hl=de (Datum der Recherche: 15.02.2020)

Firefox Addons (2018): Web Scraper – Holen Sie sich diese Erweiterung für Firefox. URL: https://addons.mozilla.org/de/firefox/addon/web-scraper/?src=search (Datum der Recherche 24.02.2020)

Google Trends (2020): Weltweite Suche nach Web Scraping in den letzten 10 Jahren. URL: https://trends.google.de/trends/explore?date=2015-01-01%202020-01-28&q=web%20scraping (Datum der Recherche: 22.02.2020)

Hamburg.de (2020): Veranstaltungskalender Hamburg. URL: https://www.hamburg.de/tickets/ (Datum der Recherche: 02.01.2020)

Helium Scraper (o. J.): Web Scraper: Best Web Scraping Tool to Extract Data from Websites. URL: https://www.heliumscraper.com/eng/ (Datum der Recherche: 20.02.2020)

Hir-Infotech (2019): What is Web Scraping: Introduction, Applications and Best Practices, URL: https://hirinfotech.com/what-is-web-scraping/ (Datum der Recherche: 18.02.2020)

Kroker, Micheal (2018): Weltweite Datenmengen sollen bis 2025 auf 175 Zetabyte wachsen. URL:https://blog.wiwo.de/look-at-it/2018/11/27/weltweite-datenmengen-sollen-bis-2025-auf-175-zetabyte-wachsen-8-mal-so-viel-wie-2017 (Datum der Recherche: 01.02.2020)

Lohrey, Mathias (2019): Weitere Vorteile beim Web Scraping. URL: https://trusted.de/web-scraping#h2-weitere-vorteile-beim-web-scraping (Datum der Recherche: 29.01.2020)

Lohrey, Mathias (2019): Helium Scraper - Fazit der Redaktion URL: https://trusted.de/helium-scraper (Datum der Recherche: 29.01.2020)

McGrip (2004): Marktdaten – Daten des Marktes und der Konkurrenz. URL: https://www.mcgrip.de/0-web/wissen/konkurrenzanalyse/08-marktdaten-konkurrenzanalyse.htm (Datum der Recherche: 24.02.2020)

Patel, Hiren (2018): How Web Scraping is transforming the World with its Applications. URL: www.towardsdatascience.com/https-medium-com-hiren787-patel-web-scraping-applications-a6f370d316f4 (Datum der Recherche: 16.02.2020)

Reinsel, David / Gantz, John / Rydning, John (2018): The Digitalization of the World -From Edge to Core. URL: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf (Datum der Recherche: 06.01.2020)

Ryte Wiki (2019): Scraping. URL: https://de.ryte.com/wiki/Scraping#G.C3.A4ngige_Verwendungszwecke (Datum der Recherche: 08.02.2020)

Savinkin, Igor (2015): Helium Scraper Review. URL: http://scraping.pro/helium-scraper-review/ (Datum der Recherche 27.11.2019)

Stuck, Niels (2019): Was ist Scraping. URL: https://wolf-of-seo.de/was-ist/scraping (Datum der Recherche: 12.12.2019)

Solmecke, Christian (2013): Ist Screen Scraping legal? URL: https://www.wbs-law.de/urheberrecht/ist-screen-scraping-legal-15081 (Datum der Recherche: 01.02.2020)

Semalt.com (2018): 6 Web Scraping Tools zum Erfassen von Daten ohne Codierung. URL: https://semalt.com/de/qa/7401-website-schaber-werkzeug.htm (Datum der Recherche: 19.01.2020)

Wikipedia (2005): Ajax Programmierung. URL: https://de.wikipedia.org/wiki/Ajax_(Programmierung) (Datum der Recherche: 23.02.2020)

Wikipedia (2011): Webhooks. URL: https://de.wikipedia.org/wiki/WebHooks (Datum der Recherche: 11.02.2020)

Yandex (o. J.): Using robots.txt. URL: https://yandex.com/support/webmaster/controlling-robot/robots-txt.html#crawl-delay (Datum der Recherche: 11.01.2020)

Anhang (Datenträger):

Anwendungsbeispiel1: Kulturkalender

- AB1_kulturkalender_hh2020_heliumscraper.xlsx
- AB1_kulturkalender_hh2020_webscraperio.xlsx
- AB1_kulturkalender_hh2020_parsers.xlsx

Anwendungsbeispiel2: Steam Work Shop

- AB2_steam_workshop_webscraperio.xlsx
- AB2_steam_workshop_heliumscraper.xlsx
- AB2_steam_workshop_parsers.xlsx

Anwendungsbeispiel 3: Konsolen und Zubehör

- AB3_Konsolenkost_webscraper.io.xlsx
- AB3_Konsolenkost_heliumscraper.io.xlsx
- AB3_Konsolenkost_parsers.io.xlsx

Erklärung

Ich versichere, dass ich die vorliegende Arbeit ohne fremde Hilfe selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel benutzt habe. Wörtlich oder dem Sinn nach aus anderen Werken entnommene Stellen sind unter Angabe der Quelle kenntlich gemacht.

Erklärung – Einverständnis

Ich erkläre mich damit

o einverstanden,

o nicht einverstanden

dass ein Exemplar meiner Bachelor- (Master-) Thesis in die Bibliothek des Fachbereichs aufgenommen wird; Rechte Dritter werden dadurch nicht verletzt. (Wenn das Unternehmen Bedenken gegen die Veröffentlichung der Bachelor- (Master-) Thesis hat, ist eine schriftliche Begründung der Firma erforderlich). Hamburg, den .... (Unterschrift der/des Studierenden)