

Extracting Relevant subdomains and tags for top 20 UK domains :

n°	domain	subdomain	relevant/irrelevant
1	bbc.co.uk	" , 'www'	Relevant with the tag /news/ bbc.co.uk/news/
		'careershup','account','bbcsignups.external','session','ssl','careers','search click','bbcgoodfood','shop','download s-app.iplayer.api','search','beta','click.email','swscdn','iplayerhelp.external','m'	irrelevant
2	theguardian.com	'ablink.editorial'	Relevant
		" , 'www'	Relevant with tags : /news/ /world/ /commentisfree/ /us-news/ /uk-news/ /politics/ /education/ /society/ /science/ /business/ /money/ /sport/ /australia-news/ /travel/ /technology/ /commentisfree/
3	dailymail.co.uk	'www', "	Relevant with tags : /news/ /columnists/ /debate/ /femail/ /health/ /money/ /sport/ /travel/ /tvshowbiz/
		"discountcode','c-5uwzmx78pmca09x24ntqx78jwizlx2ekwu.g00','c-7npsfqifvt34x24gbwfx2edp.g00','c-7npsfqifvt3	Irrelevant

		4x24x78x78x78x2etjmlgsfex2edpn.g00','c-6rtwjumjzx7877x24bbbx2emtrjx78jwajx2ehtr.g00','c-6rtwjumjzx7877x24fihqnpx2elx2eitzgqjhqnpx2esjy.g00','click.email.discountcode','gotolink','creative','t','secured','fff'	
4	sky.com	'news'	Relevant with tag /story/
		", 'ogwam', 'www', 'epgservices', 'helpforum', 'secure', 'buy', 'messages', 'mysky', 'my', 'trackmyorder', 'help', 'pubfinder', 'tv', 'myhelprequests', 'skyvision', 'broadbandshield', 'go', 'payments', 'election.news', 'partner.help', 'customermonth', 'careers', 'myaccount', 'contactus', 'broadbandreconnection', 'rewards', 'claim', 'www.fafe', 'community'	Irrelevant
5	thesun.co.uk	'www', "	Relevant with tags : /fabulous/ /money/ /news/ /sport/ /travel/ /tech/
		'link'	Relevant: External link for articles
		'racingpredictor', 'superdays', 'feeds', 'mypreferences', 'europeanpredictor', 'email.mail.europeanpredictor', 'directory', 'cricketpredictor', 'pixelappcollector', 'iframe'	Irrelevant
6	telegraph.co.uk	'www'	Relevant with tags: /politics/ /news/
		", 'surveys.contributors', 'secure', 't.email3', 'fantasyracing', 'contributors', 'www.hand-picked', 'expatdirectory', 'announcements', 'fantasycricket', 'premium', 'fantasyfootball', 'gardenshop', 'registration', 'compare', 'tickets', 'events', 'puzzles', 'auth',	Irrelevant

		'digitaledition', 'jobs', 'performance-data.gcpdata'	
7	express.co.uk	'www', "	All relevant but /showbiz/
8	metro.co.uk	"	Relevant
		'discountcode', 'link.news', 'www', 'api'	Irrelevant
9	itv.com	'www', "	Relevant with tag : /news/
		'itv7', 'tom', 'help', 'email.s', 't.e', 'spike', 'secure.pes'	Irrelevant
10	independent.co.uk	'www', "	Relevant with tag : /news/
		'link.e', 'sanmiguelrichlist', 'link.indy100', 'my'	Irrelevant
11	thetimes.co.uk	'www'	Relevant with tag : /article/
		'login', 'mypreferences', 'home', 'epaper', "", 'links.mailings'	Irrelevant
12	channel4.com	'www'	Relevant with tag : /news/
		'4id', 'thanks4', 'hello', 'www.core4', 'core4', 'f1', "", 'assets-corporate', 'origin-thanks4'	Irrelevant
13	inews.co.uk	"	Relevant with tag : /news/
14	buzzfeed.com	'www', "", 'link', 'shop'	Everything is irrelevant
15	thetelegraphandargus.co.uk	'www', "	Relevant with tag : /news/
16	nytimes.com	'www', "	Relevant
17	standard.co.uk	", 'www'	Relevant with tag : /news/
18	kentonline.co.uk	", 'www'	Relevant
19	ft.com	", 'www'	Relevant with tag : /content/
20	eveningtimes.co.uk	", 'www'	Relevant with tag : /news/ , /content/

The top **20** newspapers make **90,82%** of the **UK** newspapers.

	domain	occurrence	fraction \
0	bbc	35326	0.381321
1	theguardian	8696	0.093868
2	dailymail	7766	0.083829
3	sky	6473	0.069872
4	thesun	5127	0.055343
5	telegraph	3071	0.033149
6	express	2392	0.025820
7	metro	2197	0.023715
8	itv	2053	0.022161
9	independent	1691	0.018253
10	thetimes	1629	0.017584
11	channel4	1454	0.015695
12	inews	1350	0.014572
13	buzzfeed	1215	0.013115
14	thetelegraphandargus	897	0.009683
15	nytimes	768	0.008290
16	standard	552	0.005958
17	kentonline	501	0.005408
18	ft	492	0.005311
19	eveningtimes	491	0.005300
20	dailystar	455	0.004911