Analyse von Common Crawl

Erik Körner Kai Hainke Klemens Schölhorn

06.07.2015

Übersicht

Common Crawl
Organisation , Technik

Analyse
Extraktion , Daten

Ergebnisse
TLD , Public Suffix , Mime-Type , TLS , Encoding , Server

Textextraktion

Common Crawl: Organisation

- Non-Profit Organisation
- 2007 durch Gil Elbaz gegründet
- Mitglieder stellen unentgeldlich Fachwissen und Zeit zur Verfügung

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

Common Crawl: Organisation

- (Fast) monatliche Crawls des Web und freie Bereitstellung der Datensätze
- Letzter Crawl: 168 TB, 2,1 Mrd. Websites[1]
- Viele Code-Beispiele zur Verarbeitung der Daten
- Hosting gesponsert von AWS (Public Data Set), gesamt: 541 TB[2]

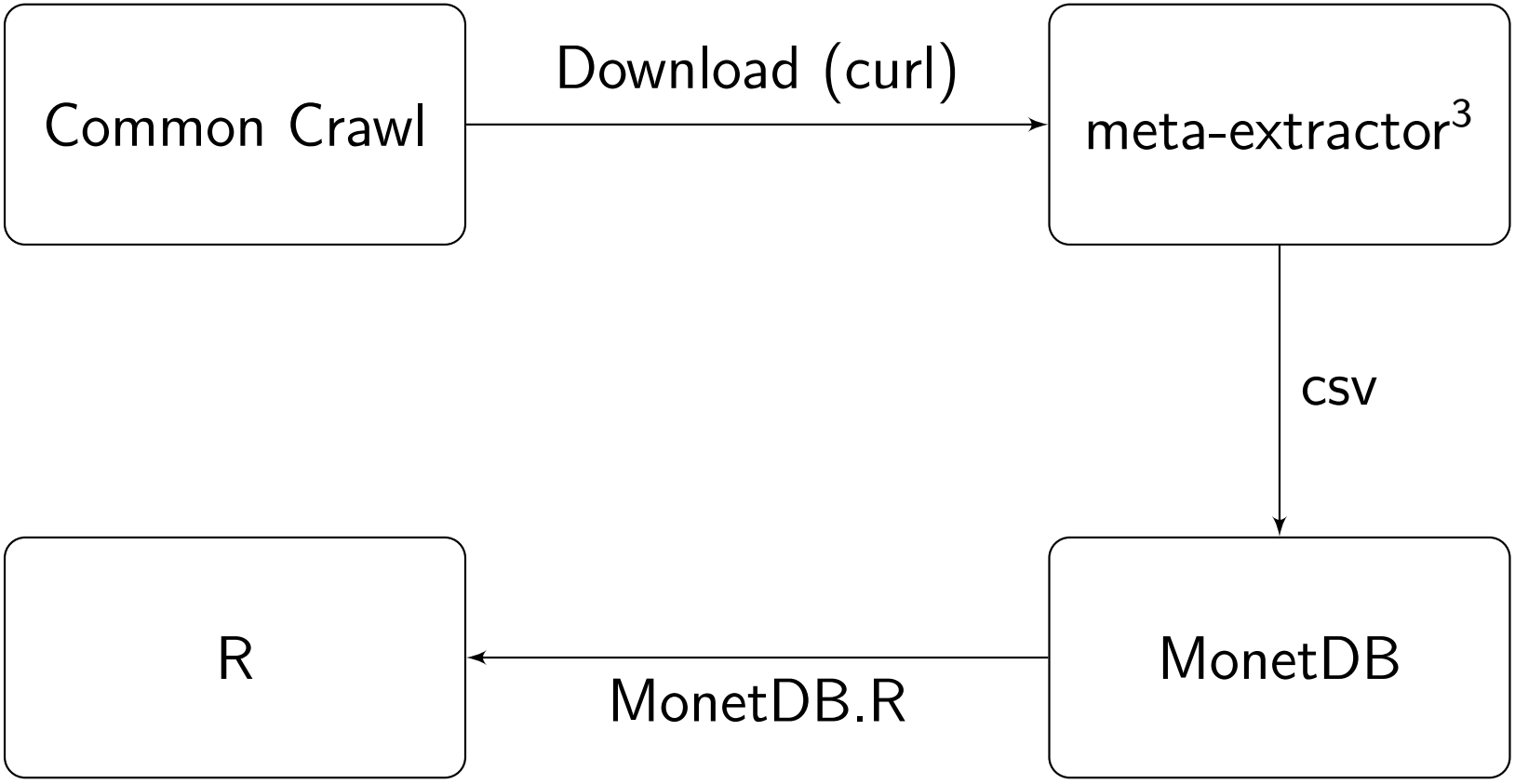
[1] `http://blog.commoncrawl.org/2015/05/april-2015-crawl-archive-available/` (Stand: Mai 2015)

[2] `https://aws.amazon.com/datasets/41740`

Common Crawl: Technik

- Speicherung im WARC-Format (Web ARChive)
- CCBot: modifizierter Apache Nutch 1.7
  - Basiert auf Lucene, Solr und Hadoop
  - Verteiltes Crawling bei AWS
  - „Freundlicher Crawler“: beachtet robots.txt und erzeugt so wenig Last wie möglich
- Drei verschiedene Ergebnisse
  - WARC: Enthält komplette HTTP Konversationen getrennt nach Anfrage und Antwort (Records)
  - WAT: Metadaten für jeden Record
  - WET: Reintext-Extrakt der Antworten

Analyse: Extraktion

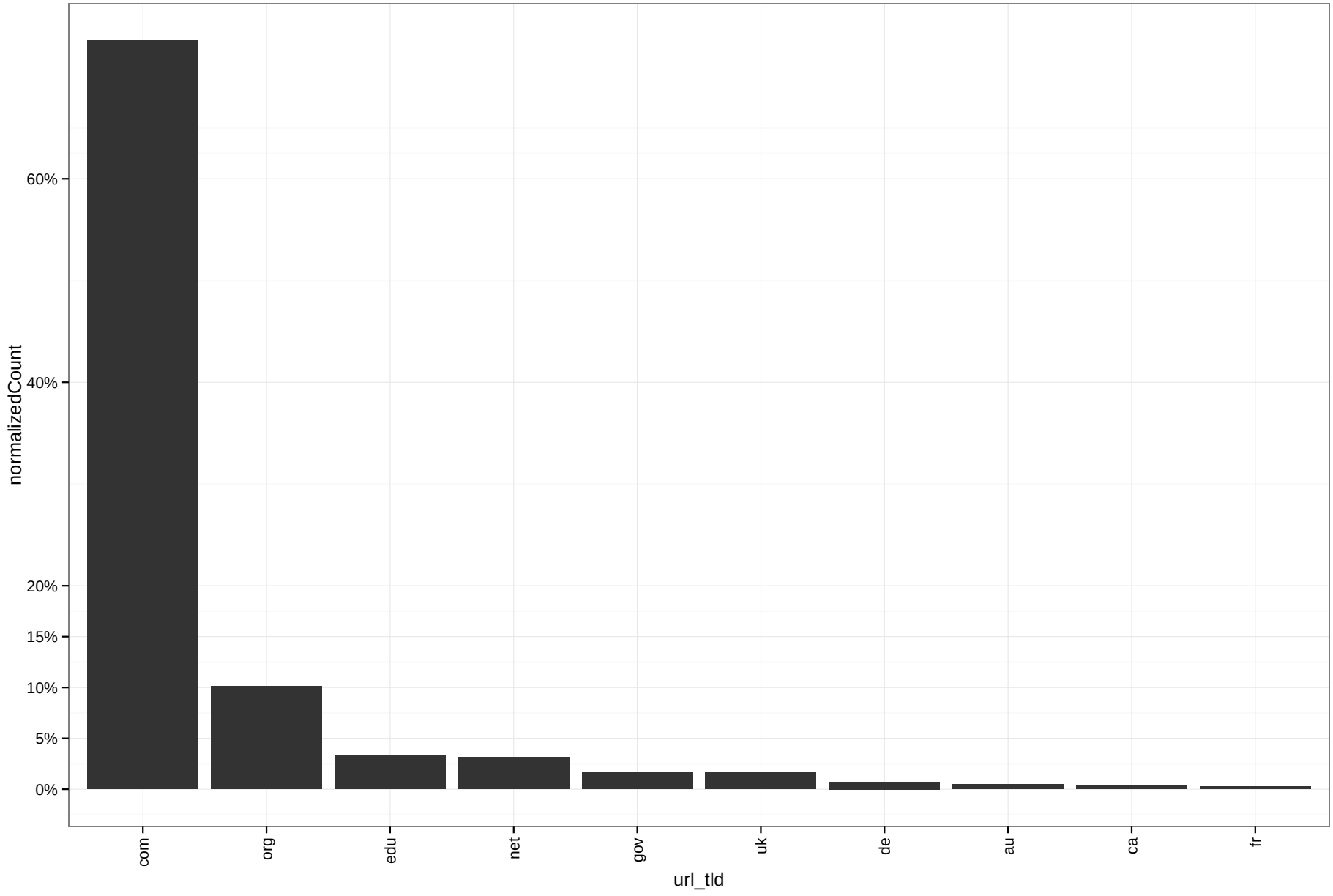


[3] https://github.com/klemens/ALI-CC

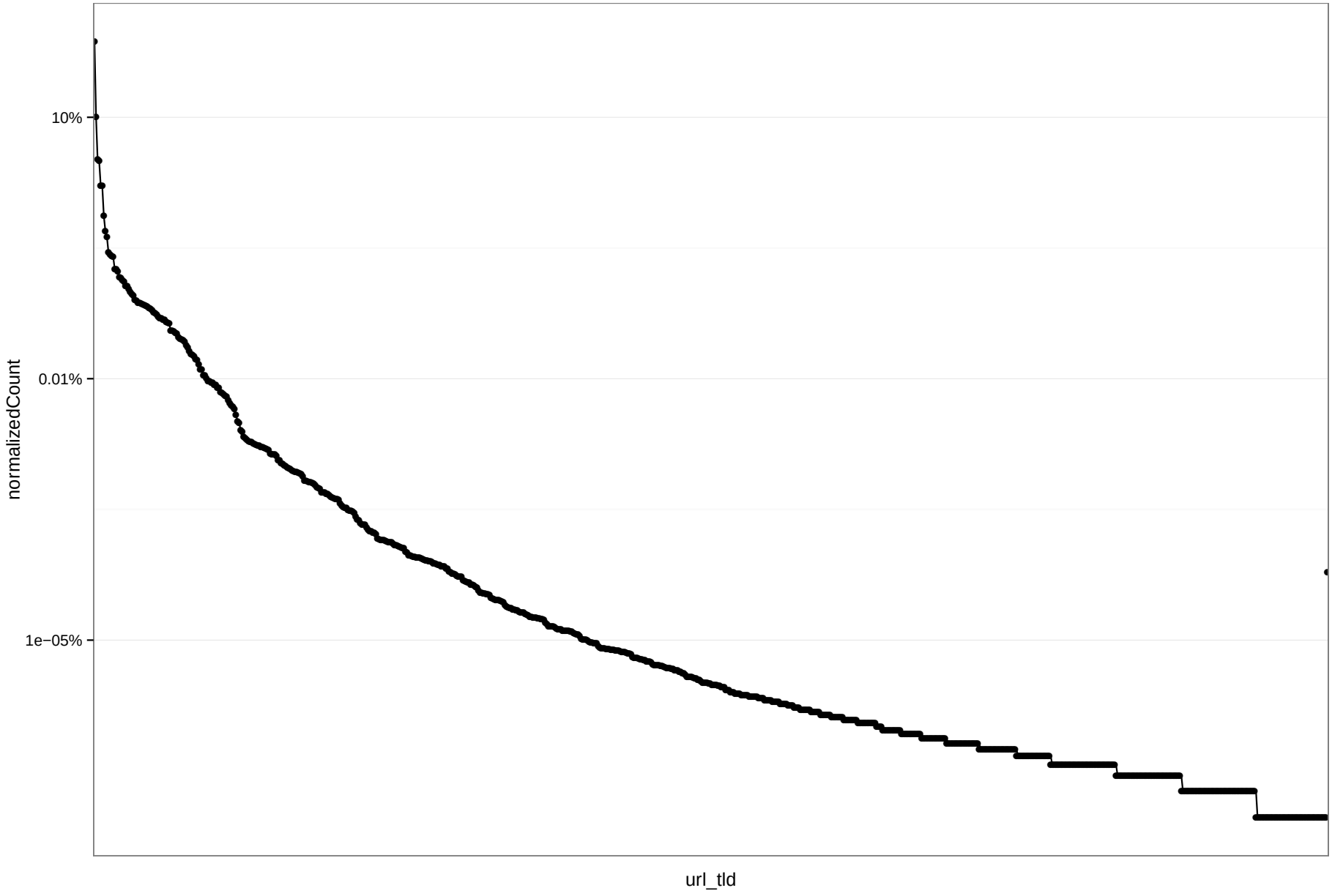
Analyse: Daten

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 von CDN	bool	CDN ja/nein
Interne Links	uint16	Anzahl relative Links + gleicher Host
Externe Links	uint16	Anzahl ausgehende Links

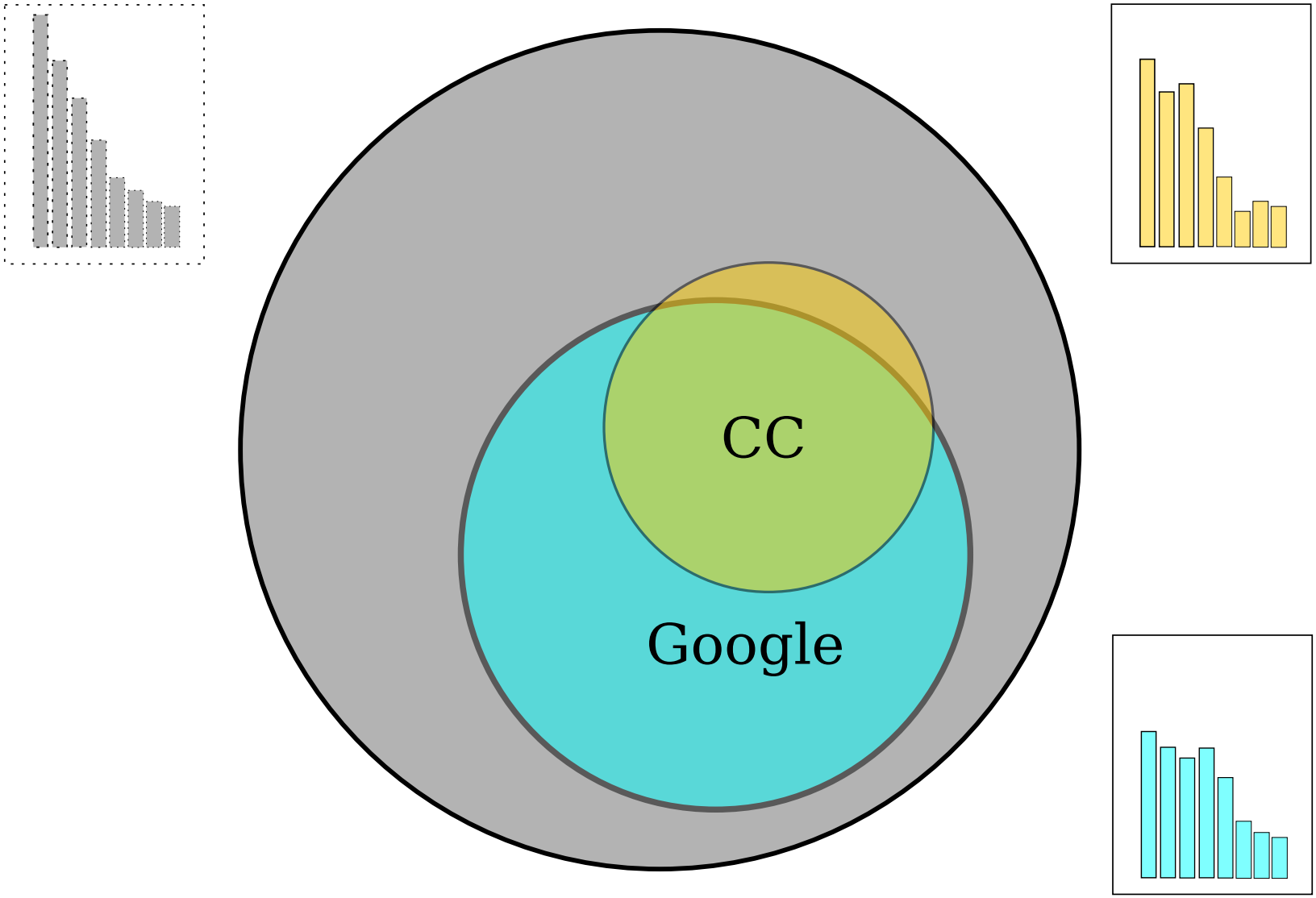
Ergebnisse: TLD



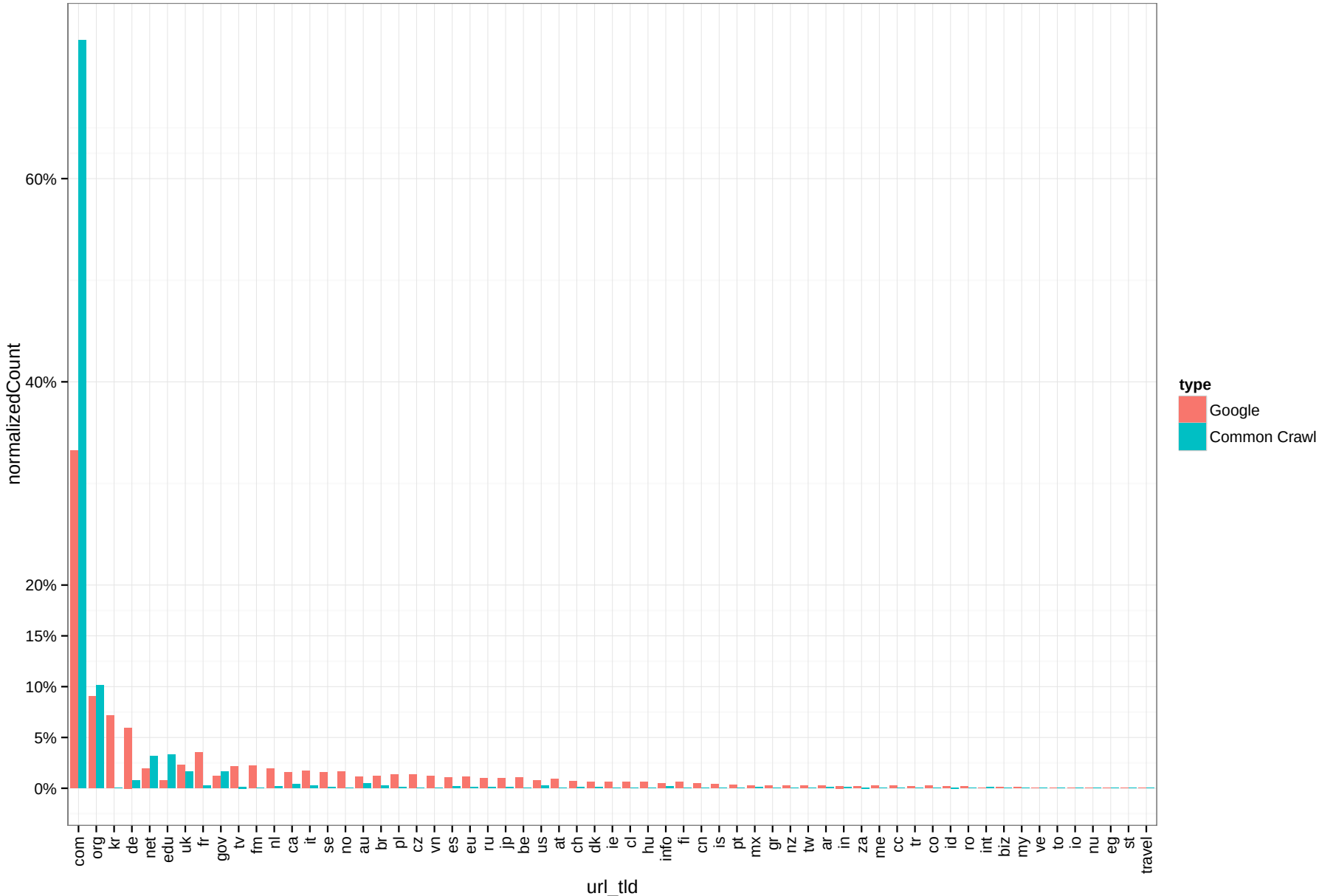
Ergebnisse: TLD



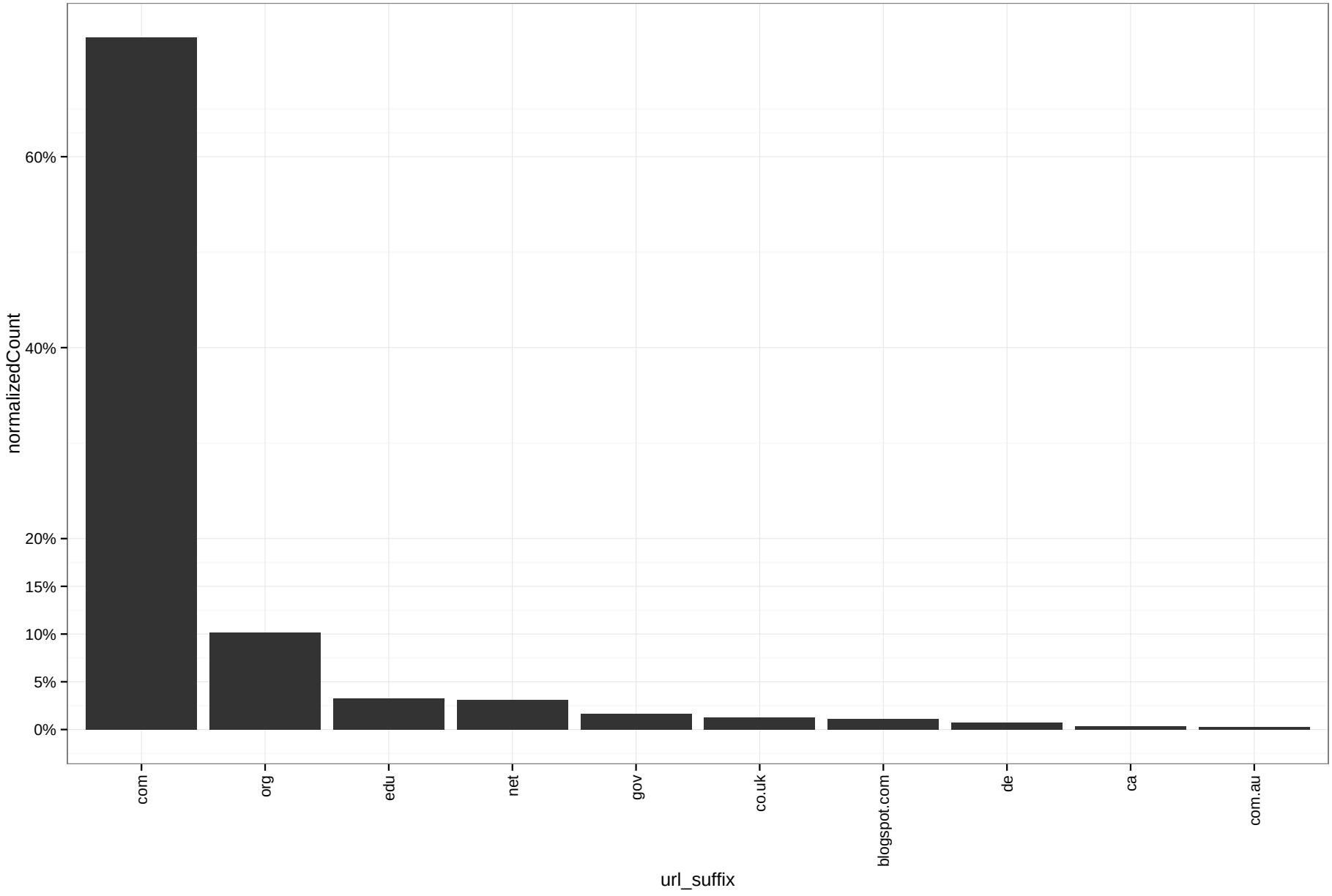
Ergebnisse: TLD



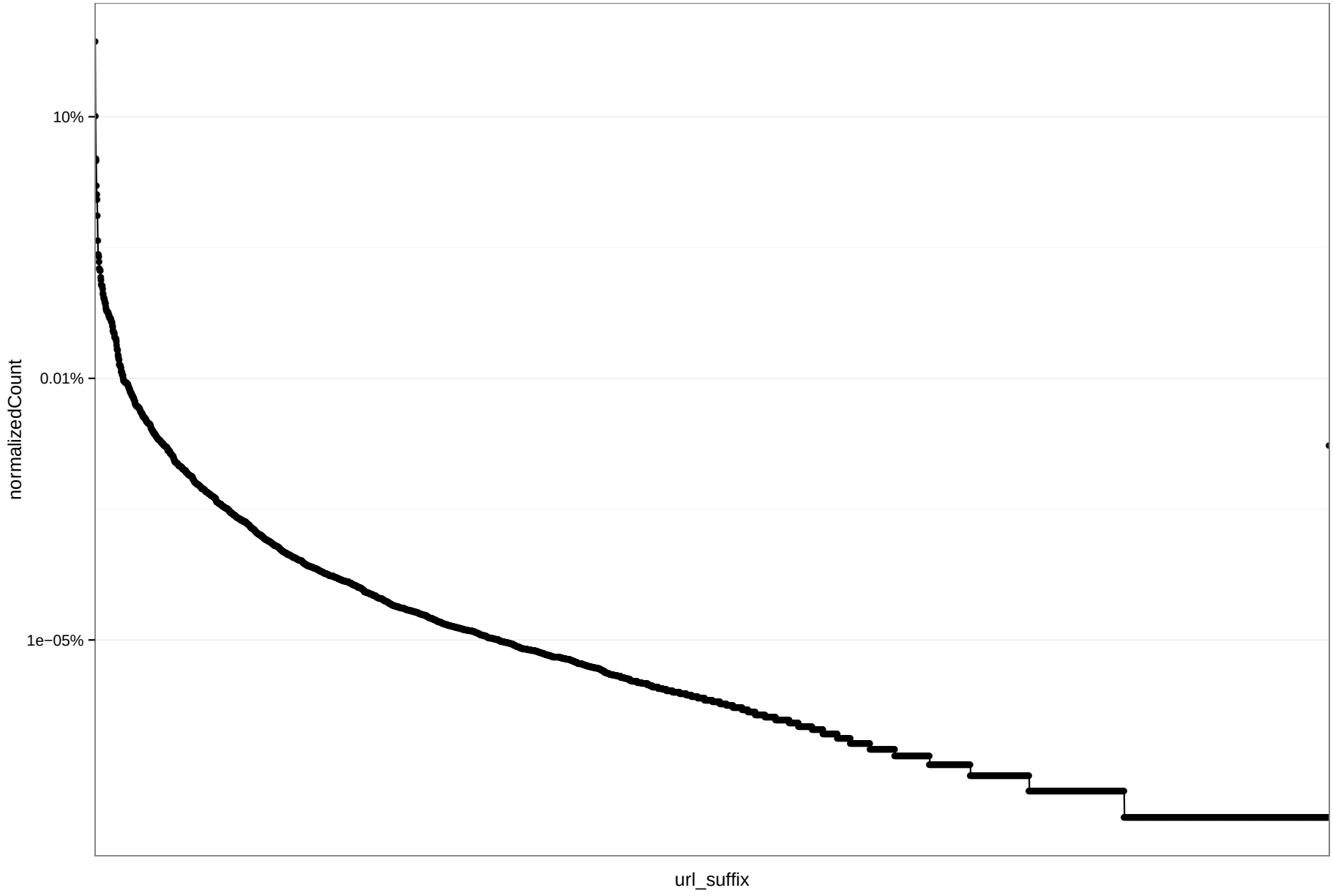
Ergebnisse: TLD



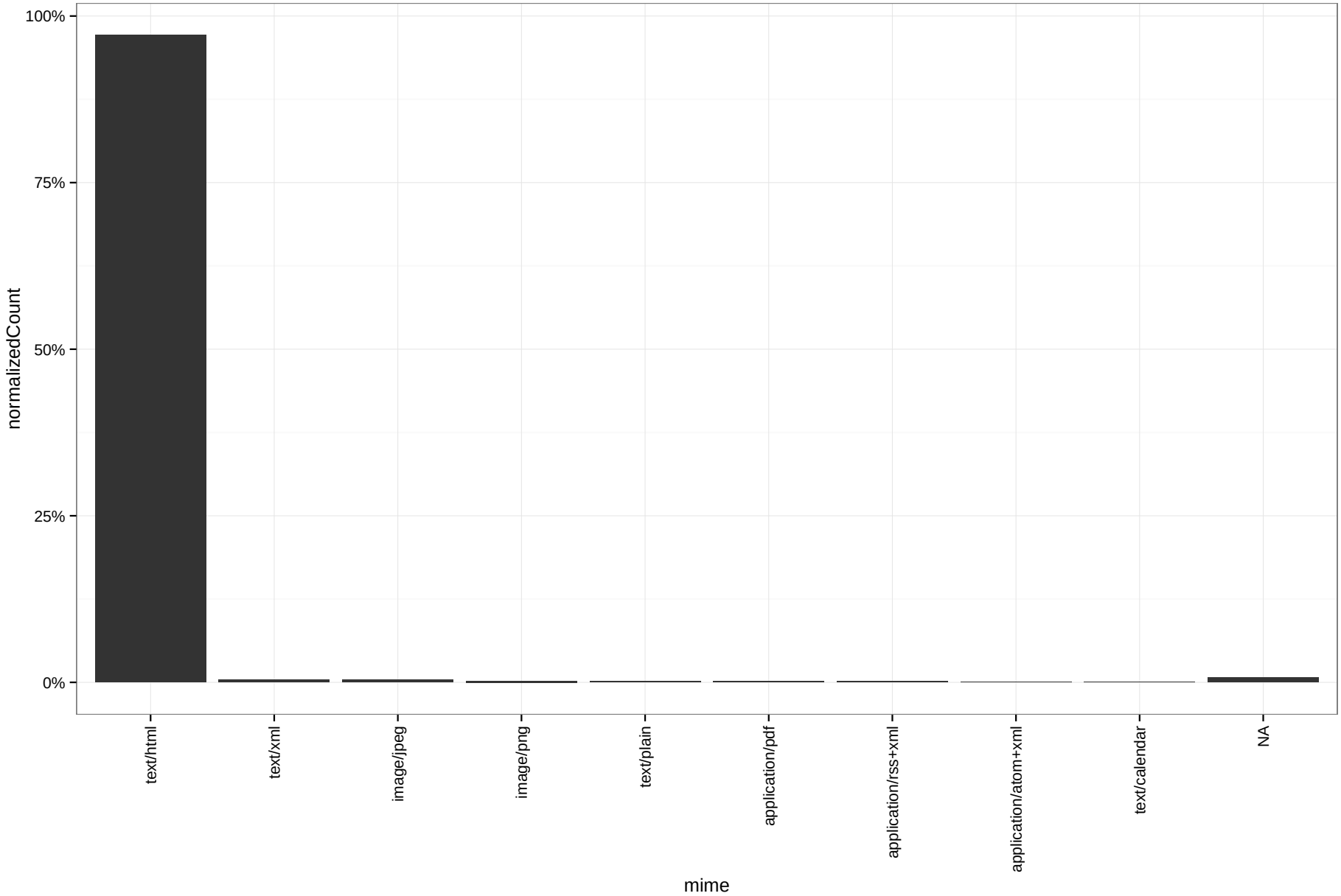
Ergebnisse: Public Suffix



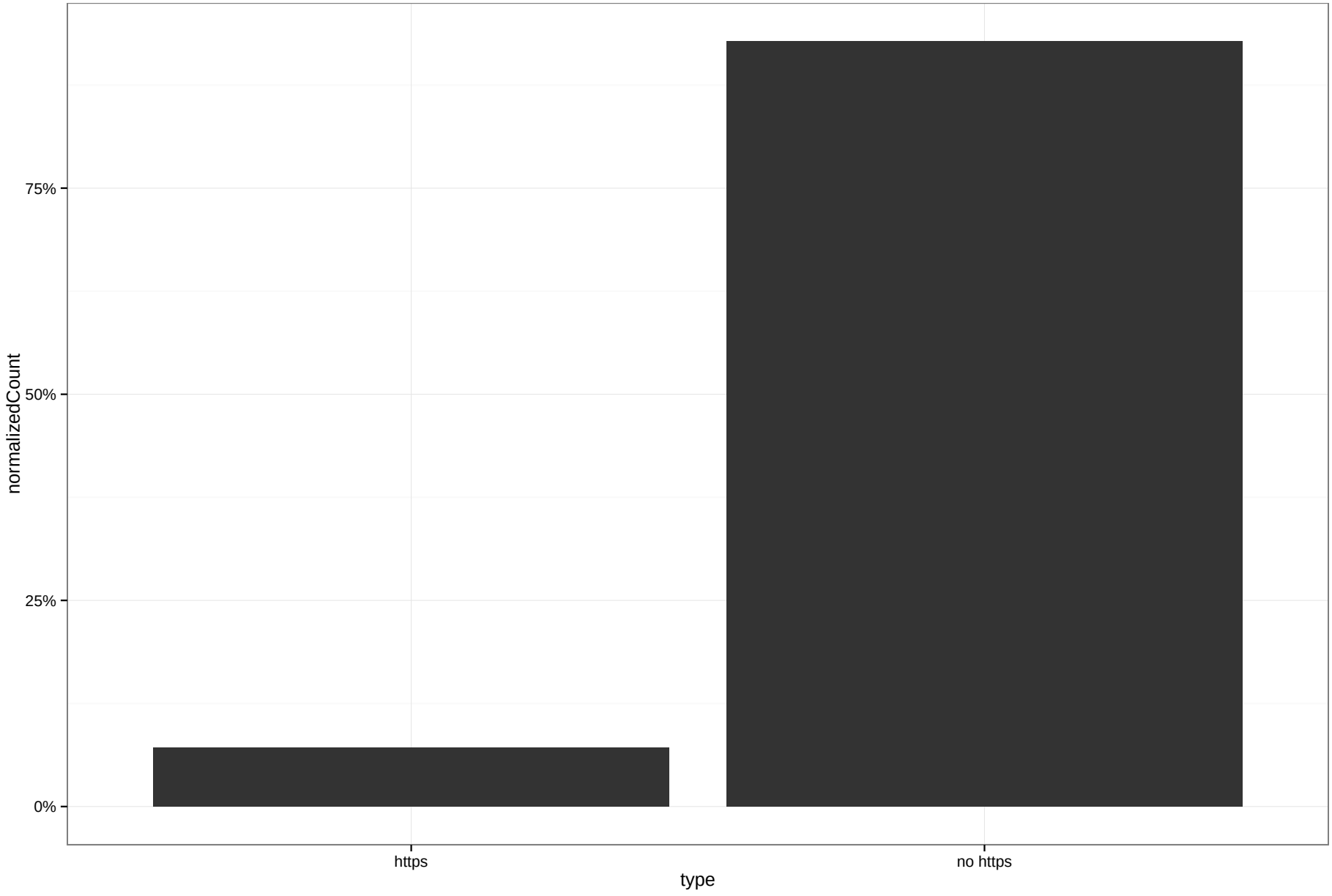
Ergebnisse: Public Suffix



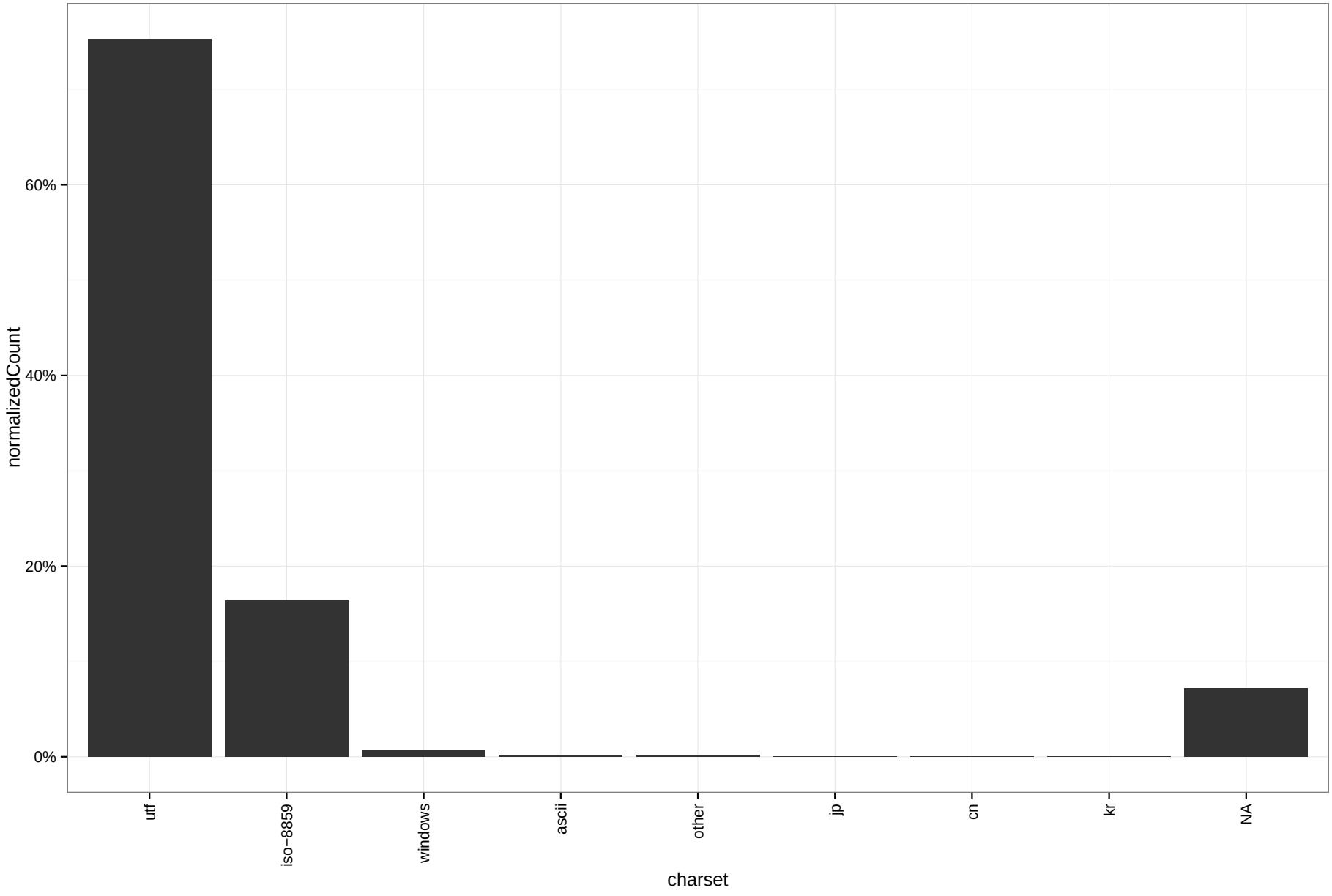
Ergebnisse: Mime-Type



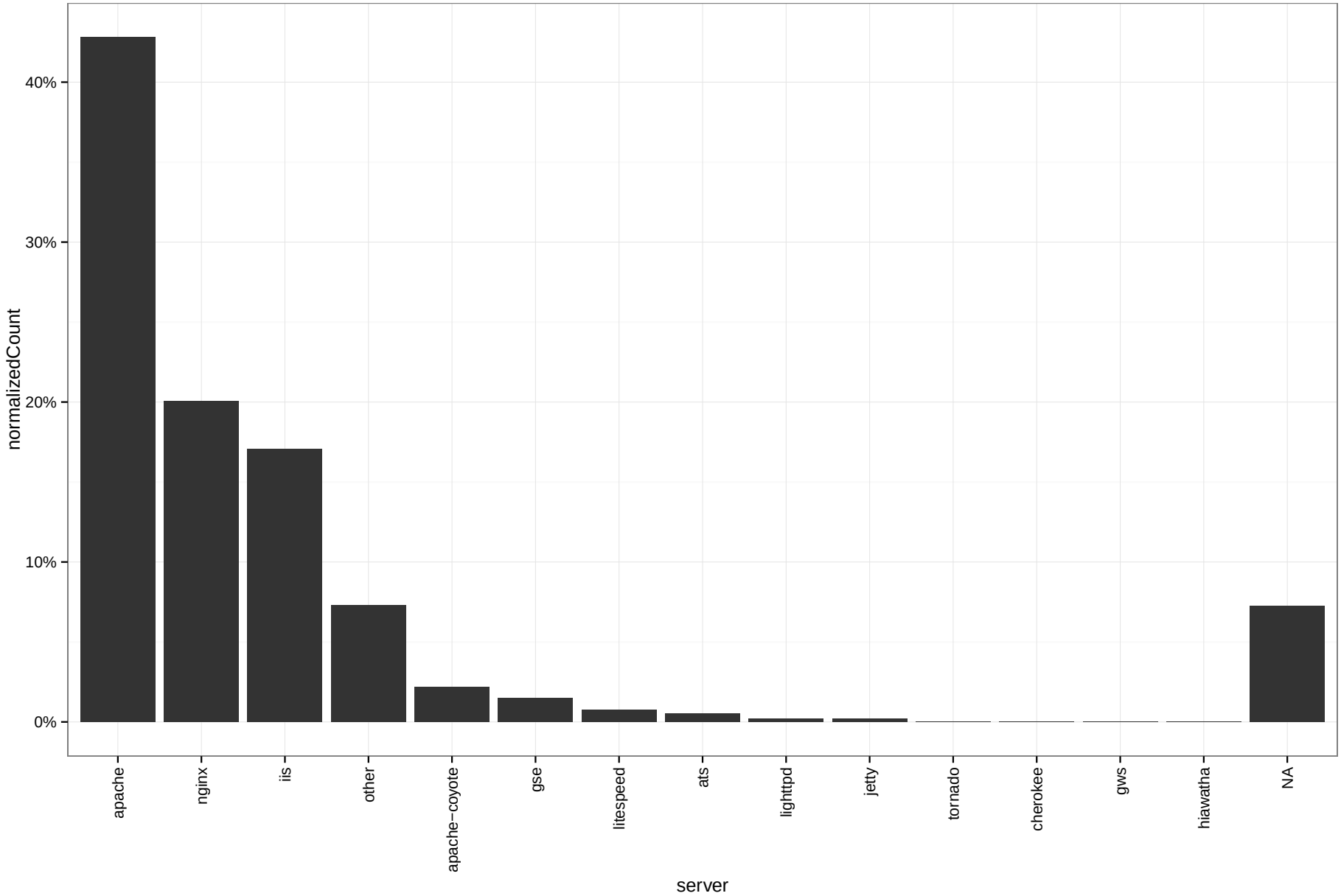
Ergebnisse: TLS



Ergebnisse: Encoding



Ergebnisse: Server



Textextraktion: Original

theguardian
Winner of the Pulitzer prize

home › US | world | opinion | sports | soccer | tech | arts | lifestyle | fashion | bus | all

The NSA files Glenn Greenwald on security and liberty

Edward Snowden: the whistleblower behind the NSA surveillance revelations

Glenn Greenwald, Ewen MacAskill and Laura Poitras in Hong Kong
Tuesday 11 June 2013 09.00 EDT

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.

Snowden will go down in history as one of America's most consequential whistleblowers,

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

Textextraktion: Common Crawl

Edward Snowden: the whistleblower behind the NSA surveillance revelations | US news | The Guardian

Textextraktion: 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.
(...)

Textextraktion: Mozilla 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.
(...)

Textextraktion: Common Crawl

Lance Armstrong, Tour de France champ, retires — abc7.com
GO
Personalize your weather by entering a location.
Sorry, but the location you entered was not found. Please try again.
Sections
Traffic
Video
Los AngelesOrange CountyInland EmpireVentura CountyCalifornia
Home
Accuweather
Traffic
Video
Photos
Mobile Apps
Local News Los AngelesOrange CountyInland EmpireVentura CountyCalifornia
Map My News
Shows Live Well Network
Follow Us
BREAKING NEWS
(...)

`http://abc7.com/archive/7962461`

Textextraktion: jWarcEx

Personalize your weather by entering a location.
Sorry, but the location you entered was not found. Please try again.

Sections Traffic Video Los AngelesOrange CountyInland EmpireVentura CountyCalifornia
Home Accuweather Traffic Video Photos Mobile Apps
Los AngelesOrange CountyInland EmpireVentura CountyCalifornia
BREAKING NEWS ABC shows live and on-demand – Download the WATCH ABC app!

Lance Armstrong, Tour de France champ, retires
Seven-time Tour de France champion Lance Armstrong is retiring. He ends his career amidst a federal doping investigation.
February 16, 2011 12:00:00 AM PST
Seven-time Tour de France champion Lance Armstrong said he's retiring from the professional cycling circuit, this time for good. The 39-year-old is calling it "Retirement 2.0" and is making it clear there is no reset button this time.

He says he wants to spend more time with his children and dedicate himself even more to the fight against cancer with his "Livestrong"Foundation.
(...)

`http://abc7.com/archive/7962461`