

Automatische Erkennung problematischer Webseiten zur Webanalyse

Fabienne Hubricht

Gutachter: Prof. Dr. Benno Stein und Prof. Dr. Ing Volker Rodehorst

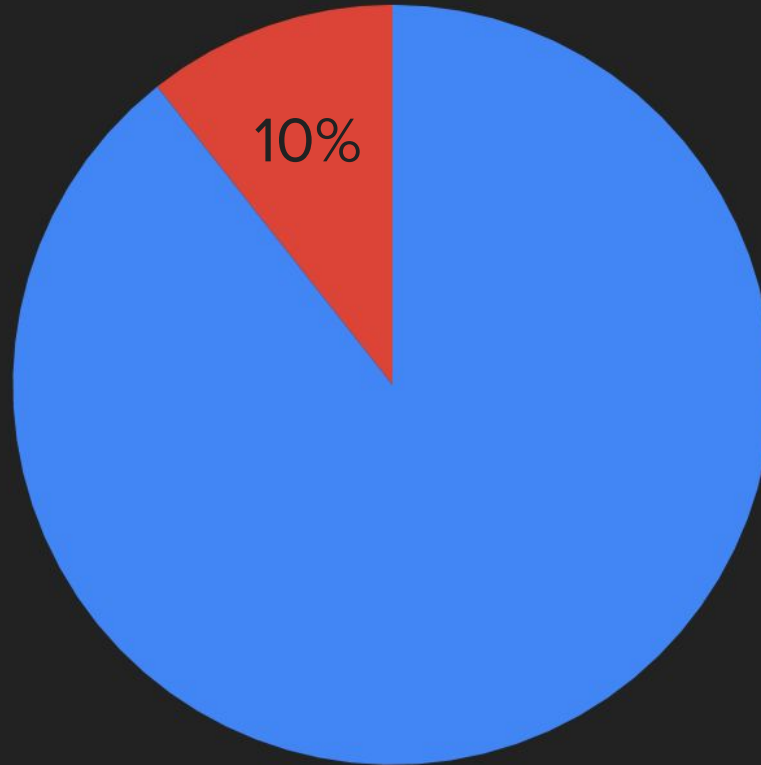
Betreuer: Johannes Kiesel

Webis-Web-Archive-17



10000 Webseiten

Webis-Web-Archive-17



gemeinnützige Organisationen

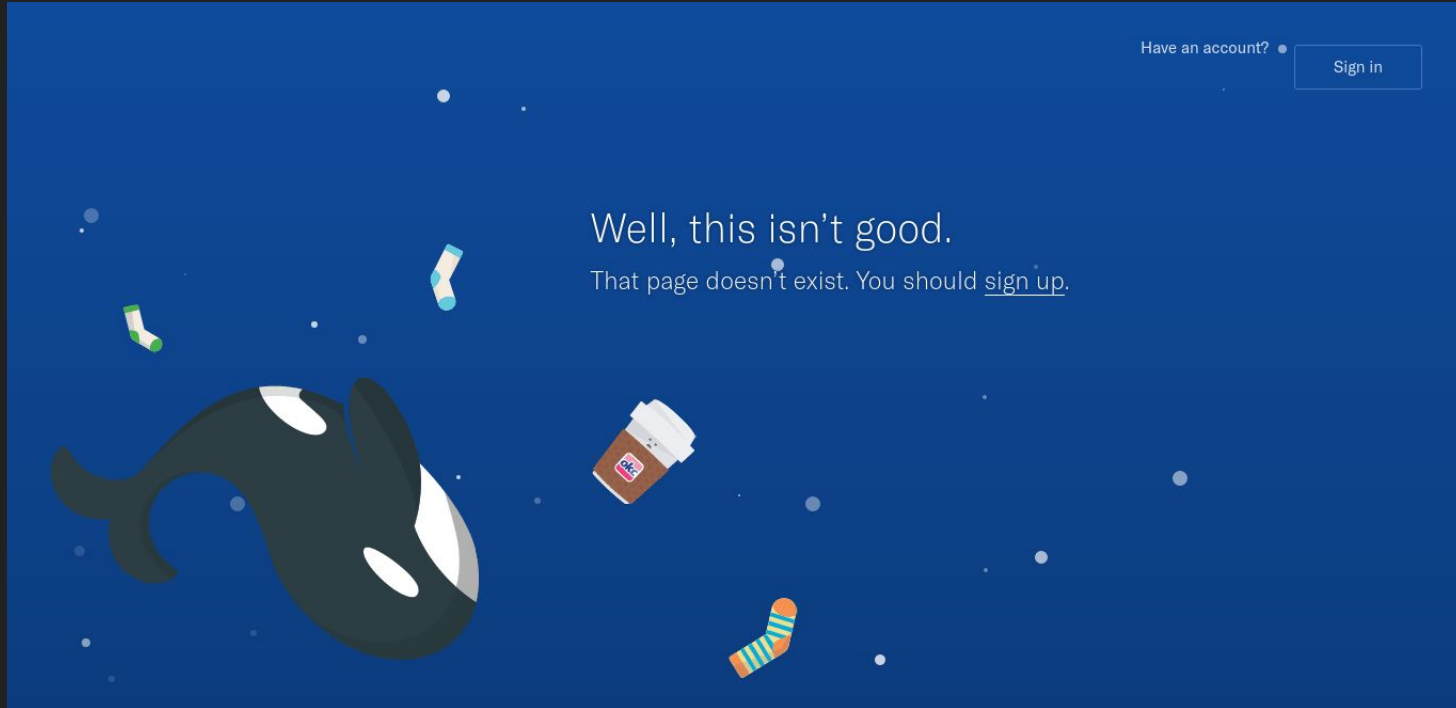
- Common Crawl
 - seit 2011 monatliches crawlen von Webseiten
 - April 2019 2.5 Milliarden Webseiten gecrawlt
- Internet Archive
 - seit 1996 349 Milliarden Webseiten gespeichert
 - Webseiten, digitalisierte Bücher, Videos, Bilder in digitaler Bibliothek für Öffentlichkeit zugänglich

Common Crawl

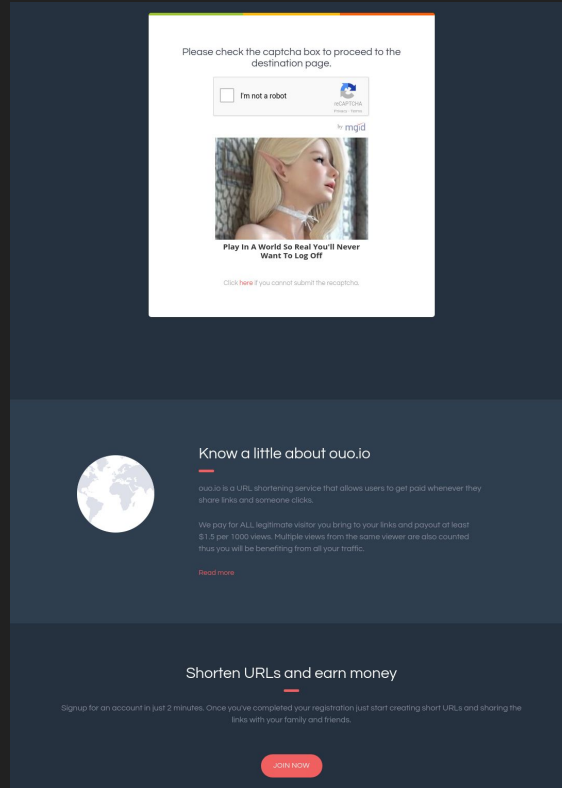


<http://library2020.blog.archive.org/>
<https://www.bellingcat.com/resources/2015/08/13/using-python-to-mine-common-crawl/>
<http://commoncrawl.org/connect/blog/>

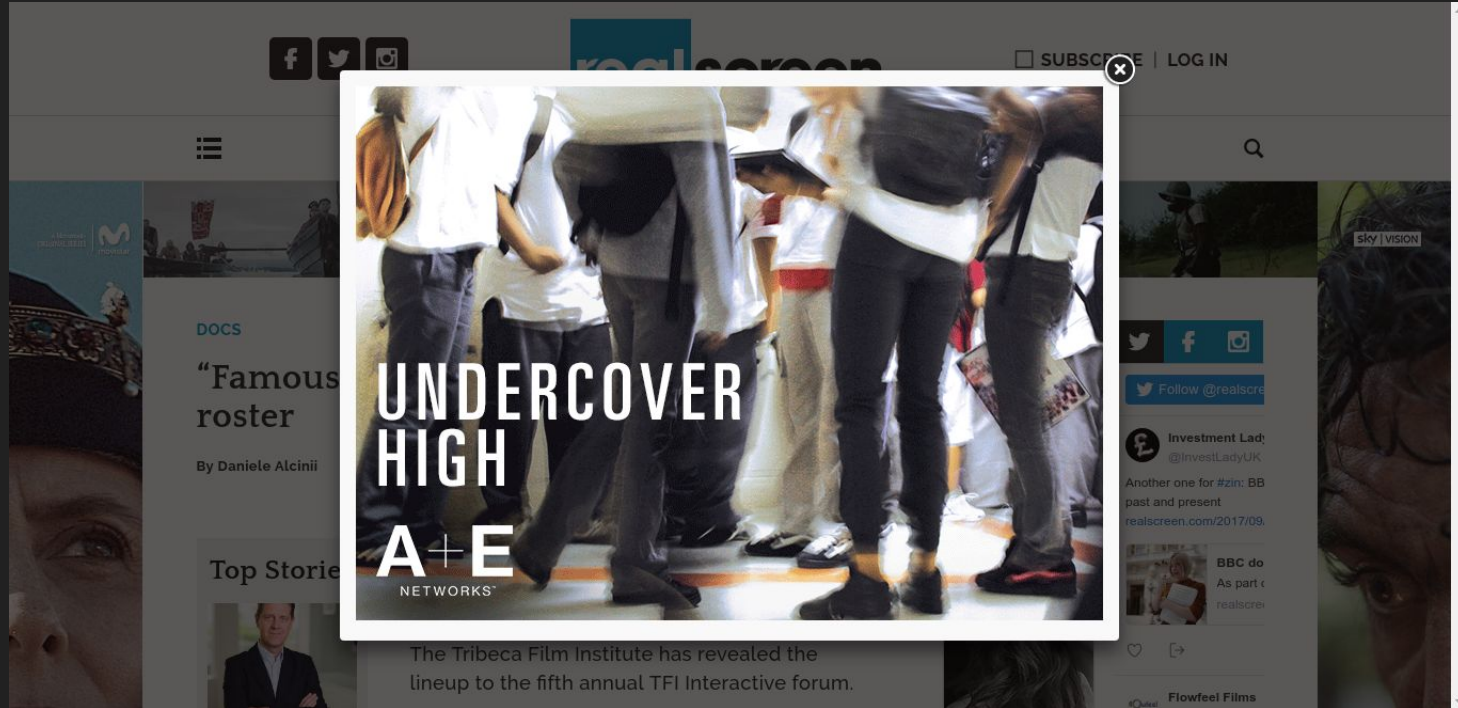
Problematische Webseiten



Problematische Webseiten



Problematische Webseiten

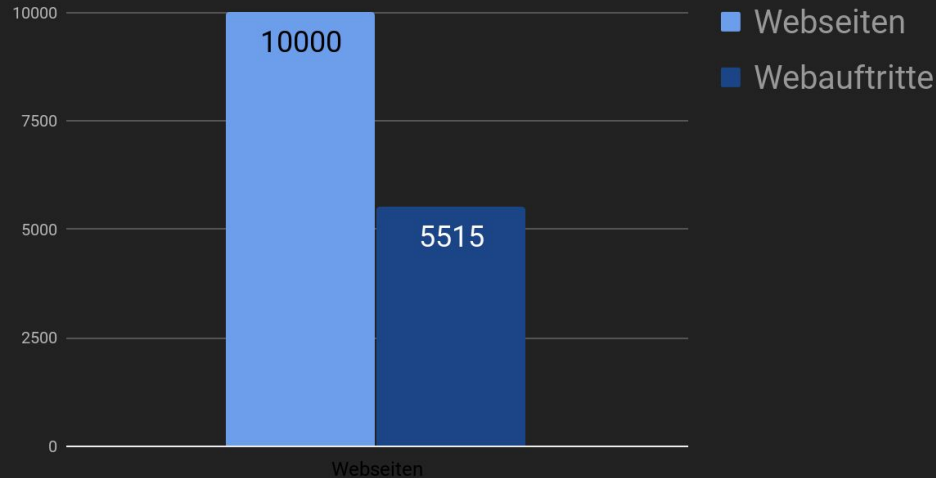


Problematische Webseiten



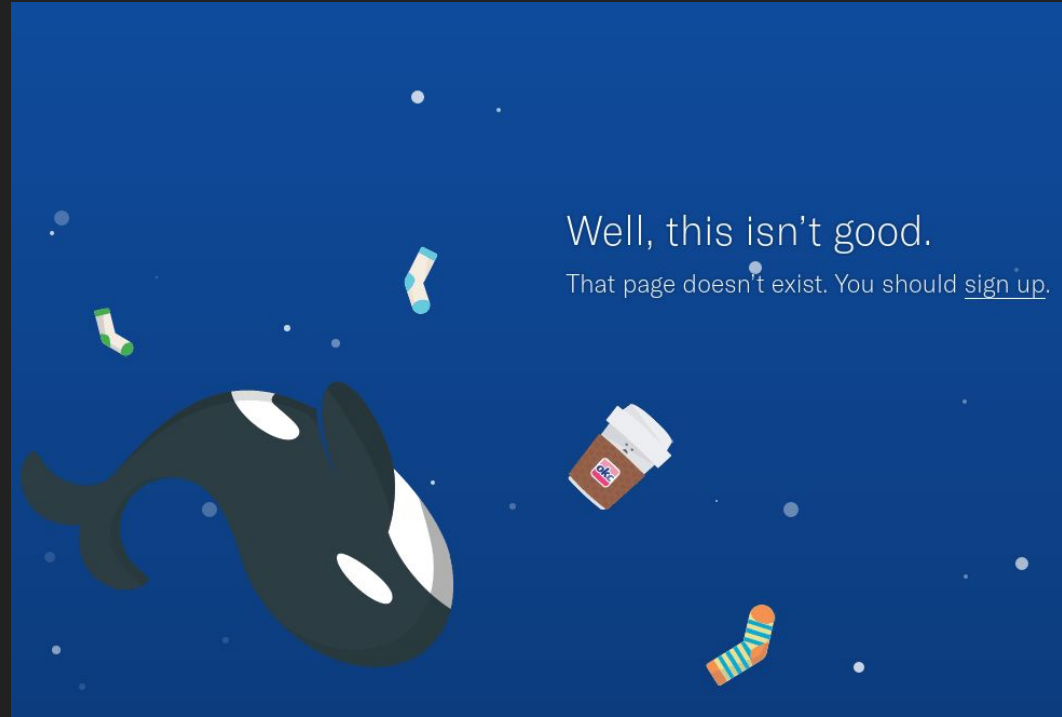
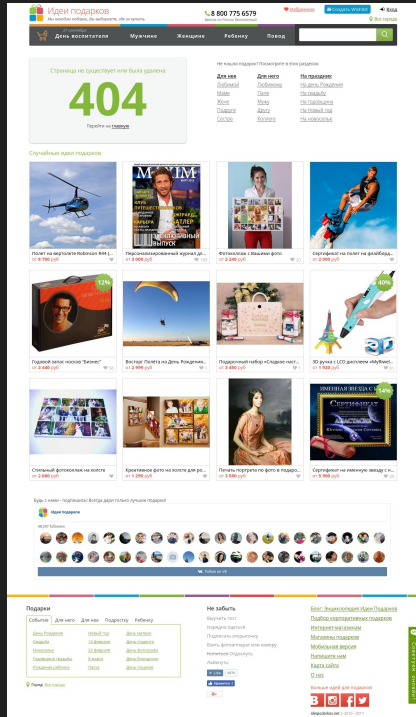
Webis-Web-Archive-17

Webis-Web-Archive-17




beinhaltet Screenshots, .html Dateien
und .warc Dateien

Differenzierung der Beeinträchtigung

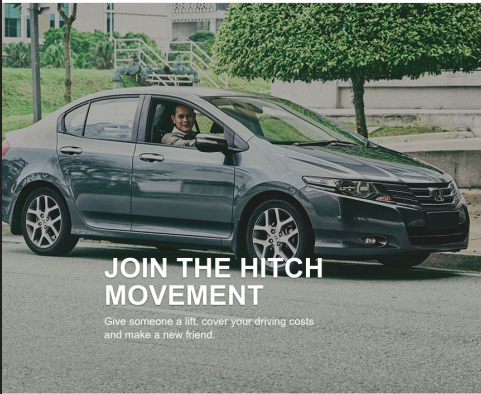


Differenzierung der Beeinträchtigung

 Singapore ▾ Services ▾ Corporate ▾

Hitch Driver ▾

Help Centre



JOIN THE HITCH MOVEMENT

Give someone a lift, cover your driving costs and make a new friend.

JOIN THE GRABHITCH DRIVER COMMUNITY!

[SIGN UP NOW!](#)

OR

Leave your particulars below and we will get in touch in 1 – 2 working days!

First Name

Last Name


+65 Phone Number

Email Address

How did you hear about Grab?

▾


☐ I'm not a robot




[I'M INTERESTED!](#)

By providing, I agree that you can collect, use and disclose the information provided by me in accordance with your [Privacy Policy](#) which I have read and understand.

What is GrabHitch?




GrabHitch is a social carpooling service that enables regular drivers like you and me to give fellow commuters a lift, meet new people, and cover the costs of owning a car.





To make it truly hassle-free, GrabHitch accepts both GrabPay and cash payments so GrabHitch drivers can choose their preferred method for cash out. All GrabHitch driver registration and bookings can be done through the regular Grab Passenger App, and do not require commercial insurance or Z10 registration.

Please check the captcha box to proceed to the destination page.

☐ I'm not a robot








Play In A World So Real You'll Never Want To Log Off

[Click here if you cannot submit the captcha.](#)



Know a little about ouo.io

ouo.io is a URL shortening service that allows users to get paid whenever they share links and someone clicks.

We pay for ALL legitimate visitor you bring to your links and payout of least \$1.5 per 1000 views. Multiple views from the same viewer are also counted thus you will be benefiting from all your traffic.

[Read more](#)

Shorten URLs and earn money

Signup for an account in just 2 minutes. Once you've completed your registration just start creating short URLs and sharing the links with your family and friends.

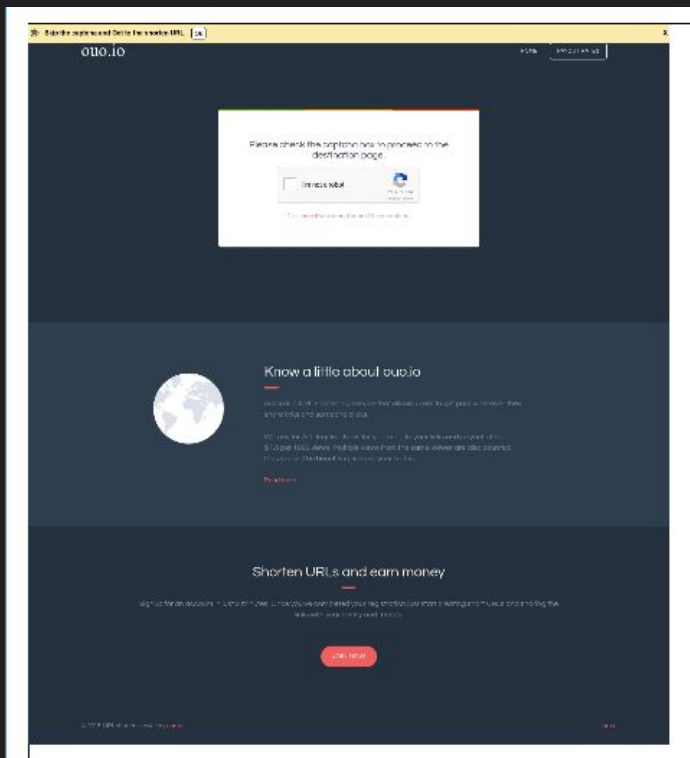
[JOIN NOW](#)

Differenzierung der Beeinträchtigung

The screenshot shows the Twitch website interface. The top navigation bar includes the Twitch logo, links for 'Browse', 'Get Desktop', 'Try Prime', and a search bar. The left sidebar features 'Followed Channels' with icons for 'frozenballz Warframe', 'CapcomFighters Street Fighter V', and 'Twitch Talk Shows'. Below this is the 'Online Friends' section. The main content area displays the channel page for 'SNoWMaN_39', with tabs for 'Videos 1', 'Clips', 'Collections', and 'Events'. Under the 'ALL VIDEOS' section, there is a video player showing a game and a list of recent videos, including one by 'goat' from 17 hours ago. At the bottom, there are three notification banners: 'TwitchCon 2017 tickets on sale now!', 'We've updated our privacy policy. Learn more.', and 'This site uses cookies. By continuing to browse the site, you are agreeing to our use of cookies. Review our Cookie Policy for more details.'

The screenshot shows the SEMBLY website. The top navigation bar includes the SEMBLY logo and a search bar. The main content area features a large banner for 'Take Your First Workshop - At No Cost'. The banner includes a form with fields for 'Email*' and 'Where would you like to learn?' (a dropdown menu). Below the form is a red 'Subscribe to Save' button. To the right of the form is a video player showing a woman working at a desk. Below the banner, there is a section titled 'SEMBLY is a pioneer in education and career' with a description of the company's mission. To the right of this section are two columns: 'Company' and 'Community'. The 'Company' column includes links for 'About', 'Locations', 'Blog', 'Careers', 'Contact', 'FAQ', and 'Press'. The 'Community' column includes links for 'Become an Instructor', 'Corporate Digital Training', 'Find a Scholarship', 'Looking for Talent?', and 'Develop 21st Century Cro...'. At the bottom, there is a red 'Sign Up Now' button.

Crowdsourcing Annotation



Select the answers that best describe the web page image on the left.

From the image, would you say the web page is...	yes	no
Mostly in a language that you understand?	<input type="radio"/>	<input type="radio"/>
Mostly advertisement? (page consists mostly of ads)	<input type="radio"/>	<input type="radio"/>
Still loading? (e.g., showing progress or activity indicator)	<input type="radio"/>	<input type="radio"/>
Pornographic? (contains sexually explicit content)	<input type="radio"/>	<input type="radio"/>
Cut off? (you can not scroll to the page bottom)	<input type="radio"/>	<input type="radio"/>

In the image, how dominant are the...	not	a bit	very
Pop-Ups? (elements that cover other elements)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Captchas? (tests to prevent bots accessing/commenting/...)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Error messages? ("404", page not found, timed out, ...)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Guidelines:

not: not on this image or barely noticeable at all

a bit: clearly noticeable, but can be ignored (e.g., pop-ups with transparent overlays, pop-ups/error messages/Captchas near the page border)

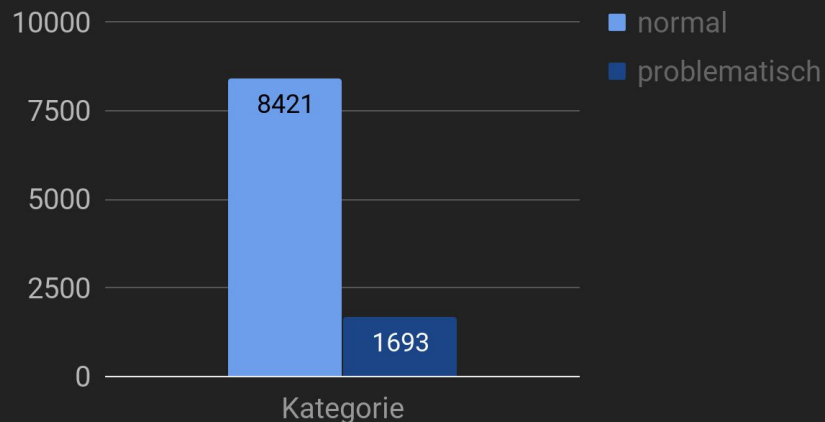
very: can not be ignored (pop-up or Captcha that requires attention before using the page, error message or Captcha instead of content)

Comments for this image (optional):

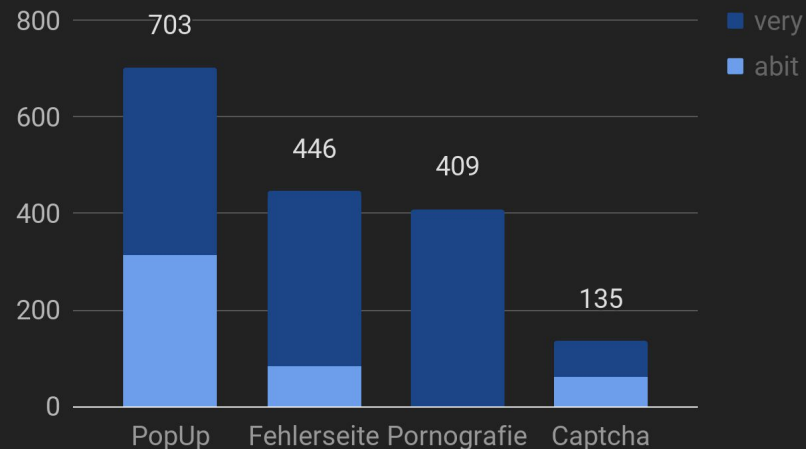
Hints: Zoom by clicking on the image. Use the spacebar to go the next image.

Ergebnisse der Crowdsourcing Annotation

Webis-Web-Archiv-17



Webis-Web-Archiv-17



Webseite

```
<html>
<head>
<title>Register |
PopCash.Net</title>
</head>
<body>
<div class="container
container--login">
</body>
</html>
```

.html Datei

A screenshot of a web registration form for 'Popcash The Popunder Network'. The form is titled 'Create a free account' and includes a link for users who already have an account. It contains several input fields: Username, Full Name, Email, Confirm Email, Password, Repeat password, Country (dropdown), Phone, IM Network (dropdown), and IM ID. There are checkboxes for 'I have read and agree with the Terms & Conditions' and 'I'm not a robot' (with a CAPTCHA image). A green 'Register' button is at the bottom.

Screenshot

```
WARC/1.0
WARC-Type: response
WARC-Record-ID:
<urn:uuid:373d6f3c-3e37-40ad-8cb0-4a87978d1fd8>
WARC-Date: 2017-09-22T18:03:04Z
WARC-Target-URI:
https://www.google.com/recaptcha/api.js?onload=onload
Callback&render=explicit
WARC-IP-Address: 172.217.16.164
Content-Type: application/http;msgtype=response
Content-Length: 956
WARC-Block-Digest:
sha1:69d0170c8a8d2830dc63d9dd050207b59034f8a6
WARC-Payload-Digest:
sha1:a0ff8ef9404db40aa2e944cd81812153b3c3b937
```

.warc Datei

Webseite



.html Datei

- title Tag
- id,
- class,
- source Attribute
- Webseiteninhalt



Screenshot

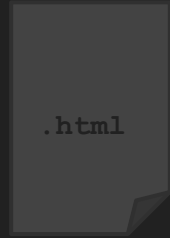
- Helligkeit
- RGB
- Bildererkennung von Captchas



.warc Datei

- Wortanzahl
Captcha

Webseite



.html Datei

- title Tag
- id,
- class,
- source Attribute
- Webseiteninhalt



Screenshot

- Helligkeit
- RGB
- Bilderkennung von Captchas

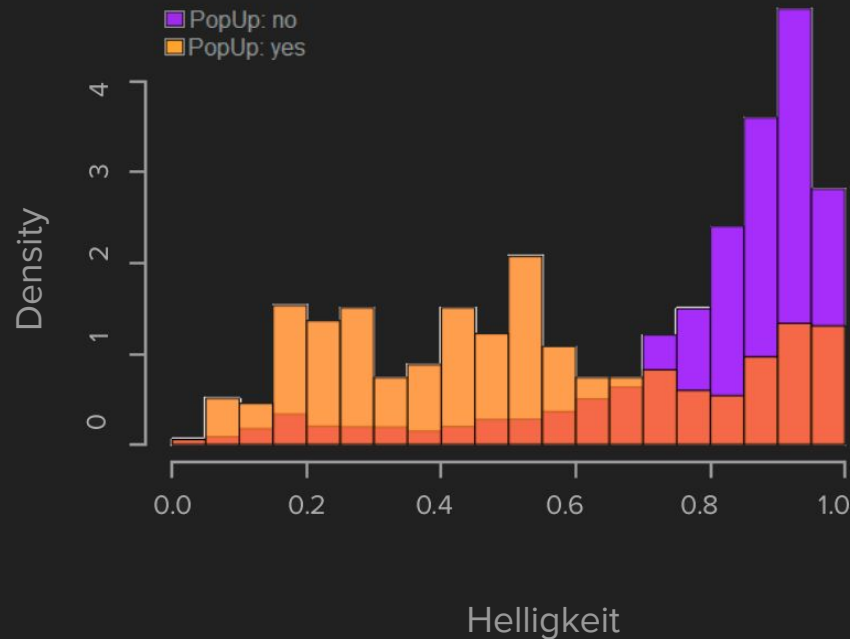
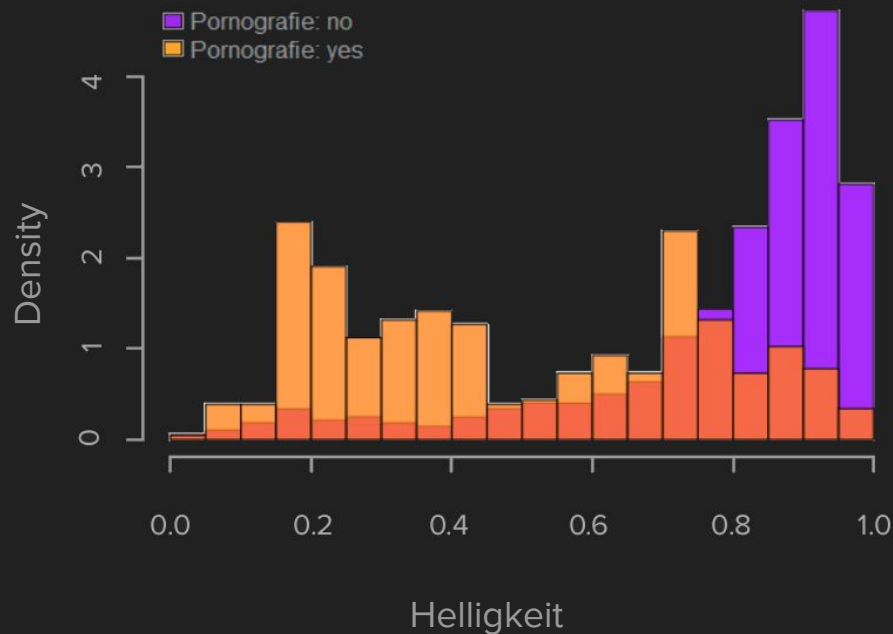


.warc Datei

- Wortanzahl
Captcha

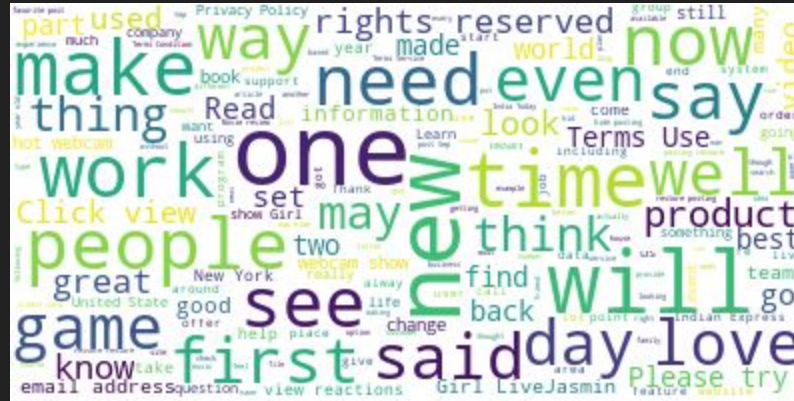


Merkmalsbasierte Lösungen

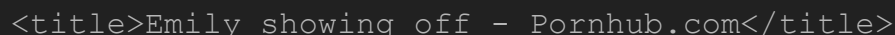
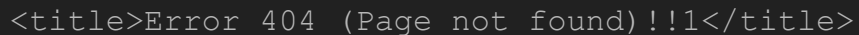


Merkmale basierte Lösungen

.html



.html



Merkmale basierte Lösungen

.html



Merkmal basierte Lösungen



WARC/1.0

WARC-Type: response

WARC-Record-ID: <urn:uuid:373d6f3c-3e37-40ad-8cb0-4a87978d1fd8>

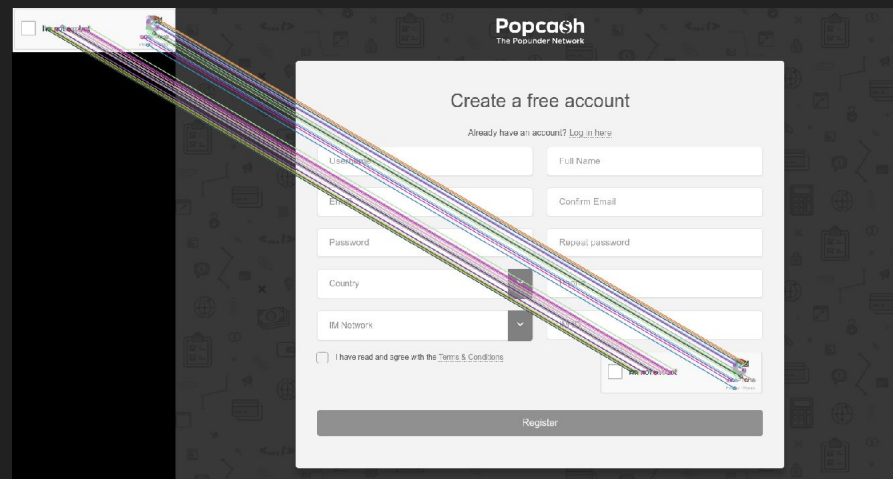
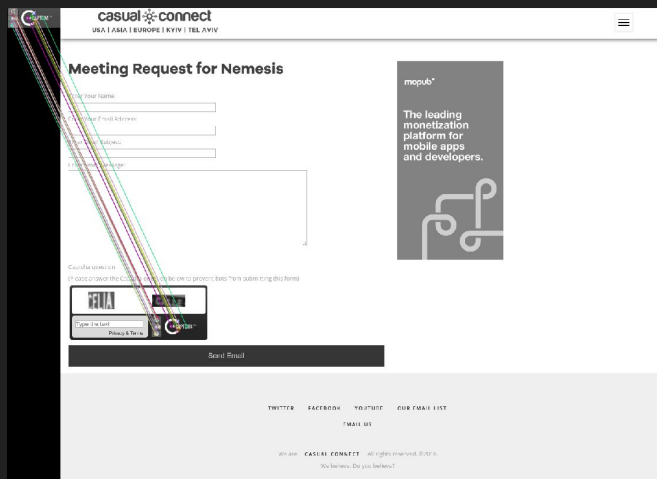
WARC-Date: 2017-09-22T18:03:04Z

WARC-Target-URI:

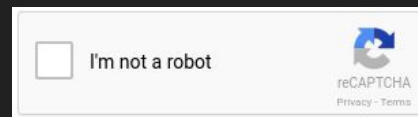
<https://www.google.com/recaptcha/api.js?onload=onloadCallback&render=explicit>



Bildererkennung von Captchas



Query Image

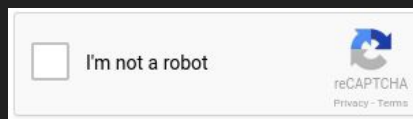
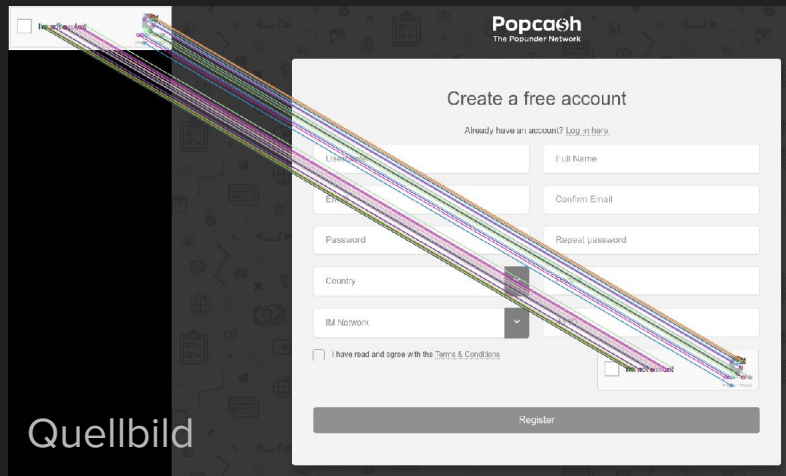


Query Image



Bildererkennung von Captchas

- Berechnung skalierungsinvarianter Schlüsselpunkte für Query Image und Quellbild
- Zuweisung von Deskriptoren um Schlüsselpunkte vergleichen zu können
- Ratio Test:
 - Vergleiche Abstand zwischen nearest neighbor und second nearest neighbor



Query Image



Bildererkennung von Captchas



Übersicht der Merkmale

Merkmal	Quelle	hilfreich für Erkennung von
Helligkeit, RGB	Screenshot	Pornografie, PopUp
Title Tag	Quellcode	Fehlerseiten, Pornografie
source Attribut	Quellcode	Captcha
id & class Attribut	Quellcode	PopUp
Webseiteninhalt	Quellcode	Fehlerseiten, Pornografie, Captcha
Feature Matching OpenCV	Screenshot	Captcha
Wortanzahl Captcha	.warc Datei	Captcha

Machine Learning

- supervised learning mit Machine Learning Framework Weka
- Klassifikator J48
 - Java Implementation des C4.5 Algorithmus
 - generieren von Entscheidungsbäumen
 - Erstellt aus Set von Trainingsdaten unter Verwendung des höchsten Information Gain
 - Attribut mit höchstem Information Gain trifft Entscheidung über Spaltung in kleinere Sets
- Beurteilt anhand der Maße Precision, Recall und F-Measure



<https://bit.ly/2YuPhjH>

Maße zur Bewertung des Klassifikators

$$\text{Precision} = \frac{\# \text{TruePositive}}{\# \text{TruePositive} + \# \text{FalsePositive}}$$

Anteil der als Captcha very klassifizierten Seiten, die tatsächlich Captcha very sind

$$\text{Recall} = \frac{\# \text{TruePositive}}{\# \text{TruePositive} + \# \text{FalseNegative}}$$

Von den Seiten die tatsächlich Captcha very sind, wie viele davon wurden als Captcha very klassifiziert

$$F_1\text{-Measure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

harmonisches Mittel aus Precision und Recall

Kategorie: Pornografie (Precision(P), Recall(R) und F-Measure(F) angegeben in Prozent)

	no			yes		
	P	R	F	P	R	F
Helligkeit, RGB	97,8	94,0	95,8	19,6	40,5	26,4
title Tag	98,0	98,2	98,1	47,3	45,7	46,5
Webseiteninhalt	98,3	97,7	98,0	45,6	53,4	49,2
alle Merkmale	98,3	98,5	98,4	56,0	52,6	54,2
naiver Klassifikator	96,5	100	98,2	03,5	100	06,7

Kategorie: Captcha (Precision(P), Recall(R) und F-Measure(F) angegeben in Prozent)

	not			a bit			very		
	P	R	F	P	R	F	P	R	F
source Attribute	99,4	98,2	98,8	09,1	06,3	07,4	10,5	46,2	17,1
Wortanzahl	99,6	98,5	99,0	11,5	18,8	14,3	17,9	53,8	26,9
OpenCV	99,6	97,5	98,5	08,5	43,8	14,3	11,8	15,4	13,3
alle Merkmale	99,6	99,5	99,6	21,1	25,0	22,9	53,8	53,8	53,8
naiver Klassifikator	99,3	100	99,6	0,5	100	0,9	0,4	100	0,8

Kategorie: PopUp (Precision(P), Recall(R) und F-Measure(F) angegeben in Prozent)

	not			a bit			very		
	P	R	F	P	R	F	P	R	F
Helligkeit, RGB	95,8	87,6	91,5	16,1	31,3	21,2	21,9	44,5	29,3
id-class Attribut	92,6	91,9	92,3	06,0	07,0	06,5	10,4	10,2	10,3
Webseiteninhalt	92,6	90,5	91,5	06,9	06,3	06,6	03,7	05,8	04,5
alle Merkmale	94,1	92,3	93,2	07,7	07,8	07,8	20,5	29,2	24,1
naiver Klassifikator	92,0	100	95,6	03,8	100	07,4	04,1	100	07,9

Fazit

Ergebnisse

- Erkennung der Kategorie Pornografie und Captcha very funktioniert am besten
- Erkennung der Kategorie PopUp funktioniert nicht gut
- Klassifikator in den Fällen (abit, very, yes) besser als naiver Klassifikator

Future Work

- Erkennung von anderen Arten von Captchas und PopUps
- größerer Datensatz
- Overfitting berücksichtigen, Post Pruning des Entscheidungsbaums

Fazit

Ergebnisse

- Erkennung der Kategorie Pornografie und Captcha very funktioniert am besten
- Erkennung der Kategorie PopUp funktioniert nicht gut
- Klassifikator in den Fällen (abit, very, yes) besser als naiver Klassifikator

Future Work

- Erkennung von anderen Arten von Captchas und PopUps
- größerer Datensatz
- Overfitting berücksichtigen, Post Pruning des Entscheidungsbaums

Ich danke für Ihre Aufmerksamkeit

Kategorie: Pornografie (Precision(P), Recall(R) und F-Measure(F) angegeben in Prozent)

	no			yes		
	P	R	F	P	R	F
Helligkeit, RGB	97,8	94,0	95,8	19,6	40,5	26,4
id,class Attribute	97,7	97,2	97,5	32,6	37,9	35,1
source Attribut	97,9	98,4	98,2	48,5	41,4	44,7
title Tag	98,0	98,2	98,1	47,3	45,7	46,5
Webseiteninhalt	98,3	97,7	98,0	45,6	53,4	49,2
alle Merkmale	98,3	98,5	98,4	56,0	52,6	54,2
naiver Klassifikator	96,5	100	98,2	03,5	100	06,7

Kategorie: Captcha (Precision(P), Recall(R) und F-Measure(F) angegeben in Prozent)

	not			a bit			very		
	P	R	F	P	R	F	P	R	F
Helligkeit, RGB	99,2	96,9	98,0	00,0	00,0	00,0	02,7	07,7	04,0
id-class Attribut	99,5	98,3	98,9	06,8	18,8	10,0	30,8	61,5	41,0
source Attribut	99,4	98,2	98,8	09,1	06,3	07,4	10,5	46,2	17,1
title Tag	99,3	96,3	97,8	00,0	00,0	00,0	08,6	46,2	14,5
Webseiteninhalt	99,2	97,6	98,4	00,0	00,0	00,0	06,5	23,1	10,2
Wortanzahl	99,6	98,5	99,0	11,5	18,8	14,3	17,9	53,8	26,9
OpenCV	99,6	97,5	98,5	08,5	43,8	14,3	11,8	15,4	13,3
alle Merkmale	99,6	99,5	99,6	21,1	25,0	22,9	53,8	53,8	53,8
naiver Klassifikator	99,3	100	99,6	0,5	100	0,9	0,4	100	0,8

Kategorie: Fehlerseiten (Precision(P), Recall(R) und F-Measure(F) angegeben in Prozent)

	not			a bit			very		
	P	R	F	P	R	F	P	R	F
Helligkeit, RGB	96,3	89,9	93,0	00,0	00,0	00,0	12,9	24,4	16,9
id-class Attribut	96,9	91,3	94,0	00,0	00,0	00,0	19,2	44,7	26,8
source Attribut	96,7	80,4	87,8	00,0	00,0	00,0	09,1	48,8	15,3
title Tag	96,9	95,3	96,1	00,0	00,0	00,0	38,2	40,7	39,4
Webseiteninhalt	97,1	88,1	92,4	00,0	00,0	00,0	13,8	44,7	21,1
alle Merkmale	96,9	93,3	95,1	00,0	00,0	00,0	20,4	37,4	26,4
naiver Klassifikator	95,5	100	97,7	0,8	100	01,5	03,7	100	07,2

Kategorie: PopUp (Precision(P), Recall(R) und F-Measure(F) angegeben in Prozent)

	not			a bit			very		
	P	R	F	P	R	F	P	R	F
Helligkeit, RGB	95,8	87,6	91,5	16,1	31,3	21,2	21,9	44,5	29,3
id-class	92,6	91,9	92,3	06,0	07,0	06,5	10,4	10,2	10,3
source	93,5	90,7	92,1	13,4	14,1	13,7	08,1	05,1	05,8
title Tag	92,5	94,5	93,3	04,7	03,1	03,8	06,8	05,1	05,8
Webseiteninhalt	92,6	90,5	91,5	06,9	06,3	06,6	03,7	05,8	04,5
alle Merkmale	94,1	92,3	93,2	07,7	07,8	07,8	20,5	29,2	24,1
naiver Klassifikator	92,0	100	95,6	03,8	100	07,4	04,1	100	07,9