Common Crawl

Analyse des Crawling-Prozesses und der Textextraktion

Klemens Schölhorn Erik Körner Kai Hainke

17. März 2016

Inhaltsverzeichnis

1 Common Crawl

"Our goal is to democratize the data so everyone, not just big companies, can do high quality research and analysis." – *Common Crawl Foundation*

1.1 Organisation

Common Crawl ist eine 2007 durch Gil Elbaz gegründete gemeinnützige Organisation mit dem Ziel, Wissenschaftlern, Firmen und Privatpersonen kostenlos eine Kopie des Internets zu Forschungs- und Analysezwecken zur Verfügung zu stellen[1].

Dazu wird monatlich ein Crawl erstellt, der anschließend in einem AWS Public Data Set[2] von Amazon kostenfrei gespeichert wird. Dies ermöglicht die direkte Analyse in AWS, aber auch den freien Download über HTTP.

Die Organisation hat zur Zeit 25 Mitglieder, wobei die Mehrzahl Fachwissen und Zeit unentgeltlich zur Verfügung stellen.

1.2 Technik

Common Crawl verwendet für das Crawlen den CCBot, einen modifizierten Apache Nutch 1.7, welcher wiederum auf Lucene, Solr und Hadoop basiert. Das Crawlen findet verteilt in AWS statt. Der CCBot beachtet die `robots.txt` und versucht durch einen ausgeklügelten Algorithmus so wenig Last wie möglich auf den einzelnen Webservern zu verursachen.

Zur Speicherung der Crawls wird das Web Archive-Format (WARC) verwendet, das in ISO 28500[3] spezifiziert und an HTTP angelehnt ist. Eine WARC-Datei besteht aus mehreren einzelnen WARC-Records, wobei ein WARC-Record einen beliebigen Payload haben kann (z. B. HTTP oder JSON).

Während des Crawls werden drei verschiedene Ergebnisse erzeugt: Zuerst der Crawl an sich mit den kompletten HTTP-Konversationen, wobei Anfrage und Antwort als getrennte WARC-Records abgespeichert werden. Zusätzlich wird für jeden WARC-Record ein JSON-Metadaten-Record[4] (WAT) erzeugt, der u. a. die kompletten HTTP-Header sowie eine Liste aller auf der Seite vorkommenden Links (inklusive der eingebundenen CSS/JavaScript-Dateien) enthält. Schließlich wird für jede HTTP-Antwort auch ein Reintext-Record (WET) erzeugt, der den automatisch aus dem HTML-Quelltext extrahierten Text enthält.

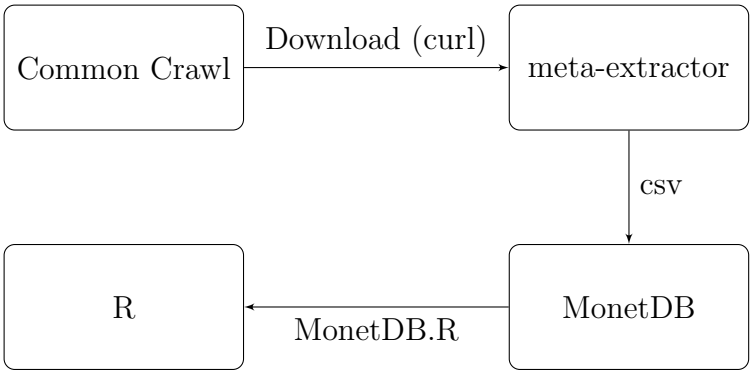
[1] `https://commoncrawl.org/faqs/`
[2] `https://aws.amazon.com/datasets/41740`
[3] `http://bibnum.bnf.fr/WARC/WARC_ISO_28500_version1_latestdraft.pdf`
[4] `https://gist.github.com/Smerity/e750f0ef0ab9aa366558#file-bbc-pretty-wat`

2 Analyse

Die Analyse erfolgte anhand des Crawls vom April 2015, des zum Zeitpunkt des Vortrages aktuellsten Crawls. Der komplette Crawl besitzt eine Größe von 168 TB und beinhaltet 2,1 Milliarden Webseiten[5]. Da eine Bearbeitung dieser Datenmenge auf eigenen Servern aufgrund von Bandbreitenbeschränkungen und eine Bearbeitung in AWS mit Hilfe von EC2-Instanzen aufgrund von finanziellen Einschränkungen nicht möglich war, haben wir unsere Analyse auf die WAT-Dateien beschränkt, die aus knapp 39000 Dateien zu durchschnittlich je 315 MiB bestehen und damit zusammen nur rund 12 TB groß sind. Davon konnten wir ca. 53 % analysieren.

2.1 Übersicht

Die nebenstehende Abbildung zeigt den strukturellen Ablauf der Analyse.



Der Download und die Extraktion der benötigten Daten mit Hilfe des `meta-extractor`s erfolgte parallel auf dem zur Verfügung gestellten Praktikumsrechner und auf einem privaten Server eines Autors. Dabei wurden keine Dateien lokal zwischengespeichert, sondern die WAT-Dateien mittels Unix-Pipes direkt von `curl` in den `meta-extractor` und anschließend die resultierenden csv-Dateien mittels `ssh` direkt auf den genannten privaten Server zur weiteren Analyse geladen.

Dazu wurden die Daten in die spaltenorientierte Datenbank `MonetDB`[6] importiert, um anschließend mittels der `R`-Anbindung darauf in `R` Analysen durchführen zu können.

2.2 `meta-extractor`

Der `meta-extractor` ist ein in `C++` geschriebenes Programm, das sequentiell Metadaten-WARC-Records aus WAT-Dateien auslesen und daraus die im Abschnitt 2.4 beschriebenen Metadaten als csv-Datei extrahieren kann.

Dazu wurden neben einem csv-Writer ein standardkonformer WARC-Parser und eine Trie-Implementierung für den effizienten Lookup des `Public Suffix` einer Domain aus der `Public`

[5] `http://blog.commoncrawl.org/2015/05/april-2015-crawl-archive-available/`
[6] `https://www.monetdb.org/`

`Suffix`-Liste[7] implementiert. Außerdem wurde auf RapidJSON[8] zum schnellen Parsen der WAT-Metadaten, auf die URI-Klasse des POCO-Projektes[9] und auf TCLAP[10] für das CLI zurückgegriffen. Das Test-Framework Catch[11] wurde für Unit-Tests der selbst geschriebenen Komponenten und für Integrationstests verwendet.

Das beiliegende `Makefile` ist für den `GCC C++`-Compiler konzipiert, der Code lässt sich jedoch mit jedem standardkonformen `C++11`-Compiler übersetzen.

2.3 Hilfsmittel

Neben dem `meta-extractor` wurden noch einige Skripte u. a. zur Steuerung des Downloads und der Extraktion, dem Import in `MonetDB` und der Analyse in `R` geschrieben. Das erste Skript ist etwas umfangreicher und daher neben dem Hauptprogramm auch Teil des Projektrepositories[12].

2.4 Gesammelte Metadaten

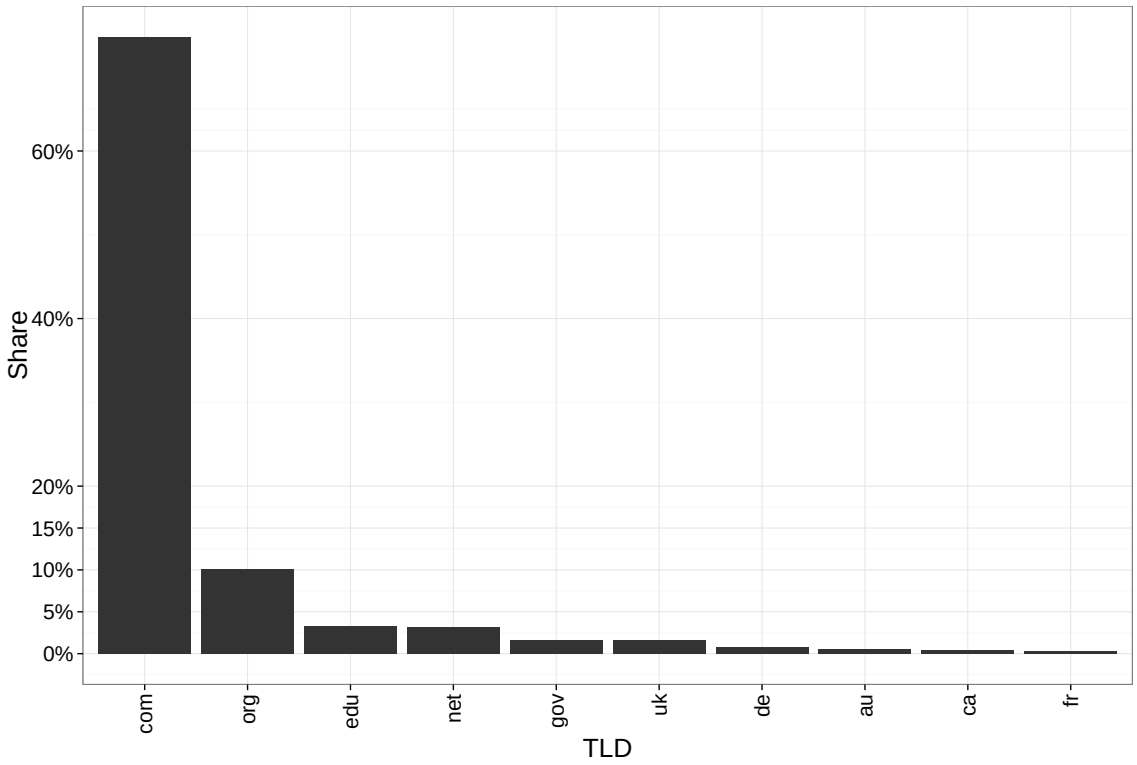
Die in folgender Tabelle beschriebenen Metadaten wurden vom `meta-extractor` extrahiert und für die weiteren Analysen verwendet:

Name	**Datentyp**	**Beispiel**
UUID	uint128	14a939f4-2355-11e5-b5f7-727283247c7f
Zeitstempel	string	2015-07-06T11:03:18T
Verwendung TLS	bool	https ja/nein
Hostname	string	amazon.co.uk
TLD	string	uk
Public Suffix	string	co.uk
Pfadtiefe	uint8	*test/cool/cool.html* → 3
Pfadlänge	uint16	*test/cool/cool.html* → 19
Server	string	apache, nginx, ...
Kompression	bool	gzip, deflate in *Content-Encoding*
Cookies	bool	*Set-Cookie* vorhanden
MIME-Typ	string	text/html, application/xml, ...
Charset	string	utf, iso-8859, ...
Verwendung CDN	bool	CDN ja/nein
Interne Links	uint16	Anzahl relativer Links + gleicher Host
Externe Links	uint16	Anzahl ausgehender Links

[7] `https://publicsuffix.org/`
[8] `https://github.com/miloyip/rapidjson`
[9] `http://pocoproject.org/`
[10] `http://tclap.sourceforge.net/`
[11] `https://github.com/philsquared/Catch`
[12] `https://github.com/klemens/ALI-CC/`

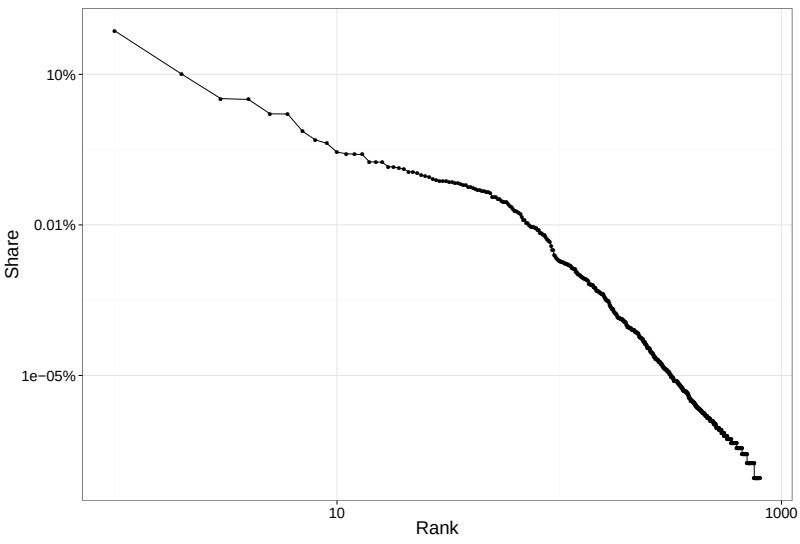
3 Ergebnisse

3.1 Top-Level-Domain



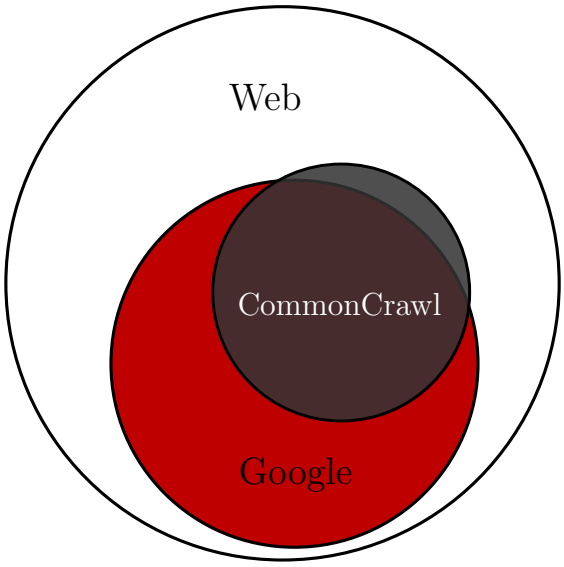
Die TLD .com, die vor allem von kommerziellen Unternehmen genutzt wird, ist mit über 70% mit Abstand am meisten vertreten. Danach folgt die eher von nicht-kommerziellen Organisationen benutzte .org.

Mit .edu und .gov befinden sich zwei TLDs unter den Top-5, die primär von US-amerikanischen Bildungs- und Regierungsorganisationen verwendet werden. Anschließend folgen länderspezifische TLDs, wobei englischsprachige und europäische Länder den überwiegenden Anteil stellen.

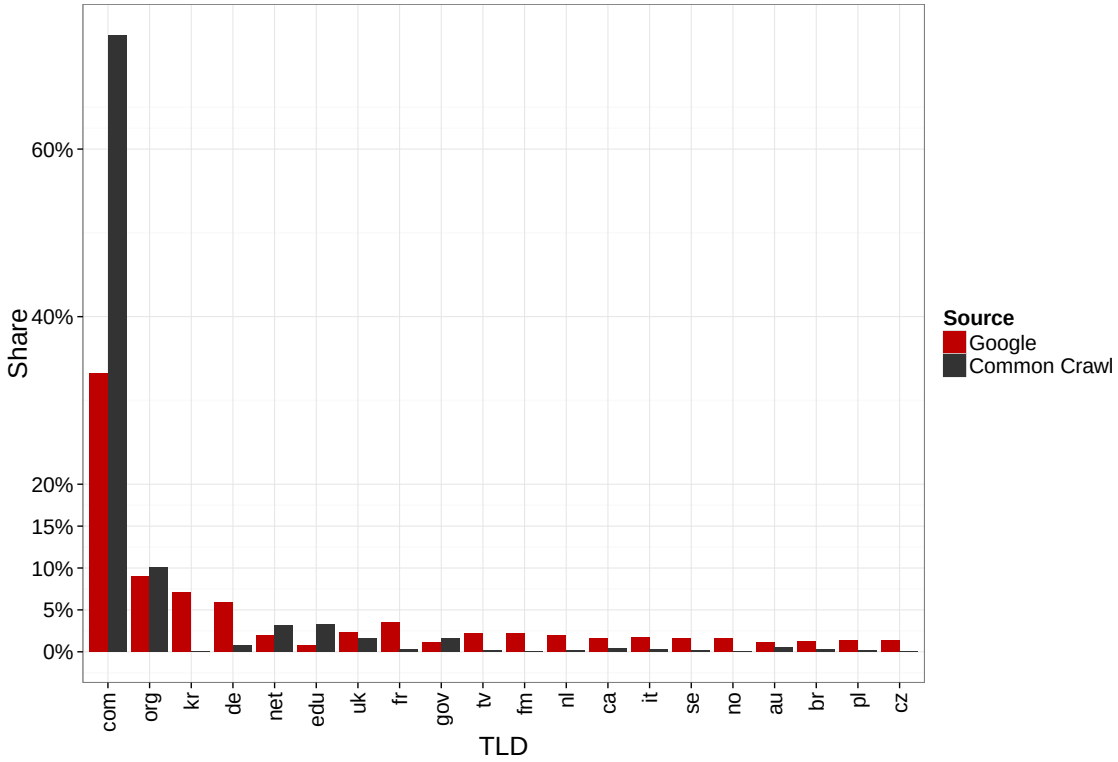


Grundsätzlich scheinen die TLDs näherungsweise zipfverteilt zu sein.

Die von Common Crawl erfassten Seiten stellen eine nicht randomisierte Stichprobe aus der Gesamtheit aller Websites dar. Daher unterliegen die Ergebnisse einer statistischen Verzerrung im Vergleich zur wahren Verteilung der Grundgesamtheit aller Websites, dem Selection Bias. Um abzuschätzen, in wie weit die Auswahl von Websites durch den Common Crawl-Prozess im Vergleich zu einer randomisierten Auswahl verzerrt ist, vergleichen wir die Verteilung der TLDs mit Daten von Google. Auch diese Daten unterliegen natürlich dem Selection Bias, haben aber eine größere Stichprobengröße.

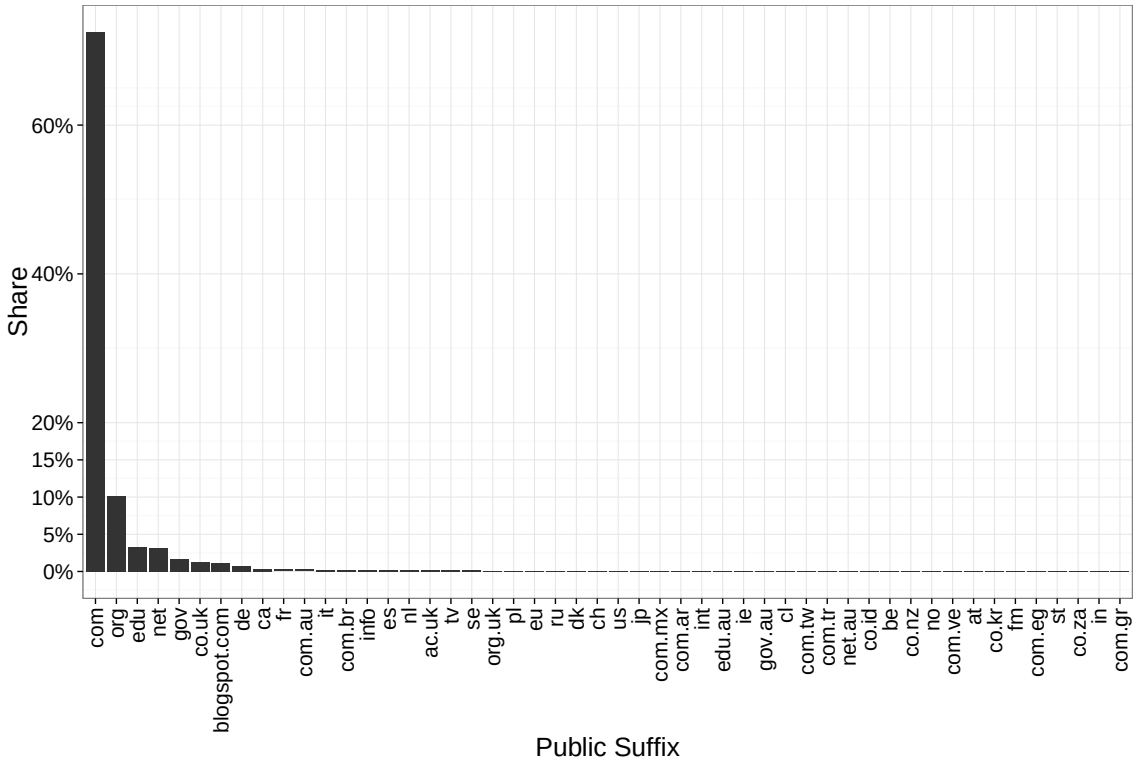


Da Google selbst keine offiziellen Statistiken über die erfassten TLD bereitstellt, wurden die Daten über eine speziell formulierte Suchanfrage der Form `site:.com` erhoben. Dies liefert alle Seiten, die in Ihrer Domain den String .com enthalten, was näherungsweise alle Seiten der TLD .com sind. Auf diese Weise wurden die Trefferzahlen der 61 häufigsten Domains ermittelt und in Relation zueinander gesetzt. Aufgrund der nicht vollständigen Abdeckung aller TLDs sind die relativen Anteile geringfügig zu hoch.



Es zeigt sich, dass mit Ausnahme englischsprachiger Länder (besonders asiatische) länderspezifische TLDs stark unterrepräsentiert sind (z. B. Südkorea .kr), während amerikanische TLDs (.edu, .gov) und die kommerzielle TLD .com stark überrepräsentiert sind. Typische TLDs von Websites für Medienstreaming (z. B. .tv und .fm) sind ebenfalls unterrepräsentiert.

3.2 Public Suffix



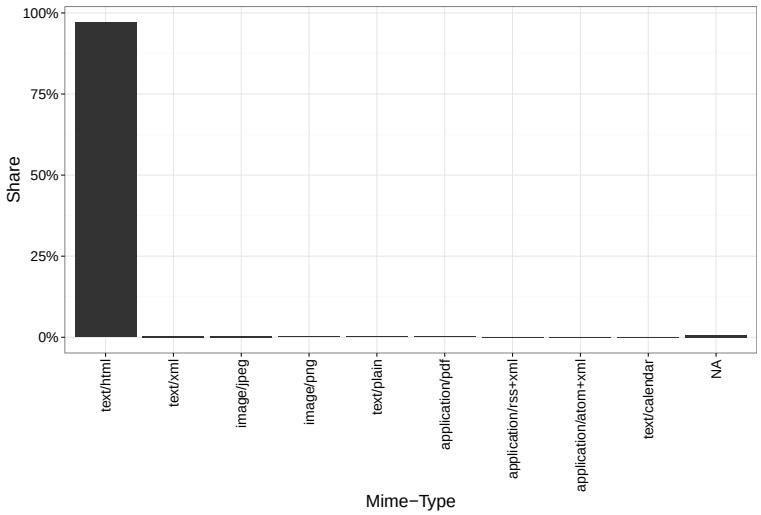
Wie zu erwarten zeigt sich bei den Public Suffixes ein ähnliches Bild wie bei den TLDs. Ein extremer Ausreißer ist der Blog-Dienst blogger.com mit seinen auf blogspot.com gehosteten Blogs.

Außerdem erscheinen nun z. B. .co.uk, .org.uk, etc. weit vor .uk (nicht im Graph), da die Registrierung von Domains direkt unter .uk erst seit kurzem erlaubt und damit noch nicht weit verbreitet ist.

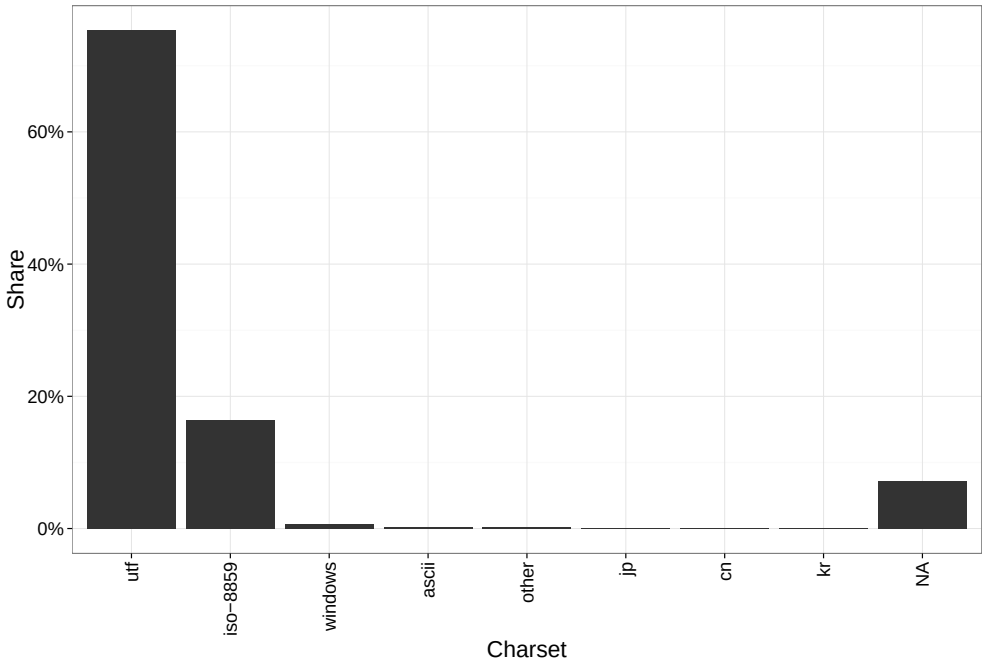
3.3 Mime-Type

Es werden fast ausschließlich HTML-Textdaten gecrawlt, was den Erwartungen entspricht. Andere typische gecrawlte Typen sind Bilder, PDFs, Feeds und Kalender.

Da die eigentlichen Daten nicht analysiert wurden und die HTTP-Header möglicherweise nicht immer korrekt sind (z. B. falsch konfigurierter Webserver), sollten diese Daten kritisch gesehen werden.



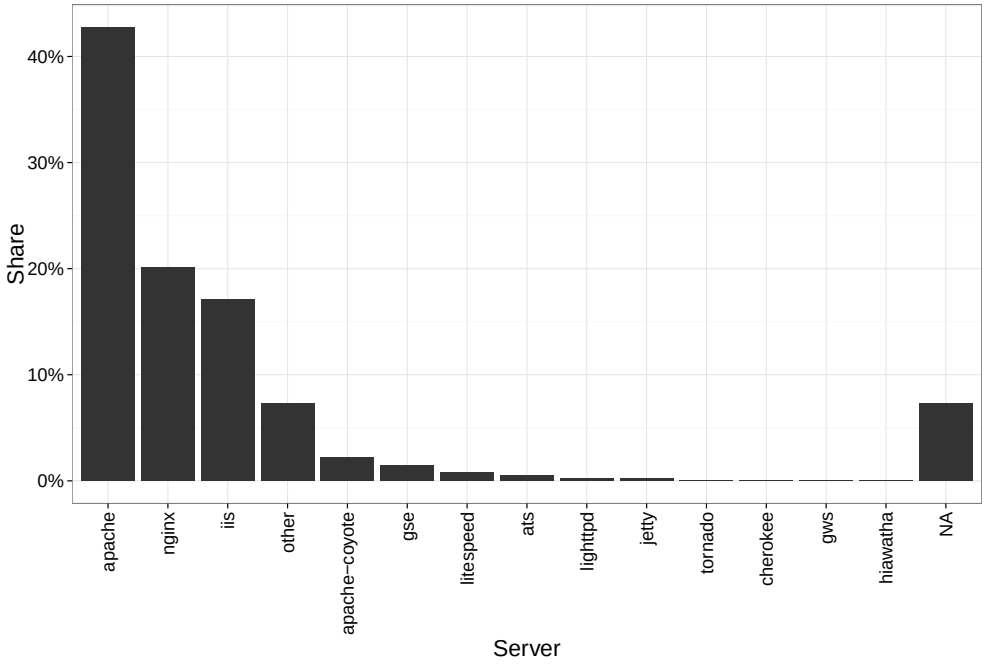
3.4 Zeichenkodierung



Während ca. 16% aller Seiten noch die regionsspezifischen Zeichenkodierungen des ISO-8859-Standards benutzen, setzt die große Mehrheit inzwischen auf Unicode.

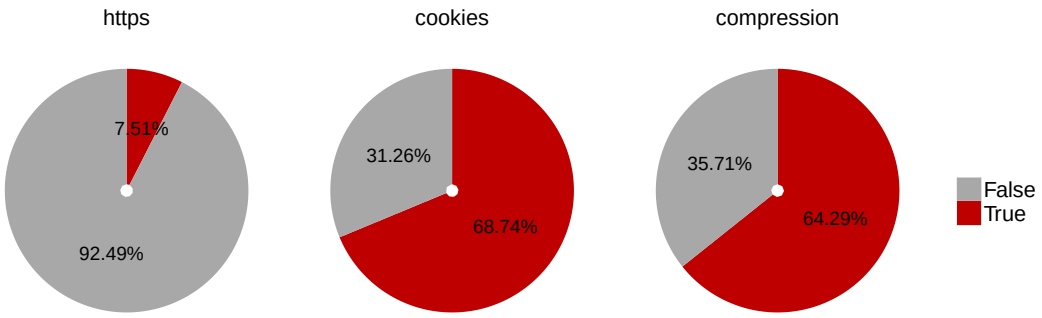
Außerdem zeigt sich hier abermals, dass asiatische Seiten stark unterrepräsentiert sind, da die dort immer noch verwendeten Zeichenkodierungen wie die des JIS-Standards oder die Big5-Zeichenkodierung (im Graph nach Ländern gruppiert) quasi nicht auftreten.

3.5 Webserver



Bei den verwendeten Webservern herrscht eine große Vielfalt (vgl. den großen `other`-Anteil), die von den beiden großen freien Webservern Apache und nginx angeführt wird.

3.6 Verschlüsselung, Cookies und Kompression

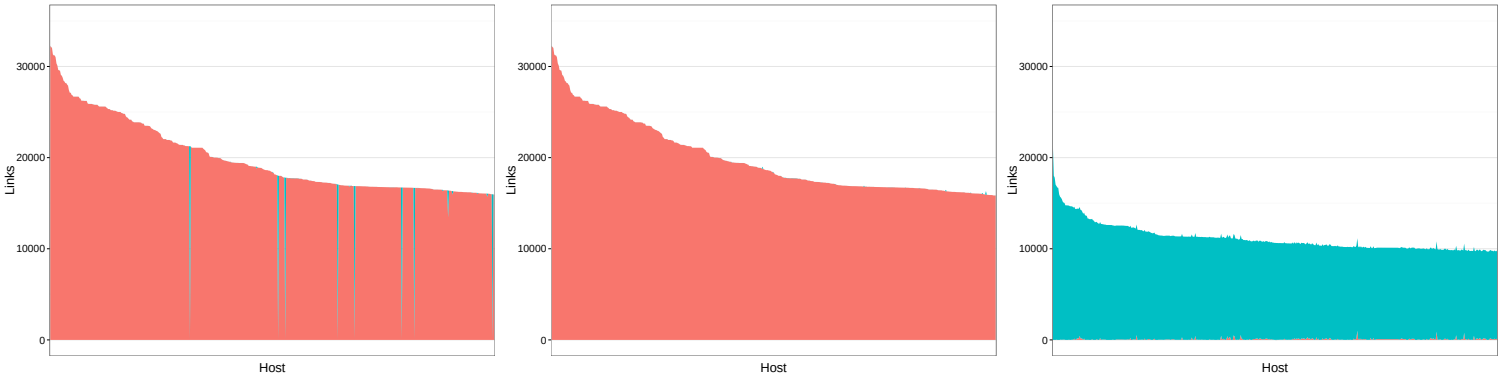


Mehr als 7% der von Common Crawl gecrawlten Seiten verwenden HTTPS zur verschlüsselten Übertragung der Inhalte. Dies zeigt eindrucksvoll den seit ca. 4 Jahren anhaltenden Trend zur Verschlüsselung: 2011 lag der Anteil der verschlüsselten Seiten noch weit unter 1%[13].

Im Gegensatz dazu versuchen mehr als zwei Drittel aller Seiten beim Besucher mindestens einen Cookie zu setzen. Es wurde hier jedoch nicht zwischen Session-Cookies und persistenten Cookies unterschieden.

Geringfügig weniger Seiten werden komprimiert (dh. mit `gzip` oder `deflate`) zum Browser übertragen.

3.7 Links



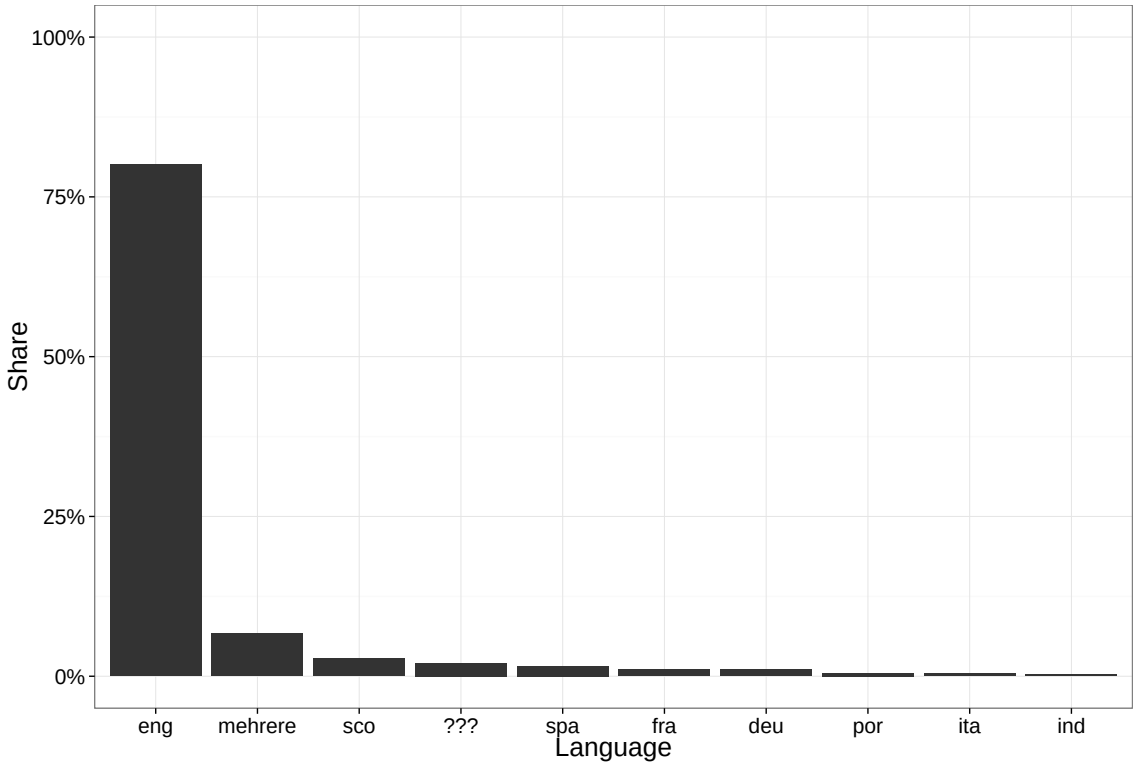
Seiten mit den meisten (links), meisten internen (mitte, rot) und meisten externen (rechts, blau) Links

Um die Reichweite des Crawls genauer zu beleuchten, betrachten wir die Verteilung und Art der Links, die auf den Seiten gefunden werden. Dabei unterscheiden wir zwischen internen Links, d.h. solche mit einem relativen Pfad bzw. mit dem gleichen Hostnamen als Ziel, und externen Links. In Bezug auf die Reichweite sind besonders externe Links interessant, da nur durch diese neue Seiten von anderen Hosts in den Crawl mit einbezogen werden.

Von allen Seiten enthalten etwa 84% Links und 67% externe Links. Ungefähr 18% aller Links sind extern, der Rest verlinkt intern. Seiten mit sehr vielen externen/internen Links enthalten jedoch nahezu ausschließlich Links des selben Typs, die Verteilung ist eher asymmetrisch.

[13] `http://www.sistrix.de/news/anteil-von-ssl-treffern-google-serps/`

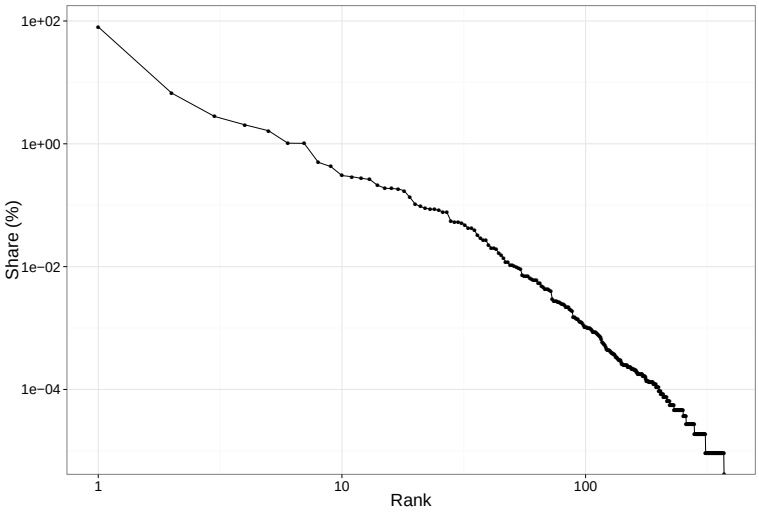
3.8 Sprache



Die Analyse der TLDs am Anfang des Kapitels kann nur einen groben Überblick über die Sprachverteilung geben. Aus diesem Grund wurde zusätzlich zu den Metadaten auch ein Teil des eigentlichen Crawls analysiert. Dabei wurde eine zufällige Stichprobe von 242 aus insgesamt ca. 33000 WARC-Dateien (gepackt je ca. 1 GB) heruntergeladen und der Text extrahiert. Für die resultierenden knapp 11 Millionen Dokumente wurde anschließend mit dem LangSepa-Werkzeug der Fakultät jeweils die Sprache bestimmt.

Wie erwartet ist der Crawl sehr stark von Englisch dominiert. Die einzigen nicht-europäischen Sprachen innerhalb der ersten 20 (über 95 % aller Dokumente) sind Indonesisch (3 ‰), Vietnamesisch (1,7 ‰) und Arabisch (1,6 ‰). Koreanisch folgt auf Platz 24 mit 0,8 ‰ und Chinesisch auf Platz 31 mit 0,5 ‰. Insgesamt wurden 367 verschiedene Sprachen erkannt.

LangSepa scheint manchmal Probleme bei der Unterscheidung ähnlicher Sprachen zu haben. So haben Scots (27,9 ‰) und das nigerianische Pidgin (1,9 ‰) einen unrealistisch hohen Anteil. Da beide jedoch dem Englischen sehr ähnlich sind, kann man hier von einer Fehlklassifizierung eigentlich englischer Dokumente ausgehen.



Die Verteilung der Sprachen entspricht – deutlicher noch als die der TLDs – einer Zipf-Verteilung.

4 Textextraktion

In diesem Abschnitt vergleichen wir die Qualität der Textextraktion von Common Crawl mit dem an der Fakultät entwickelten jWarcEx, das auch bei der Vorverarbeitung der Dokumente für die Sprachanalyse mit LangSepa im vorherigen Abschnitt zum Einsatz kam, und mit Readability.js, das von Mozilla für die Leseansicht in Firefox verwendet wird.

Eine Analyse der Textextraktion durch Vergleich aller Dokumente stellte sich aufgrund der großen Datenmenge und Schwierigkeiten bei der Definition von Metriken für einen automatischen Vergleich als zu kompliziert heraus. Aus diesem Grund erfolgte die Analyse stichprobenartig anhand einiger Beispiele, wobei die folgenden beiden exemplarisch die Unterschiede und Gemeinsamkeiten der verschiedenen Werkzeuge aufzeigen.

4.1 Beispiel: theguardian.com

Das erste Beispiel ist der erste Artikel von Glenn Greenwald, Ewen MacAskill und Laura Poitras über Edward Snowden auf `theguardian.com`[14]:

Common Crawl und jWarcEx extrahieren in diesem Fall den exakt gleichen Text, der im Wesentlichen der DOM-Eigenschaft `textContent` entspricht und z. B. auch Elemente der Seitennavigation und Alternativtexte von Medien (vlg. Hinweis zum Video) enthält.

Readability.js extrahiert dagegen exakt den Text des Artikel. Die Überschrift und die Autoren werden aus den entsprechenden HTML-Meta-Tags entnommen, wobei hier fälschlicherweise nur das letzte author-Tag beachtet wird.

[14] `http://www.theguardian.com/world/2013/jun/09/edward-snowden-nsa-whistleblower-surveillance`

Common Crawl / jWarcEx

Skip to main content

browse all sections close

Glenn Greenwald on security and liberty

Edward Snowden: the whistleblower behind the NSA surveillance revelations

The 29-year-old source behind the biggest intelligence leak in the NSA's history explains his motives, his uncertain future and why he never intended on hiding in the shadows

Q\&A with NSA whistleblower Edward Snowden: ''I do not expect to see home again''

Glenn Greenwald, Ewen MacAskill and Laura Poitras in Hong Kong

Tuesday 11 June 2013 09.00 EDT Last modified on Saturday 4 October 2014 10.54 EDT

Sorry, your browser is unable to play this video.

The individual responsible for one of the most significant leaks in US political history is Edward Snowden, a 29-year-old former technical assistant for the CIA and current employee of the defence contractor Booz Allen Hamilton. Snowden has been working at the National Security Agency for the last four years as an employee of various outside contractors, including Booz Allen and Dell.

The Guardian, after several days of interviews, is revealing his identity at his request. From the moment he decided to disclose numerous top-secret documents to the public, he was determined not to opt for the protection of anonymity. "I have no intention of hiding who I am because I know I have done nothing wrong," he said.

(...)

Readability.js

Edward Snowden: the whistleblower behind the NSA surveillance revelations

Laura Poitras

The individual responsible for one of the most significant leaks in US political history is Edward Snowden, a 29-year-old former technical assistant for the CIA and current employee of the defence contractor Booz Allen Hamilton. Snowden has been working at the National Security Agency for the last four years as an employee of various outside contractors, including Booz Allen and Dell.

The Guardian, after several days of interviews, is revealing his identity at his request. From the moment he decided to disclose numerous top-secret documents to the public, he was determined not to opt for the protection of anonymity. "I have no intention of hiding who I am because I know I have done nothing wrong," he said.

(...)

4.2 Beispiel: abc7.com

Unterschiede zwischen der Textextraktion von Common Crawl und jWarcEx zeigen sich z. B. an einem Artikel über das Karriereende von Lance Armstrong auf `abc7.com`[15].

So erkennt jWarcEx die zu einem Menüpunkt gehörenden Unterpunkte und fasst diese zusammen, während Common Crawl damit teilweise Probleme hat. Außerdem fügt Common Crawl manchmal den Seitentitel an den Anfang des Extraktes hinzu.

Common Crawl	jWarcEx
Lance Armstrong, Tour de France champ, retires | abc7.com	
GO	
Personalize your weather by entering a location.	Personalize your weather by entering a location.
Sorry, but the location you entered was not found. Please try again.	Sorry, but the location you entered was not found. Please try again.
Sections	Sections Traffic Video Los AngelesOrange CountyInland EmpireVentura CountyCalifornia
Traffic	
Video	
Los AngelesOrange CountyInland EmpireVentura CountyCalifornia	
Home	Home Accuweather Traffic Video Photos Mobile Apps
Accuweather	
Traffic	
Video	
Photos	
Mobile Apps	
Local News Los AngelesOrange CountyInland EmpireVentura CountyCalifornia	Los AngelesOrange CountyInland EmpireVentura CountyCalifornia
Map My News	
Shows Live Well Network	
Follow Us	
BREAKING NEWS	BREAKING NEWS ABC shows live and on-demand — Download the WATCH ABC app!
(...)	(...)

[15] `http://abc7.com/archive/7962461`