Redlich, Sarah

Article

Web Scraping zur Gewinnung von Testdaten für administrative Register

WISTA - Wirtschaft und Statistik

Provided in Cooperation with:
Statistisches Bundesamt (Destatis), Wiesbaden

Suggested Citation: Redlich, Sarah (2020) : Web Scraping zur Gewinnung von Testdaten für administrative Register, WISTA - Wirtschaft und Statistik, ISSN 1619-2907, Statistisches Bundesamt (Destatis), Wiesbaden, Vol. 72, Iss. 3, pp. 24-34

This Version is available at:
https://hdl.handle.net/10419/220341

Sarah Redlich

hat Survey Methodology (M.A.) an der Universität Duisburg-Essen studiert. Seit 2017 ist sie dort am Lehrstuhl für empirische Sozialforschung als wissenschaftliche Mitarbeiterin tätig und promoviert seit 2018 an der Fakultät für Gesellschaftswissenschaften. Für ihre Masterarbeit „Web Scraping zur Gewinnung von Testdaten für administrative Register", die sie in diesem Artikel vorstellt, wurde sie 2019 mit dem Gerhard-Fürst-Preis des Statistischen Bundesamtes ausgezeichnet.

WEB SCRAPING ZUR GEWINNUNG VON TESTDATEN FÜR ADMINISTRATIVE REGISTER

Sarah Redlich

↘ **Schlüsselwörter:** Traueranzeigen – Mortalität – Big Data – Web-Daten – Web Scraping

ZUSAMMENFASSUNG

Web Scraping verspricht eine kosten- und zeiteffiziente Nutzung von öffentlich zugänglichen Informationen aus dem Internet. Um Testdaten für die Erstellung eines bundesweiten Mortalitätsregisters zu generieren, könnten online gestellte Traueranzeigen mittels Web Scraping genutzt werden. Aus diesem Grund wurden für die Jahre 2015 und 2016 für eine Stadt die online extrahierten Traueranzeigen mit amtlichen Daten verglichen. Bereits bei der Konstruktion von Web Scrapern zeigten sich erhebliche Probleme. Ein Vergleich der extrahierten Informationen mit amtlichen Daten zeigt Unterschiede hinsichtlich Gesamtzahl, Geschlecht, ethnischer Zugehörigkeit und Alter der Verstorbenen.

↘ ***Keywords:*** *obituaries – mortality – big data – online data – web scraping*

ABSTRACT

Web scraping promises cost and time-efficient use of open access information from the world wide web. To generate test data for creating a national mortality register, information could be extracted from online obituaries by means of web scrapers. The data scraped from a city's online obituaries of the years 2015 and 2016 were therefore compared with official data. Even building web scrapers caused major problems. A comparison of the information extracted and official data shows differences regarding the total number of deceased, their sex, ethnicity and age.

1

Einleitung

Die Durchführung standardisierter Befragungen wird zunehmend durch den Antwortausfall (Nonresponse) und den Beantwortungsaufwand sowie die Belastung der Befragten (Response Burden) erschwert (Schnell und andere, 2018; Wallgren/Wallgren, 2014). Als alternative Datenquelle dient das Internet, jedoch wäre eine manuelle Extraktion der dort gefundenen Informationen zeitaufwendig und mitunter fehleranfällig (Hoekstra und andere, 2010). Mit Erhebungstechniken, wie dem sogenannten Web Scraping, können öffentlich zugängliche Informationen aus dem Internet nutzbar gemacht werden (Cooley und andere, 1997; Massimino, 2016).

Web Scraping wird bereits erfolgreich in der amtlichen Statistik sowie in den Bereichen Data Science, Journalismus und Marketing genutzt (Landers und andere, 2016; Blaudow/Seeger, 2019). Ein weiteres potenzielles Anwendungsgebiet von Web Scraping ist die Generierung von Testdaten für den Aufbau von administrativen Registern. Beispielsweise wurde bisher in Deutschland kein bundesweites Mortalitätsregister eingeführt. Testdaten sollten Informationen wie Namen, Geburtsdatum und Sterbedatum umfassen. Webb und andere (1966) schlugen vor, Informationen von Grabsteinen zu nutzen. Eine aktuelle Alternative dazu sind online gestellte Traueranzeigen, die mittels Web Scraping in einen nutzbaren Datensatz gebracht werden können. Traueranzeigen als Informationsquelle wurden bereits von Boak und anderen (2007), Soowamber und anderen (2016) und Sylvestre und anderen (2018) erfolgreich erschlossen.[1] Nachfolgend wird untersucht, ob Web Scraping von online verfügbaren Traueranzeigen zu brauchbaren Testdaten führt.

Zunächst wird erläutert, was unter der Datenerhebungstechnik Web Scraping zu verstehen ist (Kapitel 2) und welche ethischen und rechtlichen Grundlagen für Web Scraping gelten (Kapitel 3). Kapitel 4 gibt einen kurzen Überblick zur Entstehung und zum Aufbau von Traueranzeigen und deren Veröffentlichung in Trauerportalen. Kapitel 5 behandelt die Probleme bei der Entwicklung von Web Scrapern zur Extraktion von Traueranzeigen. Die Datensatzbeschreibung und der Vergleich der extrahierten Informationen mit amtlichen Daten folgen in den Kapiteln 6 und 7, bevor im letzten Kapitel eine abschließende Betrachtung folgt.

2

Web Scraping

Web Scraping bezeichnet die Extraktion von Information aus dem World Wide Web (Glez-Peña und andere, 2013; Cooley und andere, 1997). In der Literatur wird Web Scraping auch als Web Harvesting, Web Crawling und Web Mining bezeichnet, wobei lediglich der Begriff Web Harvesting synonym verwendet werden kann (Gatterbauer, 2009; Bharanipriya/Kamakshi Prasad, 2011; Najork, 2009). Web Scraper ahmen die Interaktion zwischen Mensch und Server nach (Glez-Peña und andere, 2013). Dabei wird zunächst auf die Seite zugegriffen. Mittels regulärer Ausdrücke[2] wird dann die gewünschte Information gesucht, anschließend extrahiert und in ein gewünschtes Datenformat gebracht.

Zu den Vorteilen von Web Scraping gegenüber anderen Datenerhebungstechniken zählen die Zeiteffizienz, die häufigere Datensammlung und die Kostenreduzierung. Letztere gilt jedoch nur, sofern der Aufwand der Datenaufbereitung gering ist und Web Scraper nie bis selten überarbeitet werden müssen (Hoekstra und andere, 2010; Landers und andere, 2016). Im Vergleich zu standardisierten Befragungen (Surveys) werden ebenfalls der Response Burden sowie Antwortverzerrungen umgangen (Hoekstra und andere, 2010; Landers und andere, 2016). In der Literatur wird auch betont, dass sich die Qualität von Statistiken verbessert (Hoekstra und andere, 2010; Blaudow/Seeger, 2019).

Von Nachteil ist, dass Web Scraping zunächst viele Ressourcen benötigt, da Personen mit Wissen und Erfahrungen im Bereich Web Scraping eingestellt (oder ausgebildet) werden müssen (Hoekstra und andere, 2010). Zudem können durch sich stark verändernde Webseiten Web Scraper schnell unbrauchbar werden und eine zeitaufwendige Anpassung erfordern (Hoekstra und andere,

1 Die Idee zu dieser Arbeit stammt von Prof. Dr. Rainer Schnell. Teile dieser Arbeit und zusätzliche statistische Tests zur Signifikanz der Ergebnisse finden sich in Schnell/Redlich (2019).

2 Mit einem regulären Ausdruck kann eine bestimmte Zeichenabfolge über syntaktische Regeln in einer Zeichenkette identifiziert werden.

2010). Hinzu kommen durch Web Scraper überladene Server und Kosten aufseiten der Betreiber der Webseite durch das Benutzen der Bandbreite. Die letztgenannten Probleme dürften mittlerweile jedoch zu vernachlässigen sein (Koster, 1993b; Thelwall/Stuart, 2006). Trotzdem ist der Ressourcenaufwand zur Aufbereitung der extrahierten Informationen nicht zu vernachlässigen (Hoekstra und andere, 2010). Die größten Probleme bei der Verwendung von Web Scrapern sind jedoch ethischer und rechtlicher Natur.

3

Ethische und rechtliche Grundlagen

Neben allgemeinen ethischen Richtlinien für Datenerhebungen[3] gilt es zu betonen, dass mittels Web Scraping Informationen von Personen ohne deren Wissen und Zustimmung erhoben werden können (van Wel/Royakkers, 2004). Daher wurde das sogenannte Robots Exclusion Protocol (robots.txt) eingeführt, welches eine Extraktion von Informationen blockiert (Koster, 1993a; Thelwall/Stuart, 2006). Trotz der Möglichkeit, auch dieses zu umgehen, sollte eine Extraktion nur erfolgen, sofern das Robots Exclusion Protocol dies gestattet (Kouzis-Loukas, 2016; Landers und andere, 2016).

Die nachfolgende Darstellung der rechtlichen Grundlagen bezieht sich auf den Stand vom Oktober 2017 und wurde im Hinblick auf die Extraktion von Informationen von online gestellten Todesanzeigen betrachtet. Für eine aktuellere Anwendung oder eine Anwendung in einem anderen Bereich ist eine individuelle Bewertung der Rechtslage zwingend erforderlich.[4]

Web Scraping ist weder explizit erlaubt noch verboten, weshalb es grundsätzlich als legal angesehen wird (Black, 2016). Durch die immer häufigere kommerzielle Nutzung von Web Scrapern werden zunehmend Regelungen getroffen und Urteile gefällt, die nicht immer eine wissenschaftliche Nutzung der Daten berücksichtigen (Black, 2016). Ob die Extraktion von Daten mittels Web Scrapern legal ist, hängt davon ab, ob der Inhalt oder die Anordnung der Daten urheberrechtlich geschützt sind. Zentrale Gesetze sind hierbei § 87 des Urheberrechtsgesetzes sowie die Richtlinie 96/9/EG des Europäischen Parlaments und des Rates über den rechtlichen Schutz von Datenbanken (hier: Kapitel zu Urheberrecht, Schutz Sui Generis). Eine Reihe von Urteilen[5] in den vergangenen Jahren lassen dabei den Schluss zu, dass Web Scraping legal ist – und zwar auch dann, wenn die Nutzungsbedingungen dies verbieten, diesen aber (zum Beispiel durch die Erstellung eines Benutzerkontos) nicht aktiv zugestimmt wurde.

Ferner ist Web Scraping unproblematisch, wenn nur ein unwesentlicher Teil der Datenbank extrahiert wird und keine Veröffentlichung der Daten erfolgt. Zudem erklären sowohl § 87c Absatz 1 Urheberrechtsgesetz als auch die Richtlinie 96/9/EG (in Artikel 6 Absatz 2 und Artikel 9), dass es Ausnahmen für den wissenschaftlichen Gebrauch gibt. Für die Extraktion von Informationen von Traueranzeigen ist zudem von Interesse, dass personenbezogene Daten von Verstorbenen extrahiert werden. Diese unterliegen nicht mehr dem Datenschutz (§ 3 Absatz 1 Bundesdatenschutzgesetz). Allerdings dürfen keine Informationen von lebenden Personen verwendet werden, beispielsweise Namen und Kontaktdaten von Angehörigen (Löwer, 2010).

4

Traueranzeigen als potenzielle Datenquelle

Bereits seit dem 18. Jahrhundert wird der Tod eines Menschen mittels einer Traueranzeige bekannt gemacht; sie ersetzte die aufwendigere Form der sonntäglichen Kanzelabkündigung (Grümer/Helmrich, 1994). In Deutschland wurde 1753 die erste (belegte) Traueranzeige veröffentlicht, erst rund zehn Jahre später war in Zeitungen eine eigene Rubrik zu Traueranzeigen zu finden (Grümer/Helmrich, 1994). Seit Anfang des 19. Jahrhunderts werden Traueranzeigen regelmäßig in eigenen Rubriken in Zeitungen abgedruckt (Grümer/Helmrich, 1994).

3 Beispielsweise der American Association for Public Opinion Research (American Association for Public Opinion Research, 2015).

4 So erlaubt zum Beispiel das Preisstatistikgesetz durch die Änderung im Dezember 2019 die Erhebung von Preisen mit automatischen Abrufverfahren zur Erstellung benötigter Statistiken (§ 7 Absatz 2 Preisstatistikgesetz).

5 Zu den Urteilen zählen 6 U 221/08 des OLG Frankfurt am Main, 308 O 162/09 des LG Hamburg und 3 U 191/08 des OLG Hamburg sowie I ZR 159/10 des Bundesgerichtshofs und Urteil C-30/14 des Europäischen Gerichtshofs.

Obwohl es keinen aufgezwungenen Standard (mit Ausnahme der Größe, die durch den Preis bedingt wird) seitens der Zeitungen hinsichtlich des Aussehens und des Informationsgehalts von Todesanzeigen gibt, hat sich im Lauf der Jahre eine Standardform etabliert. Inhaltlich lassen sich in der Regel folgende Informationen finden: Vorspruch, Symbol, persönliche Daten der verstorbenen Person (Name, Titel, Geburtsname, Geburtsdatum, Sterbedatum), Trauertext, Namensliste der Hinterbliebenen und Hinweise religiöser, organisatorischer oder sonstiger Art (Grümer/Helmrich, 1994). ↘ Grafik 1

Grafik 1
Aufbau von Traueranzeigen

<table>
<tr><td rowspan="2">Symbol</td><td>Vorspruch</td></tr>
<tr><td>Name, Titel
Geburtsname
Geburtsdatum, Sterbedatum

Trauertext</td></tr>
<tr><td colspan="2">Überleitung
Angehörige</td></tr>
<tr><td colspan="2">Sonstige Hinweise</td></tr>
</table>

Geänderte Darstellung aus Grümer/Helmrich, 1994.

2020 - 01 - 0206

Durch die zunehmende Technisierung und die zunehmende Nutzung des Internets finden sich immer mehr sogenannte Trauerportale, zumeist von Zeitungsverlagen betrieben, die eine Online-Veröffentlichung von Traueranzeigen anbieten. Es gibt auch Trauerportale, die unabhängig von Zeitungen arbeiten, jedoch ist hier nicht ersichtlich, ob ein Beleg über den Tod in Form einer Sterbeurkunde vorgelegt werden muss, wie es bei Zeitungen verlangt wird. Ein Vorteil von Trauerportalen für die Nutzung von Traueranzeigen als Datenquelle ist, dass sowohl die gedruckte Anzeige als Bild vorhanden ist, als auch die interessierenden Angaben (Name, Geburts- und Sterbedaten) aufgeführt werden. Dies erleichtert eine Extraktion der Informationen.

Eine genaue Angabe zur Anzahl an Trauerportalen gibt es nicht. Daher wurde eine Liste aller online vertretenen (Tages-)Zeitungen in Deutschland erstellt; über deren Internetpräsenzen konnten 152 Trauerportale von Zeitungen identifiziert werden, zudem ein Trauerportal, welches mehr als nur eine Zeitung/Zeitungsgruppe umfasst.[6] Fünf dieser Trauerportale sind registrierungs- oder kostenpflichtig und daher für das Web Scraping ungeeignet. Die restlichen konnten hinsichtlich ihres Layouts und der zugrunde liegenden Quellcodes in drei Gruppen eingeteilt werden. Dabei befinden sich in der letzten Gruppe Trauerportale mit jeweils individuellen Quellcodes, sodass jedes Trauerportal einen eigenen Web Scraper benötigen würde. Diese Trauerportale wurden nicht berücksichtigt. Für Gruppe 1 und Gruppe 2 wurde jeweils ein Web Scraper entwickelt, somit konnten insgesamt 109 Trauerportale abgedeckt werden. Bei den nicht berücksichtigten Trauerportalen handelt es sich jedoch ausschließlich um Portale aus anderen Regionen als der in diesem Artikel betrachteten, sodass hier keine Traueranzeigen unbeachtet blieben.

6 Stand: Oktober 2017.

5

Web Scraping von Traueranzeigen

Die Web Scraper wurden in Python 2.7 unter der Verwendung von Scrapy 1.3.3 geschrieben.[7] Die Web Scraper sollten dabei den Vornamen und Nachnamen der verstorbenen Person einschließlich Geburtsname und Zweitnamen sowie das Geburtsdatum, das Sterbedatum, die Zeitung der Veröffentlichung und das Datum der Veröffentlichung extrahieren. Dabei wurde bereits durch die Start-URL[8] nach Ort und Datum gefiltert, sodass nur Anzeigen für den interessierenden Ort und den Zeitraum extrahiert wurden.

Bei der Entwicklung von Web Scrapern zeigten sich einige Probleme. So gab es für einige Personen mehr als ein Bild der Traueranzeige, beispielsweise, weil eine zweite Traueranzeige als Jahresgedenken veröffentlicht wurde, oder weil mehrere Anzeigen in unterschiedlichen Zeitungen veröffentlicht wurden. Dadurch entstanden unzulässige Links für den Download des Bilds der Traueranzeige. In diesem Fall wurde immer das erste auftauchende Bild verwendet.

Ein weiteres Problem bestand darin, dass meist nur ein kleineres Vorschaubild der Anzeige gegeben war und größere Bilder mit einer besseren Auflösung in JavaScript eingebettet waren. Zu kleine Bilder machen

7 Eine Einführung bietet Kouzis-Loukas (2016).

8 Bei der Start-URL handelt es sich um die URL der Webseite des Portals, nachdem in der Suchmaske des Portals die Filter nach Ort und Datum gesetzt wurden.

eine Texterkennung unmöglich und sind daher nicht nutzbar. Bilder, die in JavaScript eingebettet sind, sind zwar prinzipiell extrahierbar, allerdings konnte während eines angemessenen Zeitraums keine Lösung gefunden werden, um die Bilder zu extrahieren. Des Weiteren wiesen heruntergeladene Bilder einen alphanumerischen Dateinamen auf, welcher eine Zuordnung zu einer Person nicht möglich macht. Daher musste eine zusätzliche Liste erstellt werden, die den Namen des heruntergeladenen Bilds der Person zuordnet.

Bilder der Traueranzeigen können theoretisch verwendet werden, um die darauf enthaltenen Informationen zu extrahieren. Dies ist besonders von Interesse, wenn keine Auflistung der interessierenden Information stattfindet. Mittels OCR(Optical Character Recognition)-Texterkennung ist es theoretisch möglich, alle notwendigen Informationen zu extrahieren. Bei Bildern mit schlechter Auflösung, der Verwendung von Hintergrundbildern oder der Verwendung ungewöhnlicher Schriftarten (zum Beispiel Kalligraphie-Schriftarten) unterlaufen der Texterkennung viele Fehler. Versuche, dennoch Informationen aus den Bildern zu extrahieren, waren erfolglos.

Bei Trauerportalen der zweiten Gruppe sind die Datumsangaben nicht separat aufgeführt, sondern nur in einer 300-Zeichen-Vorschau des Textes der Traueranzeige eingebettet. Hier mussten also aus einem Fließtext das Geburtsdatum und das Sterbedatum extrahiert werden. Dabei erwiesen sich die unterschiedlichen Datumsformate als problematisch, ebenso fehlende Geburtsdaten und zusätzlich vorkommende Daten, wie der Tag der Beerdigung. Dieses Problem war nur durch die Annahme zu lösen, dass es sich bei den ersten zwei gefundenen Daten um das Geburtsdatum sowie das Sterbedatum handelt und – sofern nur ein Datum gefunden wurde – nur das Sterbedatum erfasst wurde.

Das gravierendste Problem beim Web Scraping ist jedoch die mögliche Veränderung des Quellcodes von Webseiten. Hierdurch werden Web Scraper im schlimmsten Fall vollkommen unbrauchbar. Zudem kann die Anzahl der gefundenen Traueranzeigen zwischen verschiedenen Zeitpunkten variieren – beispielsweise, weil ältere Traueranzeigen später hinzugefügt oder Anzeigen aus den Trauerportalen entfernt werden. Schließlich kann sich auch der öffentliche Zugang zu den Webseiten ändern, indem das Robots Exclusion Protocol der Webseite einen Zugang unterbindet.

6

Datensatzbeschreibung

Für die Jahre 2015 und 2016 wurden für eine Großstadt 6 597 Traueranzeigen gefunden. Nachdem Anzeigen außerhalb des interessierenden Zeitraums und doppelte Anzeigen gelöscht wurden, enthielt der Datensatz 3 007 Anzeigen von verstorbenen Personen. Auf Grundlage von amtlichen Daten wurden basierend auf den Vornamen der Personen das Geschlecht und die ethnische Zugehörigkeit zugewiesen. Drei Namen, für die über die amtlichen Daten kein Geschlecht zugewiesen werden konnte, wurde manuell ein Geschlecht zugeordnet. Hierbei handelte es sich um Namen mit eindeutiger Geschlechtszuweisung.

Die Traueranzeigen wurden mit amtlichen Daten verglichen. In Deutschland besteht nach § 28 Personenstandsgesetz die Verpflichtung, den Tod eines Menschen binnen drei Tagen dem zuständigen Standesamt zu melden. Damit enthalten die Daten des Standesamts alle im Einzugsbereich des Standesamts verstorbene Personen, unabhängig von deren Wohnsitz. Die Informationen des Standesamts werden allerdings nur aggregiert als Anzahl der Personen getrennt nach Jahr, Geschlecht und ethnischer Zugehörigkeit zur Verfügung gestellt. Das Standesamt ist nach § 60 Absatz 1 Personenstandsverordnung verpflichtet, andere Behörden über den Tod zu informieren. Für Mikrodaten wurde auf amtliche Daten zurückgegriffen, die nur die verstorbenen Personen mit Wohnsitz in der interessierenden Stadt enthalten. Die Mikrodaten umfassten Nachnamen, Rufnamen, Geburtsdaten und Todesdaten. Unterschiede zeigten sich bereits zwischen diesen beiden Datensätzen. Während das Standesamt 15 001 verstorbene Personen registrierte, enthielten die amtlichen Daten der Verstorbenen mit Wohnsitz in der interessierenden Stadt nur 14 003 verstorbene Personen.

7

Vergleich der Trauerportale mit amtlichen Daten

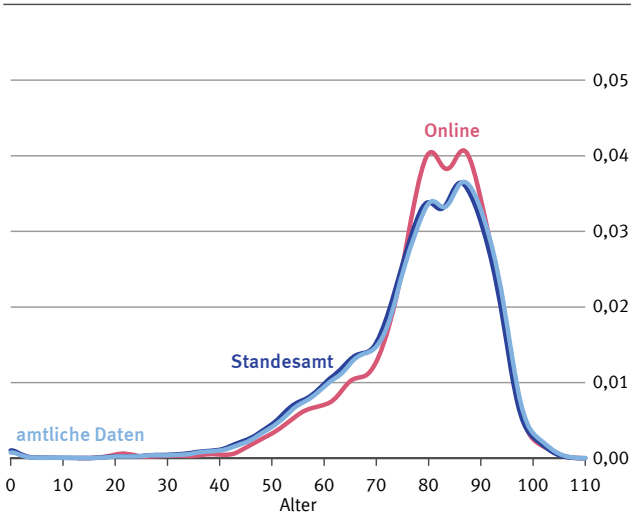
Für den Vergleich der Trauerportale mit den amtlichen Daten wurden zunächst die Datensätze hinsichtlich ihrer Geschlechterzusammensetzung, der ethnischen Zugehörigkeit und der Altersverteilung einander gegenübergestellt. Darüber hinaus wurde überprüft, ob die Datensätze übereinstimmen und inwieweit sich die übereinstimmenden Verstorbenen von den nicht übereinstimmenden Verstorbenen unterscheiden. Statistische Tests zur Signifikanz der gefundenen Unterschiede finden sich in Schnell/Redlich (2019).

7.1 Vergleich der Datensätze

Ein Vergleich der Geschlechteranteile zeigt Folgendes: Während die amtlichen Daten einen höheren Frauenanteil bei den Verstorbenen aufweisen (52,4 % nach den Standesamtsdaten; 52,9 % nach den amtlichen Daten der Verstorbenen mit Wohnsitz in der interessierenden Stadt), zeigen die online extrahierten Daten einen marginal höheren Männeranteil (lediglich 49,6 % der Verstorbenen waren demnach Frauen). Hinsichtlich der ethnischen Zugehörigkeit zeigen zwar alle Datensätze einen deutlich höheren Anteil an Verstorbenen mit deutscher Zugehörigkeit, jedoch waren laut amtlichen Daten 96,2 % (Standesamtsdaten) beziehungsweise 97,4 % (amtliche Daten der Verstorbenen mit Wohnsitz in der interessierenden Stadt) verstorbene Personen deutsch, während die online extrahierten Daten einen Anteil von 99,2 % aufweisen.

↘ Grafik 2 zeigt die Altersverteilung der Verstorbenen in den drei Datensätzen. In den Online-Daten finden sich mehr verstorbene Personen im Alter von 75 bis 90 Jahren und weniger im Alter von 40 bis 70 Jahren im Vergleich zu den amtlichen Daten. Zudem lassen sich keine Personen unter 18 Jahren finden. Aus diesem Grund wurden für eine Zeitung für das Jahr 2015 die gedruckten Todesanzeigen mit den online gestellten Anzeigen verglichen. Online konnten 1 212 Anzeigen gefunden werden, während jedoch 2 766 Anzeigen abgedruckt wurden. Die Übereinstimmung zwischen den Anzeigen betrug 1 134 Personen. Dies bestätigt, dass Angehörige wählen können, ob die Traueranzeige auch online gestellt wird. Durch den Vergleich zeigte sich, dass es durchaus Anzeigen von Minderjährigen gibt und diese auch online gestellt werden. Allerdings kann aufgrund von fehlenden Datumsangaben das Alter nicht berechnet werden.

Grafik 2
Altersverteilung der Online-Daten und amtlichen Daten
nach Dichte



Verstorbene Personen der Online-Daten sind älter.

2020 - 01 - 0193

7.2 Übereinstimmung der Datensätze

Zusätzlich fand ein Abgleich der Datensätze statt, um den Anteil der Personen zu bestimmen, die über Traueranzeigen gefunden werden können. Dies war entsprechend nur mit den Mikrodaten möglich. Einer Person wurde der Status „Duplikat“ gegeben, wenn der Name, das Geburtsdatum und das Sterbedatum übereinstimmten. Sofern das Geburtsdatum nicht vorhanden war, musste nur eine Übereinstimmung von Name und Sterbedatum gegeben sein. Von den online extrahierten Daten konnten 70,5 % der Verstorbenen in den amtlichen Daten gefunden werden, aber 84,9 % der Verstorbenen laut amtlichen Daten war keine Traueranzeige zuzuordnen.

Die genauere Betrachtung der Geschlechteranteile zeigt, dass anteilig mehr Männer in beiden Datensätzen zu finden sind. Frauen sind häufiger nur in den amtlichen Daten oder nur in den Online-Daten zu finden. Hinsicht-

lich der ethnischen Zugehörigkeit ist der Anteil an nicht deutschen Verstorbenen bei den Personen, die sich in beiden Datensätzen befinden, deutlich niedriger. Dies legt erneut nahe, dass das Veröffentlichen von Traueranzeigen nicht in allen ethnischen Gruppen praktiziert wird. ↘ Tabelle 1, Tabelle 2

Tabelle 1

Geschlechteranteile unterscheiden sich zwischen Personen, die sich in beiden Datensätzen befinden, und Personen, die sich nur in einem der Datensätze befinden

	Duplikat		Einmalig	
	Anzahl	%	Anzahl	%
Weiblich	1 050	49,6	6 785	53,2
Männlich	1 066	50,4	5 961	46,8
Insgesamt	2 116	100	12 746	100

Tabelle 2

Die ethnische Zugehörigkeit zeigt, dass für nichtdeutsche Personen weniger Traueranzeigen veröffentlicht werden

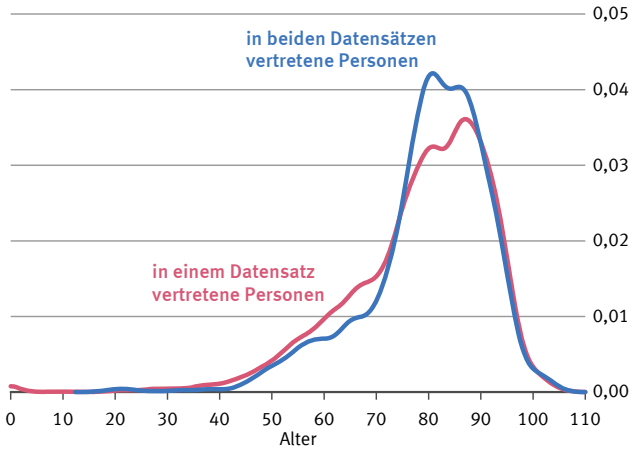
	Duplikat		Einmalig	
	Anzahl	%	Anzahl	%
Deutsch	2 103	99,4	12 376	97,2
Nichtdeutsch	12	0,6	356	2,8
Insgesamt	2 115	100	12 732	100

Eine Betrachtung der Altersverteilung zeigt, dass Personen, die sich nicht in beiden Datensätzen befinden, jünger sind. Zudem zeigt sich auch hier wieder deutlich das Problem, dass Minderjährige nicht über Traueranzeigen gefunden wurden. ↘ Grafik 3

Für Traueranzeigen, für die keine Entsprechung in den amtlichen Daten gefunden werden konnte, zeigte sich, dass die Ortseingrenzung einen unzureichenden Filter darstellte. So können Orte, deren Namen einen Teil eines anderen Orts bilden, nicht ausreichend differenziert werden; somit werden Personen eines anderen Orts berücksichtigt. Die Information des Orts ist jedoch nur aus dem Bild der Traueranzeige zu entnehmen, sodass keine Optimierung des Filters erreicht werden konnte. Zudem wurden Unstimmigkeiten in der Schreibweise von Namen gefunden sowie die Verwendung von Spitz- oder Rufnamen statt der offiziellen Vornamen. Schließlich wurden auch abweichende Geburtsdaten und Sterbedaten gefunden, die eine Zuordnung erschweren.

Grafik 3

Altersverteilung von Personen, die in beiden Datensätzen, und Personen, die in nur einem Datensatz zu finden sind
nach Dichte



Jüngere Personen lassen sich nicht in beiden Datensätzen finden.

2020 - 01 - 0208

8

Schlussbetrachtung

Im Fokus der Arbeit stand die Entwicklung eines Web Scrapers zur Extraktion von Traueranzeigen. Web Scraping von öffentlich zugänglichen Daten bietet die Möglichkeit, den Response Burden von zu befragenden Personen vollständig zu reduzieren und erlaubt eine häufigere Erhebung von Informationen, wie sie in einem Survey nicht möglichen wären (Hoekstra und andere, 2010; Landers und andere, 2016).

Traueranzeigen stellen eine mögliche Datenquelle für verstorbene Personen dar und könnten theoretisch als Testdaten für ein bundesweites Mortalitätsregister genutzt werden. Durch die zunehmende Zahl sogenannter Trauerportale, welche Traueranzeigen online zugänglich machen, können diese mittels Web Scraping extrahiert werden.

Obwohl für einen Großteil der Trauerportale zwei Web Scraper ausreichen, zeigten sich erhebliche Probleme bei deren Entwicklung, beispielsweise durch die Veränderung des zugrunde liegenden Quellcodes. Des Weiteren sind interessierende Informationen nicht bei allen Trauerportalen einzeln aufgeführt, sondern zwischen

anderem Text oder in JavaScript eingebettet, was eine Extraktion erschwert. Eine Texterkennung von Bildern der Traueranzeigen erwies sich als wenig erfolgreich.

Ein Abgleich der extrahierten Informationen mit amtlichen Daten zeigte einen höheren Anteil an männlichen Verstorbenen sowie einen deutlich geringeren Anteil an nicht deutschen Verstorbenen in den Online-Daten. Auch bei der Altersverteilung zeigen sich erhebliche Unterschiede. Verstorbene der extrahierten Todesanzeigen sind deutlich älter als die amtlichen Daten zeigen. Ein Abgleich hinsichtlich des Namens, des Geburtsdatums sowie des Sterbedatums zeigte, dass rund 85 % der laut amtlichen Daten verstorbenen Personen nicht über eine Traueranzeige gefunden werden konnten. Personen, die entweder nicht über eine Traueranzeige gefunden werden konnten oder zum Beispiel aufgrund von ungenauen Filtern nicht in den amtlichen Daten sind, sind eher weiblich, jünger und der Anteil nicht deutscher Personen ist vergleichsweise größer. Eine Verwendung von geeigneteren Verfahren zur Zusammenführung von Datensätzen[9] könnte auch diejenigen Übereinstimmungen trotz Tippfehlern finden. Allerdings dürfte auch dies kaum die Unterschiede hinsichtlich der Anzahl und der demografischen Charakteristika der Verstorbenen von online gestellten Traueranzeigen zu amtlichen Daten kompensieren.

9 In der Literatur finden sich diese Verfahren unter dem Stichwort „Record Linkage“.

LITERATURVERZEICHNIS

American Association for Public Opinion Research (AAPOR). *The Code of Professional Ethics and Practices (Revised 11/30/2015)*. 2015. [Zugriff am 20. April 2020]. Verfügbar unter: www.aapor.org

Bharanipriya, V./Kamakshi Prasad, V. *Web Content Mining Tools: A Comparatative Study*. In: International Journal of Information Technology and Knowledge Management. Ausgabe 4(1)/2011, Seite 211 ff.

Black, Michael L. *The World Wide Web as Complex Data Set: Expanding the Digital Humanities into the Twentieth Century and Beyond through Internet Research*. In: International Journal of Humanities and Arts Computing. Ausgabe 10(1)/2016, Seite 95 ff. DOI: 10.3366/ijhac.2016.0162

Blaudow, Christian/Seeger, Daniel. *Fortschritte beim Einsatz von Web Scraping in der amtlichen Verbraucherpreisstatistik – Ein Werkstattbericht*. In: WISTA Wirtschaft und Statistik. Ausgabe 4/2019, Seite 19 ff.

Boak, Marshall B./M'ikanatha, Nkuchia M./Day, Roger S./Harrison, Lee H. *Internet Death Notices as a Novel Source of Mortality Surveillance Data*. In: American Journal of Epidemiology. Ausgabe 167(5)/2007, Seite 532 ff.

Cooley, Robert/Mobasher, Bamshad/Srivastava, Jaideep. *Web Mining: Information and Pattern Discovery on the World Wide Web*. In: 9th International Conference on Tools with Artificial Intelligence, ICTAI '97, Newport Beach, CA, USA, November 3-8, 1997. Los Alamitos 1997, Seite 558 ff. DOI: 10.1109/TAI.1997.632303

Gatterbauer, Wolfgang. *Web Harvesting*. In: Liu, Ling/Özsu, M. Tamer (Herausgeber). Encyclopedia of Database Systems. New York 2009, Seite 3472 f.

Glez-Peña, Daniel/Lourenço, Anália/López-Fernández, Hugo/Reboiro-Jato, Miguel/Fdez-Riverola, Florentine. *Web Scraping Technologies in an API World*. In: Briefings in Bioinformatics. Ausgabe 15(5)/2013, Seite 788 ff. DOI: https:doi.org/10.1093/bib/bbt026

Grümer, Karl-Wilhelm/Helmrich, Robert. *Die Todesanzeige: Viel gelesen, jedoch wenig bekannt: Deskription eines wenig erschlossenen Forschungsmaterials*. In: Historical Research. Ausgabe 19(1)/1994, Seite 60 ff.

Hoekstra, Rutger/Bosch, Olav Ten/Harteveld, Frank. *Automated Data Collection from Web Sources for Official Statistics: First Experiences*. Technischer Bericht 2010-132-KOO. Statistics Netherlands 2010.

Koster, Martijn. *About /robots.txt*. 1993a. [Zugriff am 20. April 2020]. Verfügbar unter: http://www.robotstxt.org/robotstxt.html

Koster, Martijn. *Guidelines for Robot Writers*. 1993b. [Zugriff am 20. April 2020]. Verfügbar unter: http://www.robotstxt.org/guidelines.html

LITERATURVERZEICHNIS

Kouzis-Loukas, Dimitrios. *Learning Scrapy: Learn the Art of Efficient Web Scraping and Crawling with Python*. Birmingham 2016.

Landers, Richard N./Brusso, Robert C./Cavanaugh, Katelyn J./Collmus, Andrew B. *A Primer on Theory-Driven Web Scraping: Automatic Extraction of Big Data From the Internet for Use in Psychological Research*. In: Psychological Methods. Ausgabe 21(4)/2016, Seite 475 ff.

Löwer, Wolfgang. *Anhang 5: Grenzen der Erhebung und Verarbeitung von Sterbedaten durch den (postmortalen) Persönlichkeitsschutz – Überlegungen zum Schutzumfang verstorbener Personen und der lebenden Angehörigen*. In: Dicke, Peter/Dietel, Manfred/Gawrich, Stefan/Jöckel, Karl-Heinz/Klug, Stefanie/Löwer, Wolfgang/Luttmann, Sabine/Müller, Ulrich/Schelhase, Torsten/Vennemann, Mechthild/Ziese, Thomas/Schmidt-Stolte, Martina (Herausgeber). Ein Nationales Mortalitätsregister für Deutschland: Bericht der Arbeitsgruppe und Empfehlung des Rates für Sozial- und Wirtschaftsdaten (RatSWD). 2010. Seite 31 ff.

Massimino, Brett. *Accessing Online Data: Web-Crawling and Information-Scraping Techniques to Automate the Assembly of Research Data.* In: Journal of Business Logistics. Ausgabe 37(1)/2016, Seite 34 ff. DOI: https://doi.org/10.1111/jbl.12120

Najork, Marc. *Web Crawler Architecture*. In: Liu, Ling/Özsu, M. Tamer (Herausgeber). Encyclopedia of Database Systems. New York 2009, Seite 3462 ff.

Schnell, Rainer/Hill, Paul B./Esser, Elke. *Methoden der empirischen Sozialforschung*. 11. Auflage. Berlin 2018.

Schnell, Rainer/Redlich, Sarah. *Web Scraping Online Newspaper Death Notices for the Estimation of the Local Number of Deaths*. In: Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies. Volume 5: HEALTHINF. 2019, Seite 319 ff.

Soowamber, Medha L./Granton, John T./Bavaghar-Zaeimi, Fatemeh/Johnson, Sindhu R. *Online Obituaries Are a Reliable and Valid Source of Mortality Data*. In: Journal of Clinical Epidemiology. Ausgabe 79/2016, Seite 167 ff.

Sylvestre, Emmanuelle/Bouzille, Guillaume/Breton, Mathias/Cuggia, Marc/Campillo-Gimenez, Boris. *Retrieving the Vital Status of Patients with Cancer Using Online Obituaries*. In: Ugon, Adrien/Karlsson, Daniel/Klein, Gunnar O./Moen, Anne (Herausgeber). Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth. 2018, Seite 571 ff.

Thelwall, Mike/Stuart, David. *Web Crawling Ethics Revisited: Cost, Privacy, and Denial of Service*. In: Journal of the American Society for Information Science and Technology. Ausgabe 57(13)/2006, Seite 1771 ff.

LITERATURVERZEICHNIS

van Wel, Lita /Royakkers, Lambèr. *Ethical Issues in Web Data Mining*. In: Ethics and Information Technology. Ausgabe 6(2)/2004, Seite 129 ff.

Wallgren, Anders/Wallgren, Britt. *Register-based Statistics: Statistical Methods for Administrative Data*. 2. Auflage. Chichester 2014.

Webb, Eugene J./Campbell, Donald T./Schwartz, Richard D./Sechrest, Lee. *Unobtrusive Measures: Nonreactive Research in the Social Sciences*. Chicago 1966.

RECHTSGRUNDLAGEN

Bundesdatenschutzgesetz (BDSG) vom 30. Juni 2017 (BGBl. I Seite 2097), das durch Artikel 12 des Gesetzes vom 20. November 2019 (BGBl. I Seite 1626) geändert worden ist.

Gesetz über die Preisstatistik (PreisStatG) in der im Bundesgesetzblatt Teil III, Gliederungsnummer 720-9, veröffentlichten bereinigten Fassung, das zuletzt durch Artikel 1 des Gesetzes vom 10. Dezember 2019 (BGBl. I Seite 2117) geändert worden ist.

Gesetz über Urheberrecht und verwandte Schutzrechte (Urheberrechtsgesetz) vom 9. September 1965 (BGBl. I Seite 1273), das zuletzt durch Artikel 1 des Gesetzes vom 28. November 2018 (BGBl. I Seite 2014) geändert worden ist.

Personenstandsgesetz (PStG) vom 19. Februar 2007 (BGBl. I Seite 122), das zuletzt durch Artikel 17 des Gesetzes vom 20. November 2019 (BGBl. I Seite 1626) geändert worden ist.

Richtlinie 96/9/EG des Europäischen Parlaments und des Rates vom 11. März 1996 über den rechtlichen Schutz von Datenbanken (Amtsblatt der EG Nr. L 77 Seite 20).

Verordnung zur Ausführung des Personenstandsgesetzes (Personenstandsverordnung – PStV) vom 22. November 2008 (BGBl. I Seite 2263), die zuletzt durch Artikel 5 des Gesetzes vom 18. Dezember 2018 (BGBl. I Seite 2639) geändert worden ist.

Herausgeber
Statistisches Bundesamt (Destatis), Wiesbaden

Schriftleitung
Dr. Daniel Vorgrimler
Redaktionsleitung: Juliane Gude
Redaktion: Ellen Römer

Ihr Kontakt zu uns
www.destatis.de/kontakt

Erscheinungsfolge
zweimonatlich, erschienen im Juni 2020
Das Archiv älterer Ausgaben finden Sie unter www.destatis.de

Artikelnummer: 1010200-20003-4, ISSN 1619-2907