Automatic Detection of Section Titles and Prose Text in Web Pages

Dataset Manual

Aug 24, 2018

Contents

Revision History

Version	Date	Name	Description
1	05/13/2018	Abhijith	Initial document
2	08/24/2018	Abhijith	Final version

1 Introduction

This document describes the dataset used in training and testing ASDUS. The entire dataset used is provided along woth this document.

2 Dataset Details

Our dataset consists of web documents (HTML) belonging to three different categories.

Category	Number of documents
Privacy policies	152
Terms of service agreements	100
Miscellaneous documents	51

Table 1: Distribution of web documents

2.1 Web Privacy Policies

A privacy policy is a statement or a legal document (in privacy law) that discloses some or all the ways a party gathers, uses, discloses, and manages a customer or client's data. It fulfills a legal requirement to protect a customer or client's privacy**.** We based the selection of policies on the below criteria:

- We collected 80 policies from Amazon Alexa's top 100 websites list for 2016. We had to discard 20 of the top 100 as they were duplicates (same parent company owning the website).

- We collected 72 policies from the top Google trends of 2017. We extracted the privacy policies of top two sites of each trend.

Below table contains the website names, URL and collection date of the privacy policies.

Sl.	Website	URL	Date
1	9gag.com	https://about.9gag.com/privacy/	7/18/2017
2	abs-cbn.com	https://www.abscbnmobile.com/privacy-policy	7/18/2017

3	activision.com	https://www.activision.com/legal/privacy-policy	2/2/2018
4	adf.ly	https://ay.gy/privacy	7/18/2017
5	adidas	https://www.adidas.com/us/what-is-the-privacy-policy.html	7/18/2017
6	adobe.com	https://www.adobe.com/privacy/policy.html	7/18/2017
7	aenetworks.com	http://www.aenetworks.com/privacy	2/2/2018
8	akc.org	http://www.akc.org/privacy/	2/2/2018
9	alibaba.com	https://rule.alibaba.com/rule/detail/2034.htm	7/18/2017
10	aliexpress.com	https://www.alibabagroup.com/en/global/privacy	7/18/2017
11	amazon.com	https://www.amazon.com/gp/help/customer/display.html?nodeId=468496	7/18/2017
12	americanmediainc.com	https://www.americanmediainc.com/privacy-policy	2/2/2018
13	amtrak	https://www.amtrak.com/privacy-policy	2/2/2018
14	aol.com	http://www.docracy.com/0x942pd1qk6/aol-com-privacy-policy-tos	2/2/2018
15	apartments.com	http://www.apartments.com/advertise/disclaimers/privacy-statement	2/2/2018
16	apple.com	https://www.apple.com/privacy/	7/18/2017
17	ask.com	https://www.docracy.com/9myp5dolva/ask-com-privacy-policy-tos	7/18/2017
18	atlanticrecords.com	http://www.atlanticrecords.com/privacy-policy	2/2/2018
19	att.com	http://about.att.com/sites/privacy_policy	2/2/2018
20	baidu	http://usa.baidu.com/privacy/	2/2/2018
21	bankofamerica.com	https://www.bankofamerica.com/privacy/overview.go	7/18/2017
22	battle.net	http://us.blizzard.com/company/about/privacy.html	7/18/2017
23	bbc.co.uk	https://www.bbc.co.uk/privacy/	7/18/2017
24	bet.com	https://www.bet.com/privacy-policy.html	2/2/2018
25	bet365.com	https://help.bet365.com/en/privacy-policy	7/18/2017
26	billboard.com	https://www.billboard.com/p/privacy-policy	2/2/2018
27	bitcoin.com	https://www.bitcoin.com/privacy-policy	2/2/2018
28	blackboard.com	http://www.blackboard.com/footer/privacy-policy.html	2/2/2018
29	blastingnews.com	https://www.blastingnews.com/privacy/	7/18/2017
30	bleacherreport.com	https://bleacherreport.com/pages/privacy	2/2/2018
31	blogspot.com	https://support.google.com/blogger/answer/42673?hl=en	7/18/2017
32	booking.com	https://www.booking.com/content/privacy.html	7/18/2017
33	buzzfeed.com	https://www.buzzfeed.com/about/privacy	7/18/2017
34	chase.com	https://www.chase.com/digital/resources/privacy-security/privacy/online-privacy-policy	7/18/2017
35	cnet.com	https://www.cnet.com/html/aboutcnet/editorial/privacy.html	7/18/2017
36	cnn.com	https://www.cnn.com/2014/01/17/cnn-info/privacy-policy/index.html	7/18/2017
37	cocacola	https://www.coca-colacompany.com/our-company/privacy-policy	2/2/2018
38	coindesk.com	https://www.coindesk.com/privacy-policy/	2/2/2018
39	commonsensemedia.org	https://www.commonsensemedia.org/about-us/our-mission/privacy-policy	2/2/2018
40	condenast.com	http://www.condenast.com/privacy-policy/	2/2/2018
41	conservativereview.com	https://www.conservativereview.com/privacy-policy/	2/2/2018

42	cutestat.com	https://privacypolicies.com.cutestat.com/	2/2/2018
43	dailymotion.com	https://www.dailymotion.com/legal/privacy	7/18/2017
44	delta	https://www.delta.com/content/www/en_US/privacy-and-security.html	2/2/2018
45	deviantart.com	https://about.deviantart.com/policy/privacy/	7/18/2017
46	diply.com	https://diply.com/static/privacy	7/18/2017
47	discordapp.com	https://discordapp.com/privacy	7/18/2017
48	discovery.com	https://corporate.discovery.com/privacy-policy/	2/2/2018
49	disqus.com	https://help.disqus.com/terms-and-policies/disqus-privacy-policy	2/2/2018
50	doubleclick.net	https://policies.google.com/technologies/ads	7/18/2017
51	dropbox.com	https://www.dropbox.com/terms	7/18/2017
52	dummies.com	http://www.dummies.com/privacy-policy/	7/18/2017
53	espncricinfo	http://www.espncricinfo.com/ci/content/site/company/privacy_policy.html	2/2/2018
54	facebook.com	https://www.facebook.com/policy.php	7/18/2017
55	fandango.com	http://www.docracy.com/0cszygfc5pc/fandango-com-privacy-policy-tos	2/2/2018
56	fatsecret.com	http://www.fatsecret.com/Default.aspx?pa=priv	2/2/2018
57	flipkart.com	https://www.flipkart.com/privacy-policy/p/itmdysazy9vmmewk	7/18/2017
58	ford.com	https://www.ford.com/help/privacy/	2/2/2018
59	foxnews.com	http://www.foxnews.com/privacy-policy.html	7/18/2017
60	fullscreendirect.com	https://alisonkrauss.com/privacy	2/2/2018
61	generalmills.com	https://www.generalmills.com/en/Company/privacy-policies/privacy-policy-US	2/2/2018
62	getcake.com	http://getcake.com/privacy-policy/	2/2/2018
63	goodreads.com	https://www.goodreads.com/about/privacy	7/18/2017
64	greyhound	https://www.greyhound.com/en/legal/privacy-policy	2/2/2018
65	grubhub	https://www.grubhub.com/legal/privacy-policy	2/2/2018
66	gucci.com	https://www.gucci.com/us/en/st/privacy-landing	2/2/2018
67	hbo.com	https://www.hbo.com/privacy-policy	7/18/2017
68	huffingtonpost.com	https://policies.oath.com/us/en/oath/privacy/index.html	7/18/2017
69	ikea.com	https://www.ikea.com/ms/en_US/privacy_policy/privacy_policy.html	7/18/2017
70	imdb.com	https://www.imdb.com/privacy	7/18/2017
71	indeed.com	https://www.indeed.com/legal	7/18/2017
72	indiatimes.com	https://www.indiatimes.com/privacypolicy/	7/18/2017
73	instagram.com	https://help.instagram.com/155833707900388	7/18/2017
74	kayak.com	https://www.kayak.com/privacy	7/18/2017
75	kaylaitsines.com	https://www.kaylaitsines.com/pages/terms-and-conditions	2/2/2018
76	kia.com	http://www.kia.com/us/en/content/privacy-and-legal/privacy-policy	2/2/2018
77	lego.com	https://www.lego.com/en-us/legal/legal-notice/privacy-policy-full	2/2/2018
78	lg.com	http://www.lg.com/us/privacy	7/18/2017
79	mcdonalds.com	https://www.mcdonalds.com/us/en-us/privacy.html	7/18/2017
80	medium.com	https://medium.com/policy/medium-privacy-policy-	2/2/2018

		f03bf92035c9	
81	meredith.com	http://www.meredith.com/privacy.html	2/2/2018
82	merriamwebster	https://www.merriam-webster.com/privacy-policy	2/2/2018
83	messenger.com	https://www.facebook.com/privacy/explanation	7/18/2017
84	metropcs.mobi	https://www.metropcs.com/terms-conditions/privacy.html	7/18/2017
85	mlb.com	http://mlb.mlb.com/mlb/official_info/about_mlb_com/privacy_policy.jsp	7/18/2017
86	movieweb.com	https://movieweb.com/privacy/	2/2/2018
87	mozilla.org	https://www.mozilla.org/en-US/privacy/firefox/	7/18/2017
88	msdonalds.com	https://www.mcdonalds.com/us/en-us/privacy.html	2/2/2018
89	myway.com	http://mywaypill.com/privacy-policy.php	7/18/2017
90	netflix.com	https://help.netflix.com/legal/privacy	7/18/2017
91	nike	https://www.nike.com/us/en_us/c/help/privacy-policy	2/2/2018
92	nsf	https://www.nsf.gov/policies/privacy.jsp	2/2/2018
93	nymag.com	http://nymag.com/newyork/privacy/	7/18/2017
94	nytimes.com	https://help.nytimes.com/hc/en-us/articles/115014892108-Privacy-policy	7/18/2017
95	paypal.com	https://www.paypal.com/us/webapps/mpp/ua/privacy-full	7/18/2017
96	pepsico.com	http://www.pepsico.com/legal/privacypolicy	2/2/2018
97	petfinder	https://www.petfinder.com/privacy-policy/	2/2/2018
98	pinterest.com	https://policy.pinterest.com/en/privacy-policy	7/18/2017
99	pncbank.com	https://www.pnc.com/en/privacy-policy.html	2/2/2018
100	politico.com	https://www.politico.com/privacy-policy	7/18/2017
101	popads.net	https://www.popads.net/privacy-policy.html	7/18/2017
102	popcash.net	https://popcash.net/privacy-policy	7/18/2017
103	popsugar.com	https://www.popsugar.com/privacy	2/2/2018
104	premierleague	https://www.premierleague.com/privacy-policy	2/2/2018
105	puppyfind.com	https://www.puppyfind.com/privacy_policy.html	2/2/2018
106	purch.com	http://www.purch.com/privacy-policy/	2/2/2018
107	quora.com	https://www.quora.com/about/privacy	7/18/2017
108	razerzone.com	https://www.razer.com/privacy-policy	2/2/2018
109	reimageplus.com	https://www.reimageplus.com/privacy-policy/	7/18/2017
110	roblox.com	https://en.help.roblox.com/hc/en-us/articles/115004630823-Roblox-Privacy-and-Cookie-Policy-	7/18/2017
111	savefrom.net	http://en.savefrom.net/privacy-policy.html	7/18/2017
112	slack.com	https://slack.com/privacy-policy	2/2/2018
113	slideshare.net	https://www.slideshare.net/ParallelWireless/privacy-policy-39152997	7/18/2017
114	snapchat	https://www.snap.com/en-US/privacy/privacy-policy	2/2/2018
115	soundcloud.com	https://soundcloud.com/pages/privacy	2/2/2018
116	southwest	https://www.southwest.com/html/about-southwest/terms-and-conditions/privacy-policy-pol.html	2/2/2018
117	spectrum	https://www.spectrum.com/policies/your-privacy-rights.html	2/2/2018
118	speedtest.net	http://www.speedtest.net/privacy	2/2/2018

119	stackoverflow.com	https://stackoverflow.com/questions/36106163/privacy-policy-for-a-personal-website-with-user-login-form-and-comment-sections	7/18/2017
120	starbucks.com	https://www.starbucks.com/about-us/company-information/online-policies/privacy-policy	2/2/2018
121	t.co	https://www.kuali.co/privacy-policy/	7/18/2017
122	techdirt.com	https://deals.techdirt.com/privacy	2/2/2018
123	tesco.com	https://www.tesco.com/termsandconditions/privacy.htm	7/18/2017
124	tesla.com	https://www.tesla.com/about/legal	2/2/2018
125	thehistorymakers.org	http://www.thehistorymakers.org/privacy-policy	2/2/2018
126	theredup.com	https://www.thredup.com/privacy-policy	2/2/2018
127	timeinc.com	https://subscription.timeinc.com/storefront/privacy/time/generic_privacy_new.html?dnp-source=B	7/18/2017
128	totallyhermedia.com	http://www.totallyhermedia.com/privacy/	2/2/2018
129	trello.com	https://trello.com/privacy	2/2/2018
130	tripadvisor.com	https://tripadvisor.mediaroom.com/us-privacy-policy	2/2/2018
131	tumblr.com	https://www.tumblr.com/privacy/en_eu	7/18/2017
132	twitch.tv	https://www.twitch.tv/user/legal?page=privacy_policy	2/2/2018
133	twitter.com	https://twitter.com/privacy	7/18/2017
134	uber.com	https://privacy.uber.com/policy	7/18/2017
135	uefa.com	http://www.uefa.com/privacypolicy/index.html	2/2/2018
136	usatoday.com	http://static.usatoday.com/privacy/	7/18/2017
137	vimeo.com	https://vimeo.com/privacy	7/18/2017
138	w3schools.com	https://www.w3schools.com/about/about_privacy.asp	2/2/2018
139	walmart.com	http://corporate.walmart.com/privacy-security/walmart-privacy-policy	7/18/2017
140	warnerbros.com	https://www.warnerbros.com/privacy	2/2/2018
141	warnerbrosrecords.com	http://www.warnerbrosrecords.com/privacy-policy	2/2/2018
142	washingtonpost.com	https://www.washingtonpost.com/privacy-policy/2011/11/18/gIQASIiaiN_story.html?utm_term=.6b2fc1574186	7/18/2017
143	weather.com	https://weather.com/en-US/twc/privacy-policy	2/2/2018
144	wendys.com	https://www.wendys.com/privacy-policy	2/2/2018
145	wetransfer.com	https://wetransfer.com/legal/privacy	2/2/2018
146	wordpress.com	https://wordpress.org/about/privacy/	2/2/2018
147	yahoo.com	https://policies.oath.com/us/en/oath/privacy/index.html	7/18/2017
148	yelp.com	https://www.yelp.com/tos/privacy_en_us_20160131	2/2/2018
149	youtube.com	https://www.youtube.com/static?template=privacy_guidelines	7/18/2017
150	zillow.com	https://www.zillow.com/corp/Privacy.htm	7/18/2017
151	zipcar.com	https://members.zipcar.com/site/privacy	2/2/2018
152	zomato	https://www.zomato.com/privacy	2/2/2018

2.2 Terms of Service Documents

The Terms of service agreement is mainly used for legal purposes by companies which provide software or services, such as browsers, e-commerce, search engines, social media, and transport services. A legitimate terms-of-service agreement is legally binding and may be subject to change. Companies can enforce the terms by refusing service.

We collected 100 terms of service agreement documents from various websites. We used www.randomlists.com to generate a list of 200 random websites. Each website in the list was visited by hand and those that either were not in English or did not contain a terms of service agreement were filtered out. This left us with 100 websites whose terms of service documents were collected. Below table has more details.

Sl.	Website	URL	Date
1	ABC	http://about.abc.net.au/terms-of-use/	7/10/2017
2	ACBJ	https://acbj.com/privacy	7/10/2017
3	addtonay	https://www.addtoany.com/terms	7/10/2017
4	Adweek	http://www.adweek.com/terms-use/	7/10/2017
5	Alarabiya	https://english.alarabiya.net/tools/terms-of-use.html	7/10/2017
6	AllNurses	http://allnurses.com/terms-info.html	7/10/2017
7	Allstate	https://www.allstate.com/about/terms.aspx	7/10/2017
8	Amazon	https://www.amazon.com/gp/help/customer/display.html?nodeId=508088	7/10/2017
9	AOL	https://legal.aol.com/legacy/terms-of-service_full-terms/	7/10/2017
10	APA	http://www.apa.org/about/termsofuse.aspx	7/10/2017
11	Apple	https://www.apple.com/ca/legal/internet-services/itunes/ca/terms.html	7/10/2017
12	Archive	https://archive.org/about/terms.php	7/10/2017
13	Arstechnica	https://arstechnica.com/uncategorized/2009/01/user-agreement/	7/10/2017
14	Ask	https://about.ask.fm/legal/en/terms.html	7/10/2017
15	bandcamp	https://bandcamp.com/terms_of_use	7/10/2017
16	BBC	https://www.bbc.co.uk/terms/	7/10/2017
17	biblehub	http://biblehub.com/terms.htm	7/10/2017
18	Bitcoin	https://bitcoin.org/en/legal	7/10/2017
19	Bitpay	https://bitpay.com/about/terms	7/10/2017
20	BlackFriday	https://blackfriday.com/terms	7/10/2017
21	Blinklist	https://www.blinkist.com/en/privacy.html	7/10/2017
22	Bloglovin	https://www.bloglovin.com/tos	7/10/2017
23	BoardGamegeek	https://boardgamegeek.com/terms	7/10/2017
24	BoingBoing	https://boingboing.net/2018/04/18/ul-for-eula.html	7/10/2017
25	Booking	https://www.booking.com/content/terms.html	7/10/2017
26	Boston	https://www.boston.com/member-agreement	7/10/2017
27	BTR	http://www.btrtoday.com/termsofservice/	7/10/2017
28	CAGOV	https://data.chhs.ca.gov/pages/terms	7/10/2017
29	Canada	https://squareup.com/ca/legal/ua	7/10/2017
30	Cargo	http://www.sbaglobal.com/terms-and-conditions/	7/10/2017
31	CaribouCoffee	https://www.cariboucoffee.com/footer-folder/terms-of-use	7/10/2017
32	CBS	https://www.cbscorporation.com/terms-of-use/	7/10/2017

33	CBSLocal	http://newyork.cbslocal.com/terms-of-use/	7/10/2017
34	chegg	https://www.chegg.com/termsofuse/	7/10/2017
35	CincyMuseum	https://www.cincymuseum.org/terms	7/10/2017
36	Cisco	https://www.cisco.com/c/en/us/about/legal/terms-conditions.html	7/10/2017
37	CommunityCoffee	https://www.communitycoffee.com/termsofuse	7/10/2017
38	Conduit	http://app4mobilebiz.wpengine.com/como-website-terms-of-use-2.html	7/10/2017
39	ConstantContact	https://www.constantcontact.in/legal/terms	7/10/2017
40	CreativeCommons	https://creativecommons.org/terms/	7/10/2017
41	Disney	https://disneytermsofuse.com/	7/10/2017
42	Drupal	https://www.drupal.org/project/terms_of_use	7/10/2017
43	EconRes	https://www.federalreserve.gov/econres/us-models-package.htm	7/10/2017
44	Envato	https://themeforest.net/legal/market	7/10/2017
45	Europa	https://europa.eu/youreurope/citizens/consumers/unfair-treatment/unfair-contract-terms/index_en.htm	7/10/2017
46	fc2	https://help.fc2.com/common/tos/en	7/10/2017
47	FIneart	https://fineartconnoisseur.com/terms-of-service/	7/10/2017
48	Fotki	https://help.fotki.com/terms/	7/10/2017
49	FreePress	http://static.freep.com/terms/	7/10/2017
50	Freewebs	http://www.webs.com/terms-of-service	7/10/2017
51	Google	https://policies.google.com/terms	7/10/2017
52	GuardianLiv	https://www.theguardian.com/info/2014/sep/09/guardian-live-events-terms-and-conditions	7/10/2017
53	healthy	https://dashboard.healthyroster.com/Home/TermsOfUse	7/10/2017
54	hp	http://www8.hp.com/us/en/terms-of-sale.html	7/10/2017
55	ImageShack	https://imageshack.com/terms	7/10/2017
56	ISGD	http://business.ohio.gov/efiling/terms/	7/10/2017
57	ISSUU	https://issuu.com/legal/terms	7/10/2017
58	kraftrecipes	http://www.kraftrecipes.com/about/useragreement.aspx	7/10/2017
59	lastfm	https://www.last.fm/api/tos	7/10/2017
60	latimes	http://articles.latimes.com/2013/jan/01/news/lat-terms	7/10/2017
61	lego	https://www.lego.com/en-us/legal/legal-notice	7/10/2017
62	liquor	https://www.liquor.com/terms-and-conditions/#gs._bqlnJo	7/10/2017
63	MapQuest	https://developer.mapquest.com/legal	7/10/2017
64	mayoclinic	https://socialmedia.mayoclinic.org/terms-conditions/	7/10/2017
65	meetup	https://www.meetup.com/terms/	7/10/2017
66	moonfruit	https://www.moonfruit.com/tc	7/10/2017
67	MS	https://www.microsoft.com/en-us/servicesagreement/	7/10/2017
68	neworleans	http://www.neworleansonline.com/notmc/termsandconditions.html	7/10/2017
69	nymag	http://nymag.com/newyork/terms/	7/10/2017
70	OAIC	https://www.oaic.gov.au/terms-and-conditions	7/10/2017
71	oracle	https://www.oracle.com/legal/terms.html	7/10/2017
72	Overstock	https://www.overstock.com/7935/static.html	7/10/2017
73	Patch	https://patch.com/terms	7/10/2017
74	playstation	https://www.playstation.com/en-us/network/legal/terms-of-service/	7/10/2017
75	prnews	http://www.prnewsonline.com/terms-of-use/	7/10/2017
76	PRWEB	http://service.prweb.com/legal/terms-of-service/	7/10/2017
77	PumpkinHead	https://www.allmusic.com/artist/pumpkinhead-mn0000857507	7/10/2017
78	quantcast	https://www.quantcast.com/terms/measure-terms-service/	7/10/2017

79	reference	https://en.wikipedia.org/wiki/Terms_of_service	7/10/2017
80	SeattleTimes	https://company.seattletimes.com/notices/notice1.html	7/10/2017
81	ShutterFlyInc	https://www.shutterflyinc.com/terms-of-use/	7/10/2017
82	SlashDotMdeia	https://slashdotmedia.com/terms-of-use/	7/10/2017
83	SoundCloud	https://soundcloud.com/terms-of-use	7/10/2017
84	Storify	https://storify.com/tos	7/10/2017
85	TheWeatherChannel	https://weather.com/legal	7/10/2017
86	TicketMaster	http://www.ticketmaster.com/h/terms.html	7/10/2017
87	TripAdvisor	https://tripadvisor.mediaroom.com/us-terms-of-use	7/10/2017
88	tulsaworld	http://www.tulsaworld.com/site/terms.html	7/10/2017
89	ucla	http://www.ucla.edu/terms-of-use/	7/10/2017
90	UCSD	https://ucsd.edu/about/terms-of-use.html	7/10/2017
91	USAToday	https://www.usatoday.com/legal/tos.html	7/10/2017
92	Vimeo	https://vimeo.com/terms	7/10/2017
93	WebEden	https://webeden.co.uk/terms.html	7/10/2017
94	WebNode	https://www.webnode.com/terms-and-conditions/	7/10/2017
95	Wikimedia	https://wikimediafoundation.org/wiki/Terms_of_Use/en	7/10/2017
96	WooCommerce	https://woocommerce.com/terms-conditions/	7/10/2017
97	wufoo	https://www.wufoo.com/terms-of-service/	7/10/2017
98	Yahoo	https://policies.oath.com/	7/10/2017
99	Yola	https://www.yola.com/terms	7/10/2017
100	zimbio	http://zimbio.whoiskenjackson.com/zimbio-terms-of-service/	7/10/2017

2.3 Miscellaneous Web Documents

For the third set, we wanted a high amount of diversity in content and domain. So we chose web pages by collecting the top two to four search results of keywords which span across topics: news, sports, botany, web design, photography, data science, cookie policies, HTML, history, migraine, dataset, technical documentation, shoes, grammar, kids stories and cricket. We skipped web pages which did not have sectional demarcations.

Sl.	Website	URL	Date
1	abcnews.go.com	https://abcnews.go.com/Technology/windows-things-microsofts-os/story?id=17570782	3/17/2018
2	adidas.com	http://www.adidas.com/us/help-topics-terms_and_conditions.html	4/28/2018
3	algorithmia.com	https://blog.algorithmia.com/introduction-to-microservices/	5/3/2018
4	britannica.com	https://www.britannica.com/science/botany	5/3/2018
5	cbsinteractive.com	https://www.cbsinteractive.com/legal/cbsi/privacy-policy/managing-cookies	5/3/2018
6	cbsinteractive.com	https://www.cbsinteractive.com/legal/cbsi/privacy-policy/third-party-online-advertising	3/26/2018
7	cel35gal.blogspot.com	https://cel35gal.blogspot.com/2018/04/6-ways-to-become-invaluable-at-work.html	5/7/2018

8	computerhope.com	https://www.computerhope.com/jargon/h/html.htm	5/1/2018
9	conda.io	https://conda.io/docs/release-notes.html	4/30/2018
10	cox.com	https://www.cox.com/aboutus/policies.html	5/1/2018
11	creativebloq.com	https://www.creativebloq.com/web-design/examples-of-html-1233547	4/1/2018
12	dailymom.com	http://dailymom.com/capture-2/7-basic-photography-rules/	4/28/2018
13	datasciencecentral.com	https://www.datasciencecentral.com/profiles/blogs/building-outiler-resistant-centroids-in-any-dimension	4/29/2018
14	economist.com	https://www.economist.com/cookies-info	4/30/2018
15	en.wikipedia.org	https://en.wikipedia.org/wiki/Feature_engineering	5/1/2018
16	environmentalscience.org	https://www.environmentalscience.org/botany	5/2/2018
17	excelsior.edu	https://www.excelsior.edu/about/ferpa-and-privacy-rights	5/3/2018
18	gameofthrones.com	http://gameofthrones.wikia.com/wiki/Game_of_Thrones_Wiki	4/30/2018
19	grammarly.com	https://www.grammarly.com/blog/articles/	4/30/2018
20	grammarly.com	https://www.grammarly.com/blog/practice-english-skills/	4/30/2018
21	grammarly.com	https://www.grammarly.com/blog/common-adjectives/	4/30/2018
22	hibu.com	https://hibu.com/legal/cookie-policy	3/17/2018
23	history.com	https://www.history.com/news/hungry-history/8-things-you-may-not-know-about-the-real-colonel-sanders	3/18/2018
24	htmlgoodies.com	https://www.htmlgoodies.com/beyond/cms/7-ways-to-start-learning-javascript-as-a-wordpress-developer.html	3/19/2018
25	inc.com	https://www.inc.com/amy-morin/8-things-mentally-strong-people-do-every-single-day.html	4/28/2018
26	indiegogo.com	https://learn.indiegogo.com/cookie-policy/	4/28/2018
27	intel.com	https://www.intel.com/content/www/us/en/privacy/intel-cookie-notice.html	4/28/2018
28	kickstarter.com	https://www.kickstarter.com/cookies	4/28/2018
29	lemurproject.org	http://lemurproject.org/clueweb09/	5/3/2018
30	lifehack.org	https://www.lifehack.org/297991/8-things-people-with-depression-want-you-know	5/3/2018
31	lifehack.org	https://www.lifehack.org/articles/productivity/8-things-successful-people-sacrifice-for-their-success.html	5/3/2018
32	migraine.com	https://migraine.com/blog/8-things-everyone-needs-to-know-about-migraine/	5/3/2018
33	ndsu.edu	https://www.ndsu.edu/gradschool/graduating_students/dtp/format/headings/#c151852	5/3/2018
34	newwoldencyclopedia.org	http://www.newworldencyclopedia.org/entry/Botany	3/26/2018

35	nike.com	https://help-en-us.nike.com/app/answer/article/productcare-shoes/a_id/43/country/us	3/26/2018
36	nltk.org	http://www.nltk.org/howto/chat80.html	4/17/2018
37	plainlanguage.gov	https://www.plainlanguage.gov/guidelines/organize/add-useful-headings/	5/5/2018
38	premierleague.com	https://www.premierleague.com/cookie-policy	5/5/2018
39	scientificamerican.com	https://www.scientificamerican.com/page/use-of-cookies/	5/1/2018
40	seoland.in	https://seoland.in/14-resources-to-improve-your-writing-and-editing-skills/#.Wvho1kO5tdo	4/1/2018
41	snap.com	https://www.snap.com/en-US/cookie-policy/	4/28/2018
42	squarespace.com	https://www.squarespace.com/cookie-policy/	5/2/2018
43	statcounter.com	https://pt_br.statcounter.com/about/cookies/	5/2/2018
44	stonegableblog.com	http://www.stonegableblog.com/7-rules-for-choosing-perfect-curtains/	5/4/2018
45	storiestogrowby.org	https://www.storiestogrowby.org/story/pinocchio-fairy-tale-story-english-kids/	5/5/2018
46	storiestogrowby.org	https://www.storiestogrowby.org/story/early-reader-snow-queen-fairy-tale-english-stories-kids/	4/30/2018
47	symantec.com	https://www.symantec.com/privacy/gps-english	4/30/2018
48	thehindu.com	http://www.thehindu.com/sport/cricket/kkr-vs-csk-a-battle-of-the-batsmen-on-the-cards/article23751496.ece	5/5/2018
49	travelingspud.com	http://www.travelingspud.com/2015/06/01/8-things-to-do-in-moscow-idaho/	5/5/2018
50	w3schools.com	https://www.w3schools.com/html/html_responsive.asp	5/5/2018
51	wwwdb.inf.tu	https://wwwdb.inf.tu-dresden.de/misc/dwtc/	5/5/2018

3 Organization of Data

Each web document is inside its own folder and it contains the below two files.

Filename	**Purpose**
priv.html	Original web document
gold.html	Sanitized gold version