Goal

Check that Websegmenter is able to extract information of the website as well as the classic pure html `extractor`.

Associated User Stories

- https://dev.azure.com/alpha10x/ALPHA10X/_workitems/edit/9529
- https://dev.azure.com/alpha10x/ALPHA10X/_workitems/edit/9528

In [ ]:
```
from utils import WebSegmenter,visualize_graph,get_summary
import pandas as pd
import pprint
import seaborn as sns
from datetime import datetime
import pickle
import numpy as np
```

Get Output graph of websegmenter

In [ ]:
```
pdf = pd.read_csv('PE_relevant_sample_2023_05_23.csv')
pdf.head()
print(pdf.shape)
```

```
(1000, 2)
```

In [ ]:
```
# pdf = pdf.iloc[0:300,:]
```

In [ ]:
```
def get_graph(url):
    websegmenter = WebSegmenter(url=url)
    websegmenter.run()
    try :
        graph = websegmenter.graph
    except:
        graph = None
    return graph
```

In [ ]:
```
#pdf['graph'] = pdf['url'].apply(lambda x: get_graph(x))

out = []
for i, url in enumerate(pdf['url'].values.tolist()):
    print(i, '->', url)
    graph = get_graph(url)
    out.append(graph)

pdf['graph'] = out
```

0 -> http://churchillglass.co.uk
1 -> http://www.ppwoniu.com
2 -> http://www.hansecom.com
3 -> http://costimp.it
4 -> http://ortesia.com
5 -> http://www.afit.ro
6 -> http://www.phone580.com
7 -> http://www.garage-gatti.fr
8 -> http://jaipurhaveli.com
9 -> http://www.bbci.eu
10 -> http://publish.manheim.com
11 -> http://www.elektronik-produkt.de
12 -> http://www.dsjet.com
13 -> http://seamfix.com
14 -> http://tailwindapp.com
15 -> http://www.zeusagro.com
16 -> http://yoappstore.com
17 -> http://bucknerbarrel.com
18 -> http://chiropodyandpodiatry.co.uk
19 -> http://astroprint.cz
20 -> http://www.jabe.net
21 -> http://www.dekonconstruction.com
22 -> http://www.aldautomotive.co.uk
23 -> http://www.JARCO.com
24 -> http://www.kyocera-hardcoating.com
25 -> http://soneva-beauty.co.uk
26 -> http://www.sauter-personal.de
27 -> http://www.instrumentariumdental.com
28 -> http://pharmacyinsurance.com.au
29 -> http://corbets.com.au
30 -> http://chenerycontractorsltd.co.uk
31 -> http://pmcrobot.cn
32 -> http://alcaautomotive.com.au
33 -> http://www.pmsidirect.com
34 -> http://www.fcac.org
35 -> http://www.growhomes.in
36 -> http://www.ndxcards.com
37 -> http://raymillardinsurancebrokers.co.uk
38 -> http://www.censo.nl
39 -> http://limesdc.com.au
40 -> http://www.unirobot.com
41 -> http://twicetime.com
42 -> http://www.thevictorcloset.com
43 -> http://www.yiamas-remscheid.de
44 -> http://www.waysidewatergardens.co.uk
45 -> http://www.factorydoorservices.co.uk
46 -> http://www.universalcoal.com
47 -> http://www.fundacaotorino.com.br
48 -> http://www.weipl.net
49 -> http://scoop-international.com
50 -> http://www.french-mortgage.com
51 -> http://www.klara.hr
52 -> http://www.carlsonhall.com
53 -> http://www.dmyec.com
54 -> http://anamma.be
55 -> http://www.contractrecruiter.com
56 -> http://www.musikverein-lyra-stupferich.de
57 -> http://kookaeye.com.au
58 -> http://www.advancedclimatesolutionshvac.com
59 -> http://www.mediplus.co.uk

60 -> http://knightguardsecurity.co.uk
61 -> http://sofys.com
62 -> http://www.australboreal.ca
63 -> http://drcarsons.com
64 -> http://wholedesignstudios.com
65 -> http://xenoware.be
66 -> http://www.besserco.com.au
67 -> http://maistro.ru
68 -> http://restaurant-attwenger.at
69 -> http://www.makingsense.co.nz
70 -> http://renukajewellers.com
71 -> http://www.pasticceria-etna.de
72 -> http://bluestackcloud.com
73 -> http://maldronhotelpearsestreet.com
74 -> http://keder-rails.co.uk
75 -> http://camperdowncoll.vic.edu.au
76 -> http://www.turunculevye.com
77 -> http://yoalearning.com
78 -> http://www.comparethemanandvan.co.uk
79 -> http://aaautocarelv.com
80 -> http://www.manifest.com
81 -> http://welltech.hu
82 -> http://saptham.com
83 -> http://wildwesthelicopters.com
84 -> http://www.mcc-machinery.com
85 -> http://www.thegrandhouse.com
86 -> http://www.cabaneindigo.com
87 -> http://fortcollinsshoes.com
88 -> http://www.lvo-gmbh.de
89 -> http://abento.de
90 -> http://www.nextgenerationdata.com
91 -> http://lemons.com.tr
92 -> http://www.americas.technology
93 -> http://www.psgcable.com
94 -> http://www.yipled.com
95 -> http://mobilerobotguide.com
96 -> http://thakuralelectric.com
97 -> http://janesafricanviolets.com.au
98 -> http://spatzennest-kerken.de
99 -> http://www.kulturhaus-kaefertal.de
100 -> http://petroases.com
101 -> http://www.ar-corporation.com
102 -> http://www.indiamart.com
103 -> http://tecsupri.net
104 -> http://avsbd.net
105 -> http://palmbeachmarinefuel.com
106 -> http://stmichaelinthehamletschool.com
107 -> http://www.synerbuild.com
108 -> http://www.nsin.us
109 -> http://bishvilaych.org
110 -> http://capoom.com
111 -> http://amaccontabil.com.br
112 -> http://jcqh.en.alibaba.com
113 -> http://jonesfinancialadvisorygroup.com
114 -> http://www.teste.com
115 -> http://www.first-global.com
116 -> http://www.qualitywaterair.com
117 -> http://kitabeli.id
118 -> http://vasanthamfoundry.com
119 -> http://mainstreethigh.com

120 -> http://korfezhalideediportaokulu.meb.k12.tr
121 -> http://alsafartours.co.uk
122 -> http://visagelaser.co.uk
123 -> http://svglaw.ca
124 -> http://palaiseau-basket.fr
125 -> http://ktbridal.co.uk
126 -> http://euhost.com
127 -> http://drfabianofalasque.com.br
128 -> http://www.lomondhillshotel.com
129 -> http://www.simef.fr
130 -> http://www.capstonecrea.com
131 -> http://jequitibaconstrutora.com.br
132 -> http://www.chaos-coworking.de
133 -> http://www.limosceneservices.com
134 -> http://adriangainesdecoratingservices.co.uk
135 -> http://www.global-rent-a-car.com
136 -> http://www.frontiernr.com
137 -> http://www.yatemanandsons.co.uk
138 -> http://www.visitaurora.com
139 -> http://www.redpathtechnicalservices.com
140 -> http://www.grand-hotel-des-bains.com
141 -> http://www.myabundantliving.com
142 -> http://bouwwerken-hst.be
143 -> http://northhertfordshirecollegehitchin.2day.uk
144 -> http://softicu.com
145 -> http://findorffschule-ohz.de
146 -> http://www.tmconnected.com
147 -> http://www.evercold.be
148 -> http://www.plisa.com.mx
149 -> http://www.dbtss.com
150 -> http://www.independentminingservices.com.au
151 -> http://www.hotel-schwarzbachtal.de
152 -> http://www.puff-inoue.jp
153 -> http://ppb.be
154 -> http://www.microco.sm
155 -> http://www.trovea.com
156 -> http://blackstoreboutique.negocio.site
157 -> http://ortechsecurityservices.com.au
158 -> http://pgcvfra.org
159 -> http://www.ozelyalovahastanesi.com.tr
160 -> http://cnpeg.cn
161 -> http://lovelydoggyhouse.ca
162 -> http://ohooho.net
163 -> http://cure.link
164 -> http://corriginpharmacy.com.au
165 -> http://mymuhc.org
166 -> http://www.befeld.de
167 -> http://apredef.com
168 -> http://www.quanim.fr
169 -> http://www.salamurano.com
170 -> http://petrolube.co.id
171 -> http://rapid.com.tw
172 -> http://www.grandbearresort.com
173 -> http://hintermueller.at
174 -> http://erodecancercentre.com
175 -> http://hollowaysauctioneers.co.uk
176 -> http://www.city.shinjuku.lg.jp
177 -> http://www.bluelotus360.com
178 -> http://www.blueskykids.com.cn
179 -> http://www.ascentair.com

180 -> http://www.relationmediasales.dk
181 -> http://ict.ae
182 -> http://ffvs.org.au
183 -> http://www.wreducons.com
184 -> http://www.brachin.com
185 -> http://holzweiler-schreinerei.de
186 -> http://thewebhub.com.au
187 -> http://brius.com
188 -> http://stevenliphotography.com
189 -> http://www.indiamart.com
190 -> http://srmagma.co.rs
191 -> http://adviesgroeppinson.be
192 -> http://deltaativa.com.br
193 -> http://www.oxhn.com
194 -> http://www.adeptclippingpath.com
195 -> http://goego.in
196 -> http://houstonlegalaidteam.wordpress.com
197 -> http://dianepancelmusic.com
198 -> http://elementsbarandgrill.com.au
199 -> http://buttercupdairies.co.nz
200 -> http://www.intersagal.de
201 -> http://krawetzke.de
202 -> http://e-deutschland.de
203 -> http://www.profix24.de
204 -> http://www.cristeaarhitectura.ro
205 -> http://www.jswhite.co.uk
206 -> http://zermatt.se
207 -> http://vicway.com.au
208 -> http://hicoa.com.br
209 -> http://creationsmicheline.com
210 -> http://www.makelotion.com
211 -> http://dewzilla.com
212 -> http://www.lovingcareanimalhospital.com
213 -> http://indigoasie.com
214 -> http://www.southwestaan.com
215 -> http://westindiaexports.com
216 -> http://www.nwegar.com
217 -> http://www.espace-cuisines.com
218 -> http://www.treccorp.com
219 -> http://www.indiamart.com
220 -> http://www.sutton-estates.co.uk
221 -> http://wildwoodtrees.com
222 -> http://kubernesys.com
223 -> http://www.viriksontravel.co.uk
224 -> http://7groaster.pt
225 -> http://www.sportstv.com.tr
226 -> http://andycaseyphotography.co.uk
227 -> http://motoactionimola.it
228 -> http://www.darululoomonline.org
229 -> http://factorychic.com
230 -> http://auto700.com.br
231 -> http://beachhouse25.nl
232 -> http://www.petermaksymuksurveying.co.uk
233 -> http://illawarraophthalmology.com.au
234 -> http://www.berlinersingles.de
235 -> http://www.epicbounties.com
236 -> http://caritas-singen.org
237 -> http://joingoldstarmortgage.com
238 -> http://www.littledaffodilspreschool.com
239 -> http://eeleonice.blogspot.com

240 -> http://www.deroyal.com
241 -> http://pledgeit.org
242 -> http://bespokekitchenworkshop.co.uk
243 -> http://bayerische-massivhaus.de
244 -> http://www.ashbyandatkinson.com
245 -> http://www.playtopstreet.com
246 -> http://www.eureka-resources.com
247 -> http://wiarquitetura.com.br
248 -> http://rangestreeworks.com.au
249 -> http://www.42photo.com
250 -> http://colegioloreto.com
251 -> http://www.halkwinds.com
252 -> http://www.balticexchange.com
253 -> http://www.sernkou.com
254 -> http://www.brandschutz-beratung.de
255 -> http://www.intentinterior.com
256 -> http://newgroundchurches.org
257 -> http://zelhealth.com
258 -> http://www.geertjefoth.de
259 -> http://stb-klose.de
260 -> http://nutsfactorynyc.com
261 -> http://www.atelierairsoft.fr
262 -> http://www.smilesdentalcentre.co.uk
263 -> http://bluestonefmc.com.au
264 -> http://www.instantanet.net
265 -> http://www.itering.io
266 -> http://www.sunkean.com
267 -> http://www.wallacefoundation.org
268 -> http://www.lpclaw.com
269 -> http://toit-du-monde.com
270 -> http://45000feet.com
271 -> http://www.restaurant-marinella.com
272 -> http://alecell.com.br
273 -> http://guaramangueiras.com.br
274 -> http://montespatagonicos.com.ar
275 -> http://www.whutweshare.com
276 -> http://www.pelene.co.uk
277 -> http://wsg-wohnen.de
278 -> http://buzzbouz.com
279 -> http://www.ethnikos-bc.gr
280 -> http://realschule-ahrweiler.de
281 -> http://ultimatelasertag.ca
282 -> http://senderoverdepr.com
283 -> http://siroglutim.business.site
284 -> http://dollinger-realschule.de
285 -> http://www.allawgp.com
286 -> http://essofuelfinder.co.uk
287 -> http://www.designthinkersacademy.com
288 -> http://deine-eisbar.de
289 -> http://washmywhip.com
290 -> http://welcu.com
291 -> http://costadecaparica.com
292 -> http://alfristondaycentre.co.uk
293 -> http://allgardenmachines.com
294 -> http://www.hagl-vaterstetten.de
295 -> http://www.aptimus.com
296 -> http://www.minarikdrives.com
297 -> http://movserv.com.br
298 -> http://laradiogospel.ca
299 -> http://furnhouse.com.au

300 -> http://www.caponamibia.com
301 -> http://rtvsutter.ch
302 -> http://roguethink.com
303 -> http://www.mygamez.com
304 -> http://ilmilanoditoto.be
305 -> http://www.velcomex.com
306 -> http://pul2.com
307 -> http://galrmarketing.com
308 -> http://invoiceplatform.com
309 -> http://daviesfurniture.co.uk
310 -> http://watersideantiques.co.uk
311 -> http://4008812356.com
312 -> http://residenciamarerafols.wordpress.com
313 -> http://marmorariabelasartes.com.br
314 -> http://orioncymbals.com.br
315 -> http://yarraoncology.com.au
316 -> http://svsystems.blogspot.com
317 -> http://www.vitalthings.no
318 -> http://matermed.com.br
319 -> http://www.elepants.com.ar
320 -> http://moirmedical.com.au
321 -> http://www.jxzhixian.com
322 -> http://frescointeriors.ca
323 -> http://www.toptruecn.com
324 -> http://emeraviaggi.it
325 -> http://www.ai-con.co.jp
326 -> http://l-osales.co.za
327 -> http://icbsl.blogspot.com
328 -> http://septechconsulting.com
329 -> http://embellish.com.au
330 -> http://coinflex.com
331 -> http://prosperpetroleum.com
332 -> http://www.studio-silver.com
333 -> http://www.eelpieislandmusic.com
334 -> http://mattig.cc
335 -> http://www.incubasia-ventures.com
336 -> http://ilfracombelifeboat.org.uk
337 -> http://www.abricop.com
338 -> http://www.dwellstudent.com.au
339 -> http://bhenandco.com.au
340 -> http://www.premieraudiovisual.com
341 -> http://aroundtheclockfamily.com
342 -> http://www.idspl.com
343 -> http://www.bjwishing.com.cn
344 -> http://nambuccadentalsurgery.com.au
345 -> http://builders-supply.co.uk
346 -> http://www.artcandlesuk.co.uk
347 -> http://aixplatform.com
348 -> http://www.fenceworkshop.com
349 -> http://osgodbyvillageinstitute.btck.co.uk
350 -> http://www.wcblueprints.com
351 -> http://lacavadesign.ca
352 -> http://www.mobilifiver.com
353 -> http://www.elexcon.com
354 -> http://www.pinkqueen.com
355 -> http://kinderopvangjipenjanneke.be
356 -> http://sonntag-morgenmagazin.de
357 -> http://lamolisana.de
358 -> http://cgtpv.org
359 -> http://www.thyssenkruppdelicate.de

360 -> http://legiteamargot.com
361 -> http://www.mathiasfossum.com
362 -> http://thealtius.com
363 -> http://nordicpics.co.uk
364 -> http://www.hamiltoncountylandbank.org
365 -> http://eneasmarques.atende.net
366 -> http://www.easinteriors.com
367 -> http://rowlandsavenuepreschool.co.uk
368 -> http://greendoor-mortgages.co.uk
369 -> http://www.glidersports.com
370 -> http://www.megamart2-ecsel.eu
371 -> http://mlwlf.com
372 -> http://www.venenklinik-blaustein.de
373 -> http://www.sekuurshop.be
374 -> http://www.adaptiveseeds.com
375 -> http://www.taloturva.fi
376 -> http://www.stivabak.fr
377 -> http://reklama-centrum.cz
378 -> http://agence-securise.com
379 -> http://www.diswest.fi
380 -> http://difereconsultoria.com.br
381 -> http://www.elpitazo.net
382 -> http://bollatel.com.br
383 -> http://www.auto-wieser.eu
384 -> http://www.nimbus9usa.com
385 -> http://www.messiahentertainment.com
386 -> http://ihlsee-restaurant.de
387 -> http://conservatorio.ch
388 -> http://jaktpasset.se
389 -> http://leepeckmedia.com
390 -> http://www.d-ambra.com
391 -> http://www.wsbengineers.com
392 -> http://awf.edu.pl
393 -> http://www.indiamart.com
394 -> http://gusnerassociates.com.au
395 -> http://www.opusonesolutions.com
396 -> http://finestel.com
397 -> http://sdsoftsolutions.com
398 -> http://www2.dabpremiumfinance.com
399 -> http://glasgowmaritimeacademy.co.uk
400 -> http://induma.com.br
401 -> http://mula-berlin.de
402 -> http://itriadi.com
403 -> http://chauthiduniya.com
404 -> http://www.cdu-rotenburg.de
405 -> http://cielofm.com
406 -> http://mack-fellbach.de
407 -> http://mtgravattsubaru.com.au
408 -> http://www.vhspringfield.com
409 -> http://www.decoconseil.com
410 -> http://paryagraj-carpenter.business.site
411 -> http://jdroofing.com.au
412 -> http://jp-mi.com
413 -> http://www.construction-dynamic-systems.be
414 -> http://www.holidaycaravanparade.com
415 -> http://locaguincho.com.br
416 -> http://advmotor.co.uk
417 -> http://cityincolour.com.au
418 -> http://www.touchofmodern.com
419 -> http://trotan.no

420 -> http://evs-safety.de
421 -> http://smartwaysystem.com.au
422 -> http://sushizero.it
423 -> http://afvac.com
424 -> http://ibram.ind.br
425 -> http://www.tyretracks.co.uk
426 -> http://www.blackeight.com
427 -> http://www.og.com.cn
428 -> http://misrcontract.com
429 -> http://www.rbpa.co.uk
430 -> http://colegioalemansevilla.com
431 -> http://mevamekitchenexpress.ca
432 -> http://www.servibocage.pt
433 -> http://springfieldspartans.org
434 -> http://pcdobpe.org.br
435 -> http://www.info@zwaan-son.nl
436 -> http://stickerei-funk.de
437 -> http://flextem.de
438 -> http://victorcontractor.com.sg
439 -> http://www.threetiermedia.com
440 -> http://heerlenmijnstad.nl
441 -> http://www.cioatotransportes.com.br
442 -> http://www.templelogic.com
443 -> http://le-bouclart.overblog.com
444 -> http://pjhallett.com.au
445 -> http://edenbrewhouse.com.au
446 -> http://craftcoffeehouse.co.uk
447 -> http://pecanzfashion.com
448 -> http://www.styleowner.com
449 -> http://www.mortgageoptions.co.uk
450 -> http://www.gxpynm.com
451 -> http://mpreis.at
452 -> http://www.stanbicibtcpension.com
453 -> http://barasatcollege.org
454 -> http://joaoxxiii.com
455 -> http://resolutiontelevision.com
456 -> http://www.celledge.in
457 -> http://vialecanova.com.au
458 -> http://igrejavirtude.com.br
459 -> http://www.verticalnerve.com
460 -> http://louises-kitchen.business.site
461 -> http://www.ferienhaus-breyer.de
462 -> http://bv-bausysteme.de
463 -> http://discountdriver.fr
464 -> http://www.agentur-beziehungsweise.de
465 -> http://fnk-i.com
466 -> http://www.pbconception.com
467 -> http://www.bigredbus.biz
468 -> http://www.vtc-solutions.com
469 -> http://ledermair.at
470 -> http://fasi-qm.com
471 -> http://opioidsolutions.org
472 -> http://escoladharma.com.br
473 -> http://www.ajyaguru.com
474 -> http://gt-medien.de
475 -> http://barloventoviajes.com.ar
476 -> http://orangestudio.co.uk
477 -> http://www.watco.ie
478 -> http://unitedcanvas.com
479 -> http://poplars.suffolk.sch.uk

480 -> http://renovatorsparadise.com.au
481 -> http://www.ledpower.cl
482 -> http://www.aplend.com
483 -> http://www.hjg-sim.de
484 -> http://pastelfm.com
485 -> http://eletricamartins.com.br
486 -> http://www.bbcontabil.com.br
487 -> http://tornearіasaojorge.com.br
488 -> http://baryo-tours.business.site
489 -> http://www.shlomo-bit.co.il
490 -> http://www.fretigo.com
491 -> http://www.icentapp.com
492 -> http://finetec.in
493 -> http://champagnedumont.com
494 -> http://www.czzy.cn
495 -> http://thedecofactory.be
496 -> http://licomarques.com.br
497 -> http://www.learnenglishplus.com
498 -> http://www.lowcostskips.co.uk
499 -> http://www.strandresort-ostsee.de
500 -> http://windclouds.net
501 -> http://magnoliahairsalon.com.au
502 -> http://kommunalfahrzeuge.de
503 -> http://www.tk.co.th
504 -> http://www.premier-bearings.org
505 -> http://confeccoesmauricio.com.br
506 -> http://www.paradiseglobal.com.au
507 -> http://mk-j.co.jp
508 -> http://rainbowaviation.in
509 -> http://alhuraizgroup.ae
510 -> http://todolar.com.br
511 -> http://www.meganelec.be
512 -> http://grafikart.it
513 -> http://kettgen.de
514 -> http://www.lic-bcbc.com
515 -> http://mpc-management.com
516 -> http://www.prophotonix.com
517 -> http://pfcustomcountertops.com
518 -> http://www.maverickentrepreneurs.com
519 -> http://www.maakindustries.com
520 -> http://mservice.kz
521 -> http://mulheresdocafe.com.br
522 -> http://www.softsyshosting.com
523 -> http://www.evju-loebzi.de
524 -> http://www.rmsistemas.net
525 -> http://tateybred.com
526 -> http://uhtrading.com
527 -> http://www.elmdonlodge.com
528 -> http://www.greatergiving.com
529 -> http://mitreocraftmerida.es
530 -> http://lashbergtowing.ca
531 -> http://www.playkinderworld.com
532 -> http://gkh-versicherungsmakler.de
533 -> http://midiaoffice.com.br
534 -> http://mirotech.com
535 -> http://www.vineyardotr.org
536 -> http://www.yanmar.com
537 -> http://www.fjordstjernen.dk
538 -> http://vestacon.ca
539 -> http://magrisacanarias.es

540 -> http://www.tamatransport.com
541 -> http://www.lifeguardli.com
542 -> http://www.theabbeyinn.com
543 -> http://www.orca-ai.io
544 -> http://www.tisserant.fr
545 -> http://treppenmeister.com
546 -> http://chezsarah.net
547 -> http://giraflorstore.com.br
548 -> http://www.lyhlhg.com
549 -> http://www.agricircle.com
550 -> http://www.worldcontainerindex.com
551 -> http://financies.com.br
552 -> http://www.medicalert.ca
553 -> http://phindia.com
554 -> http://www.surge-electricalestimating.co.uk
555 -> http://www.portnou.com
556 -> http://mesquitegaming.com
557 -> http://gsti.com.mx
558 -> http://www.skcpl.in
559 -> http://flashlight.dance
560 -> http://www.wilnerassociates.com
561 -> http://elymartins.com.br
562 -> http://www.kiesjeoutplacementbureau.nl
563 -> http://volleyball-ntsv.de
564 -> http://www.southdownsleisure.co.uk
565 -> http://sydneysurvey.co
566 -> http://www.assemcorp.com
567 -> http://bumhanvina.com
568 -> http://www.kauno-atikas.com
569 -> http://findajewishschool.co.uk
570 -> http://monopatin.co.kr
571 -> http://www.mietomnibusse.de
572 -> http://universalfamilyofschools.org
573 -> http://www.natlinear.com
574 -> http://awlconsulting.com.au
575 -> http://www.indiamart.com
576 -> http://www.haulwel.com
577 -> http://jaipurjewelleryshow.org
578 -> http://www.eselondon.ac.uk
579 -> http://www.sw91.com
580 -> http://www.tgs-klissenbauer.de
581 -> http://tgcltd.co.uk
582 -> http://pacificdriveins.com
583 -> http://katal-defense.com
584 -> http://www.con4m.com
585 -> http://code4hr.org
586 -> http://zolidmanufacturing.co.uk
587 -> http://jxzywl.com.cn
588 -> http://maxsysbrasil.com.br
589 -> http://www.galledia.ch
590 -> http://pr.business
591 -> http://www.herbfarm.co.uk
592 -> http://gloudemansgevelprodukten.nl
593 -> http://www.CollTree.co.za
594 -> http://www.sparklightuae.com
595 -> http://datahelpsoftware.com
596 -> http://www.lynxmedia.info
597 -> http://paradiseandaman.com
598 -> http://itcbr.com.br
599 -> http://melitta.ch

600 -> http://weddingexhibitions.wordpress.com
601 -> http://www.jtn.cc
602 -> http://www.myservicesusa.com
603 -> http://theplugdrink.com
604 -> http://www.facebook.com
605 -> http://recantocataratasresort.com.br
606 -> http://www.fpasociados.es
607 -> http://atelierfollaco.com
608 -> http://www.tradersedgellc.com
609 -> http://tlsrobot.com
610 -> http://www.pophealthman.com
611 -> http://www.dorridgevillagehall.org
612 -> http://shanghaibrewery.com
613 -> http://prestigeauto.com.au
614 -> http://www.kuda.co.za
615 -> http://fussmedikal.com
616 -> http://northpointedentist.com
617 -> http://evolutiontrainingcenter.ca
618 -> http://www.penghuothgroup.com
619 -> http://rlxsolutions.com
620 -> http://peritosgrafotecnicos.blogspot.com.br
621 -> http://salesianos.edu
622 -> http://cliniquephysio.ca
623 -> http://akaldiesel.com
624 -> http://kwmwlaw.com
625 -> http://www.spctogether.co.uk
626 -> http://akrotea.ch
627 -> http://skilllauncher.com
628 -> http://www.orbitsoftltd.com
629 -> http://www.led4light.co.uk
630 -> http://golden-radiance.co.uk
631 -> http://greyfinch.com
632 -> http://dentalclinicaavanzada.com
633 -> http://abracadabracoffeeco.com
634 -> http://www.coaching-institutes.net
635 -> http://vikasgargco.icai.org.in
636 -> http://www.edinburghthistlehotel.com
637 -> http://wohlertdemexico.com
638 -> http://www.ruskin-privatehire.com
639 -> http://www.aimsun.com
640 -> http://raytell.co.uk
641 -> http://www.isranientertainment.com
642 -> http://www.alter-fritz.com
643 -> http://expertsupplydelcentro.com
644 -> http://www.planwithinspire.com
645 -> http://bluebellhotel.com
646 -> http://surmanja.com
647 -> http://lpbuildingservices.co.uk
648 -> http://www.theclaimsspot.com
649 -> http://www.mychina.org
650 -> http://LonnieHohl.com
651 -> http://www.singerynagoya.com
652 -> http://fgenergy.pt
653 -> http://gxhongxin.cn
654 -> http://www.blackboxdev.com
655 -> http://www.stonehouseproperty.co.uk
656 -> http://pcaudiovisual.com
657 -> http://gebrueder-schmidt.de
658 -> http://galariamd.com
659 -> http://haustechnik-eckert.de

660 -> http://legastchocolatier.com
661 -> http://awesomebazar.com
662 -> http://thewirelessbox.com
663 -> http://travel-center.co
664 -> http://www.cantinadeipapi.com
665 -> http://translatoruk.co.uk
666 -> http://www.sc-zxkj.com
667 -> http://lambonline.org
668 -> http://chocamama.com
669 -> http://vfl-herrenberg.de
670 -> http://www.castillosecurityservices.com
671 -> http://rrg.com.br
672 -> http://www.quikhiring.com
673 -> http://www.remindacap.com
674 -> http://malaga.hammamalandalus.com
675 -> http://grukom.de
676 -> http://www.cafler.com
677 -> http://www.silc.com
678 -> http://www.sherpashaw-golf.co.uk
679 -> http://www.facebook.com
680 -> http://volvotrucks.lt
681 -> http://www.mirarobot.com
682 -> http://bordeaux-fete-le-fleuve.com
683 -> http://www.dontworry.com.mx
684 -> http://guarderiachupetin.es
685 -> http://acampamentovaledasaguas.com.br
686 -> http://mama.eu
687 -> http://huskyadventure.no
688 -> http://padarnalarms.com
689 -> http://pappaya.com
690 -> http://machinedtechnologycnc.com
691 -> http://www.losttavernbrewing.com
692 -> http://digital-jukeboxes.com
693 -> http://aeichelbaum.de
694 -> http://www.onenottingham.org.uk
695 -> http://alu-unna.de
696 -> http://aragonesa.net
697 -> http://www.zahnarzt-dr-warnick.com
698 -> http://barbqplaza.com
699 -> http://saluswellness.ca
700 -> http://www.itzaarchaeology.com
701 -> http://www.uitdefileaanhetwerk.nl
702 -> http://bkk-bpw.de
703 -> http://metro-constructions.com.au
704 -> http://hofoled.en.alibaba.com
705 -> http://www.dowaysoft.com
706 -> http://marksmenaquatic.com
707 -> http://www.indiamart.com
708 -> http://hoyer-architekten.de
709 -> http://thedevonport.com
710 -> http://www.fashionid.de
711 -> http://circa-art.com
712 -> http://wrecords.pt
713 -> http://www.fo-cus.nl
714 -> http://biscuiterieantoine.com
715 -> http://www.babur.info
716 -> http://www.artejewellery.com
717 -> http://gzeal.co.jp
718 -> http://www.deverslist.com.au
719 -> http://shoppingdoscaminhoes.com.br

720 -> http://dealab.it
721 -> http://www.stewartbuilderskc.com
722 -> http://ecafez.tatasteel.co.in
723 -> http://oces-group.com
724 -> http://hovelagoon.co.uk
725 -> http://www.itwreddipac.com
726 -> http://TheWordChoice.com
727 -> http://www.tomesen.com
728 -> http://agk-world.com
729 -> http://www.pearls-a.com
730 -> http://ateca-sl.com
731 -> http://warongbrasil.com.br
732 -> http://veterinariasantana.com.br
733 -> http://www.ifcgegypt.com
734 -> http://www.huake3d.com
735 -> http://seifert-immo.de
736 -> http://cevo.cat
737 -> http://kalishop.de
738 -> http://keltonarapiraca.com.br
739 -> http://southernautomatics.co.nz
740 -> http://www.weinlabor-braun.de
741 -> http://www.pinnaclemtgcorp.com
742 -> http://bioentorno.org
743 -> http://g.co
744 -> http://visionaryeyecentre.com
745 -> http://superiorsummit.com
746 -> http://sendbox.com.ar
747 -> http://www.idyllwineco.com.au
748 -> http://www.expressionsinsilk.com
749 -> http://royaltonsuites.com
750 -> http://poffernier.nl
751 -> http://gx181.com
752 -> http://www.gepsolutions.co.uk
753 -> http://williamsconsulting.ca
754 -> http://tstsat.com
755 -> http://www.kiddycreche.fr
756 -> http://alchawmetprint.com
757 -> http://www.voaba.org
758 -> http://www.dermatologyconsultantslondon.com
759 -> http://pramym.com.br
760 -> http://www.velabikes.com
761 -> http://www.dptradeking.com
762 -> http://bbq-jungs.com
763 -> http://www.nitewerk.com
764 -> http://www.novemsol.com
765 -> http://vst-hd-weinheim.de
766 -> http://www.ralcotyre.com
767 -> http://innoting.eu
768 -> http://cg-frohnhausen.de
769 -> http://www.tormaco.com
770 -> http://pass-sport-control.co.uk
771 -> http://www.thelatinamericagroup.com
772 -> http://www.cambridge-hr.com
773 -> http://theoakslakes.co.uk
774 -> http://www.xilinglab.com
775 -> http://www.maryleboneproperties.com
776 -> http://www.priceblocks.com
777 -> http://huddlecreative.com
778 -> http://worldreliefmoline.org
779 -> http://jacmotors.by

780 -> http://armylegalservices.co.uk
781 -> http://www.indiamart.com
782 -> http://www.camelotinvestigations.com
783 -> http://portosystem.com.br
784 -> http://www.palaisbelvedere.org
785 -> http://valleyyouthrugby.com
786 -> http://uvidi.com
787 -> http://updeed.co
788 -> http://consultsys.com.br
789 -> http://rottalinnkliniken.de
790 -> http://camiseteca.com.br
791 -> http://dimmer7.com.br
792 -> http://www.his999.com
793 -> http://www.cleem.com
794 -> http://www.projectcasting.com
795 -> http://ambrusconstruction.com
796 -> http://tvsmalhas.com.br
797 -> http://hpwf.co.uk
798 -> http://weilerandco.com
799 -> http://www.liveonbiolabs.com
800 -> http://doubleyolk.com
801 -> http://pebblesrockandstone.com.au
802 -> http://therapyskincare.co.uk
803 -> http://classicski.co.uk
804 -> http://milestonetires.com
805 -> http://signalmash.com
806 -> http://www.mlrarchitecture.com
807 -> http://budapestmedical.eu
808 -> http://sbwl.in
809 -> http://weldingandrefurbsltd.co.uk
810 -> http://www.gallowayfiresafetyservices.co.uk
811 -> http://www.schefenacker-druck.de
812 -> http://eyeboltindia.com
813 -> http://sigal-yoga.com
814 -> http://www.balthasar-neumann.com
815 -> http://awantu.placestars.us
816 -> http://peakpainting.com.au
817 -> http://www.WhiteHouseImpressions.com
818 -> http://www.suncarrieromega.com
819 -> http://harveyag.wa.edu.au
820 -> http://geekfam.asia
821 -> http://dynamicauto.com.au
822 -> http://monumentcapitalgroup.com
823 -> http://www.cosmeceutique.com
824 -> http://www.melsherwood.com
825 -> http://www.thirdplacedesigncoop.com
826 -> http://www.kwpalmetto.com
827 -> http://strgrp.com.au
828 -> http://www.blutspende-kassel.de
829 -> http://www.charliesoap.ca
830 -> http://www.automakler-schulz.de
831 -> http://www.mommaeat.co.kr
832 -> http://www.bdrprofitcoach.com
833 -> http://kick-film.de
834 -> http://gays.com
835 -> http://www.anomoz.com
836 -> http://www.propic.com.au
837 -> http://malteser-buxtehude.de
838 -> http://www.fhp-design.de
839 -> http://ladivinamisericordia.edu.pe

```
840 -> http://heilpraktiker-rosenheim.de
841 -> http://ipov.pt
842 -> http://www.elektro-niehoff.de
843 -> http://www.wilkesgreenhill.co.uk
844 -> http://icones8.fr
845 -> http://monmouthfabrics.com
846 -> http://sanijato.pt
847 -> http://lapraim.com
848 -> http://www.maflow.com
849 -> http://www.idjardin.com
850 -> http://www.accurecord-direct.com
851 -> http://zoskinhealth.com
852 -> http://rivanto.be
853 -> http://www.streblgreencarbon.com
854 -> http://reche.cl
855 -> http://bushchemist.com.au
856 -> http://brandtronik.de
857 -> http://antoniomarazuelallorente.es
858 -> http://www.harthill-village.com
859 -> http://www.designworlds.com
860 -> http://techsparkacademy.ch
861 -> http://www.cotsmech.co.uk
862 -> http://scrubscardetailing.com.au
863 -> http://www.bb-stuck.de
864 -> http://www.qualityquickprint.com
865 -> http://reefhotelgladstone.com.au
866 -> http://www.skyemeadowsweet.com
867 -> http://chemicalrecycling.com.au
868 -> http://buggenhoutlacrosse.be
c:\Users\jbapt\Documents\ALPHA10X\Web-segmenter\utils.py:216: MarkupResemblesLoca
torWarning: The input looks more like a filename than markup. You may want to ope
n this file and pass the filehandle into Beautiful Soup.
  soup = BeautifulSoup(r.text, 'html.parser').body
```

869 -> http://kmwloaders.com
870 -> http://bodiesbyryan.com.au
871 -> http://www.beechhillgrange.co.uk
872 -> http://britenglishschool.com
873 -> http://kaiserwetter.eu
874 -> http://www.bcsmechanical.com
875 -> http://identisys.co.uk
876 -> http://www.aeroquipcu.com
877 -> http://greenmountainimports.com
878 -> http://thestockbuyer.com
879 -> http://www.mayenexpress.com
880 -> http://www.deula-warendorf.de
881 -> http://www.speakeasytheatre.co.uk
882 -> http://www.facebook.com
883 -> http://angerer-steuerberatung.at
884 -> http://www.namastaysober.com
885 -> http://catalog.kita.org
886 -> http://www.stimmel-sports.de
887 -> http://www.ceifx.com
888 -> http://www.elite.link
889 -> http://www.mkann.com
890 -> http://pandiansurfactants.com
891 -> http://www.tuskastl.de
892 -> http://aromacleanliving.com.au
893 -> http://chingfordhouseschool.co.uk
894 -> http://www.profi-servicewerkstatt.de
895 -> http://www.magda-bittner-simmet-stiftung.de
896 -> http://www.avamarble.com
897 -> http://promoda.ca
898 -> http://wxlhwc.com
899 -> http://www.esstuning.de
900 -> http://www.wrightproductions.com
901 -> http://cybershield.com.au
902 -> http://www.ibrainbaby.com
903 -> http://www.waunarwfarmlivery.co.uk
904 -> http://www.sarvodaya.nl
905 -> http://www.houbensteyngroep.nl
906 -> http://vitroplant.at
907 -> http://coastskate.com.au
908 -> http://spgc.com.au
909 -> http://www.fesskobbi.com.br
910 -> http://www.talestech.com
911 -> http://aderhelsclub.com
912 -> http://www.directproduction.net
913 -> http://sundmedia.se
914 -> http://leclochardcafe.com
915 -> http://wabionrose.com
916 -> http://bellaflorafloricultura.negocio.site
917 -> http://www.hotelyastrebets.bg
918 -> http://www.hitechled.cn
919 -> http://datahouse.asia
920 -> http://www.digitalid.co.uk
921 -> http://atoshi.org
922 -> http://www.test.de
923 -> http://franciscoconte.com.br
924 -> http://asadorsanmartin.com
925 -> http://www.hugoetlia.com
926 -> http://kraeton.com
927 -> http://echostec.com.br
928 -> http://www.cpusoftware.de

929 -> http://cesvov.it
930 -> http://www.vaext.nl
931 -> http://www.yiqizou.com
932 -> http://balliagreencitydeveloperspvtlt.elisting.in
933 -> http://renaza.com
934 -> http://kubiseg.aggilizador.com.br
935 -> http://mtour-travel.de
936 -> http://www.exquisit-herrenmoden.de
937 -> http://mckennamechanical.com.au
938 -> http://www.wowkai.cn
939 -> http://www.atria-eindhoven.nl
940 -> http://yusuf-cicek-insaat-otamotivemlak.business.site
941 -> http://www.keepers.com.kw
942 -> http://paragindustries.com
943 -> http://dungcuphunxamhani.com
944 -> http://inwardbound.ca
945 -> http://moskitchenandtavern.com
946 -> http://construtorabelga.com.br
947 -> http://jameelenterprises.com
948 -> http://cayetanocifuentes.com
949 -> http://www.futurebuildsea.com
950 -> http://www.tbelec.com
951 -> http://www.pcblltd.com
952 -> http://www.clearsense.com
953 -> http://oxfordoralsurgeon.org
954 -> http://beautybybe.co.uk
955 -> http://libertyarc.org
956 -> http://magamuseu.org
957 -> http://simasajans.com
958 -> http://www.baman.club
959 -> http://ontrackstudio.online
960 -> http://angelcalcados.business.site
961 -> http://www.diagnosticrobotics.com
962 -> http://www.rench-chemie.de
963 -> http://ronaldoeassociados.com.br
964 -> http://www.lametallerie.net
965 -> http://www.teleclinic.com
966 -> http://obsco.ca
967 -> http://engelbert-kaempfer-apotheke.de
968 -> http://friendsoftheblind.org
969 -> http://porthonda.ca
970 -> http://www.jardineroofingltd.co.uk
971 -> http://www.robertsonmgt.com
972 -> http://www.westfalia-hopsten.de
973 -> http://www.controlledfluidics.com
974 -> http://adairmotos.com.br
975 -> http://covered6.com
976 -> http://myebus.ca
977 -> http://cjcgroup.cl
978 -> http://www.medoranger.com
979 -> http://www.buchanantrading.com
980 -> http://ab-insurance.com.com
981 -> http://globecancer.com
982 -> http://dlm-associes.fr
983 -> http://www.surfsideraceplace.com
984 -> http://aiminggate.com
985 -> http://hotmouse.co.uk
986 -> http://ksmrig.com
987 -> http://DandMPacking.com
988 -> http://businessangel.no

```
989 -> http://lemearmazens.com.br
990 -> http://www.freshblood.com
991 -> http://www.jsrcjt.cn
992 -> http://www.isleepprogram.com
993 -> http://www.soundcrete.us
994 -> http://www.tamuconsultingclub.com
995 -> http://www.ovay.com.cn
996 -> http://allcopy.com.br
997 -> http://www.wellservicetechnology.co.uk
998 -> http://www.ciceron.com
999 -> http://www.goocampus.in
```

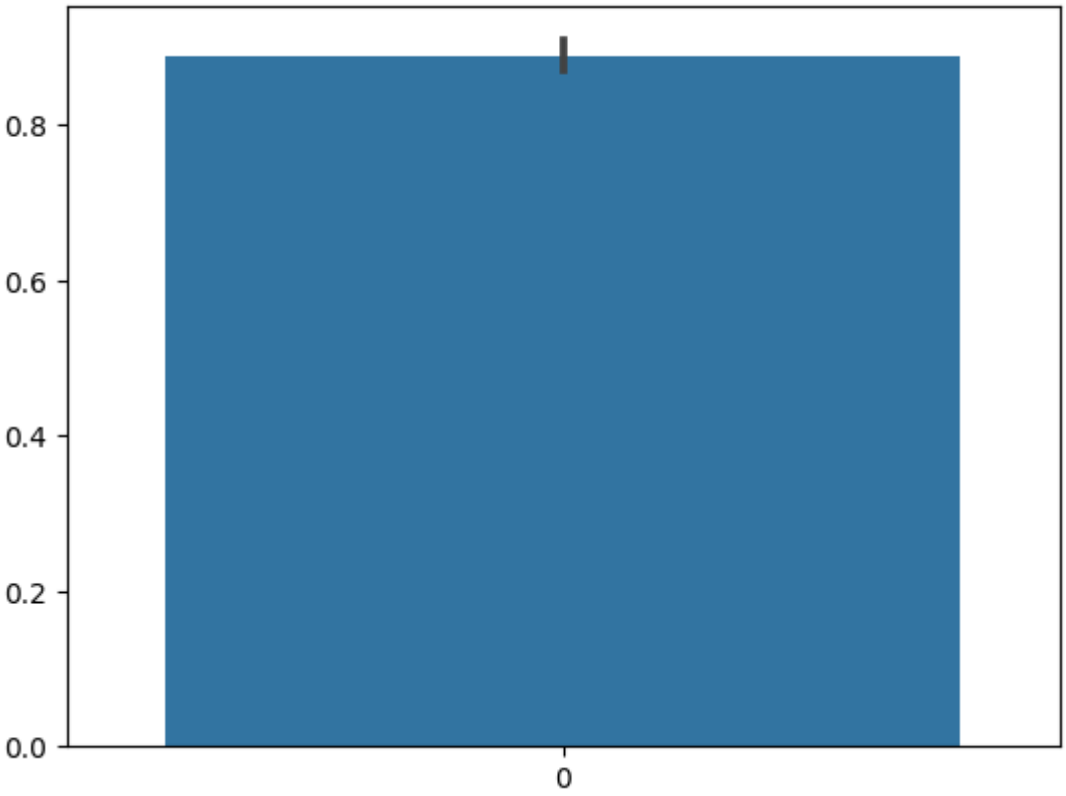
In [ ]:
```
pdf['graph'] = out
```

In [ ]:
```
pdf.to_csv(f'PE_relevant_sample_2023_05_23_with_graph_tmp_{datetime.now().strfti
```

In [ ]:
```
pdf['any_graph'] = pdf['graph'].apply(lambda x: 0 if x is None or len(x) == 0 el
```

In [ ]:
```
sns.barplot(pdf['any_graph'])
```

Out[ ]:
```
<Axes: >
```



In [ ]:
```
pdf['graph_size'] = pdf['graph'].apply(lambda x: len(x) if x is not None else 0)
```

In [ ]:
```
pdf
```

Out[ ]:

	url	content	
0	http://churchillglass.co.uk	\n ; Fast Emergency Glaziers Croydon CR0 | All...	(21aa7af207b82794bfbc02e0b561
1	http://www.ppwoniu.com	For full functionality of this site it is nece...	
2	http://www.hansecom.com	; IT + Software Lösungen für den ÖPNV - Hans...	(3b74338a7e136d2e7a4791ea3349
3	http://costimp.it	\n ; Impresa di Costruzioni Piacenza Lodi – Ri...	(698f1cae7b3208b2df9856db2d95
4	http://ortesia.com	; Ortesia Pain Relief & Skin Care CBD Produc...	(26723d5bfc738bc8c8249b49c36a
...	...	...	
995	http://www.ovay.com.cn	\n ; 首页	
996	http://allcopy.com.br	\n ; All Copy ; Abrir Chamado ; FALE COM NOSSO...	(35f49c5ca7d90ba975d48cde43b2(
997	http://www.wellservicetechnology.co.uk	\n ; WELL SERVICE TECHNOLOGY ; HOME ; PROFILE ...	(698f1cae7b3208b2df9856db2d95
998	http://www.ciceron.com	\n ; Ciceron - The Digital Agency For Brands T...	(ea33753ed70136287dbdcc0d3f60
999	http://www.goocampus.in	GooCampus | Medical career choices made simple	

1000 rows × 5 columns

In [ ]:
```
pdf.shape
```

Out[ ]: (1000, 5)

Save output to `.pickle` and `.csv`

In [ ]:
```
pickle.dump(pdf, open('PE_relevant_sample_2023_05_23_with_analysis.pkl', 'wb'))
```

In [ ]:
```
pdf.to_csv('PE_relevant_sample_2023_05_23_with_analysis.csv', index=False)
```

Analysis

Load data

In [ ]:
```
#read the pickle file
pdf = pickle.load(open('PE_relevant_sample_2023_05_23_with_analysis.pkl', 'rb'))
```

Overall raw success

In [ ]:
```
print(f"On a sample of {pdf.shape[0]} pages, we have {pdf['any_graph'].sum()} pa
      f"resulting in an average success rate of {pdf['any_graph'].mean()*100:.1f
```

```
On a sample of 1000 pages, we have 889 pages with a graph resulting in an average
success rate of 88.9%
```

Sample of unsuccesful websegmenter

In [ ]:
```
pdf[pdf['any_graph']== 0].head()
```

Out[ ]:

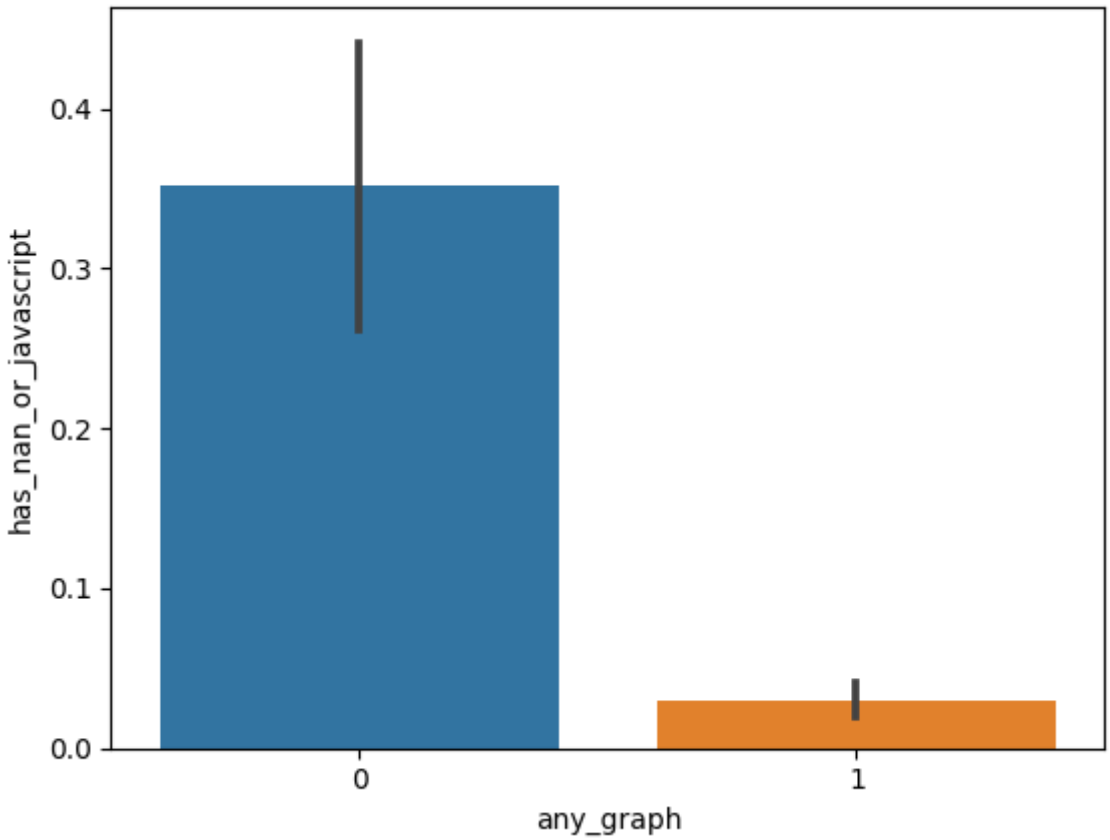
	url	content	graph	any_graph	graph_size	has_javascript
1	http://www.ppwoniu.com	For full functionality of this site it is nece...	()	0	0	1
6	http://www.phone580.com	蜂助手-以数字科技助力便捷生活 ; You need to enable JavaScrip...	()	0	0	1
17	http://bucknerbarrel.com	NaN	()	0	0	0
20	http://www.jabe.net	NaN	()	0	0	0
23	http://www.JARCO.com	\n ; Jarco | Setting the Propane Truck Standar...	()	0	0	0

In [ ]:
```
# Add flag to check for the presence of javascript in the content, this is robus
pdf['has_javascript'] = pdf['content'].apply(lambda x: 1 if 'javascript' in str(
```

In [ ]:
```
# Add flag for the presence of NaN in the content column
pdf['has_nan'] = pdf['content'].apply(lambda x: 1 if str(x).lower() == 'nan' els
```

In [ ]:
```
sns.barplot(data=pdf, x='any_graph', y='has_nan_or_javascript')
```

Out[ ]:
```
<Axes: xlabel='any_graph', ylabel='has_nan_or_javascript'>
```



A third of the unsuccessful websegmenter are due to the fact that the website is probably dynamic.

Sample of unsuccessful and not NaN or Javascript

In [ ]:
```
pdf[(pdf['any_graph']== 0)&(pdf['has_nan_or_javascript']== 0)].head()
```

Out[ ]:

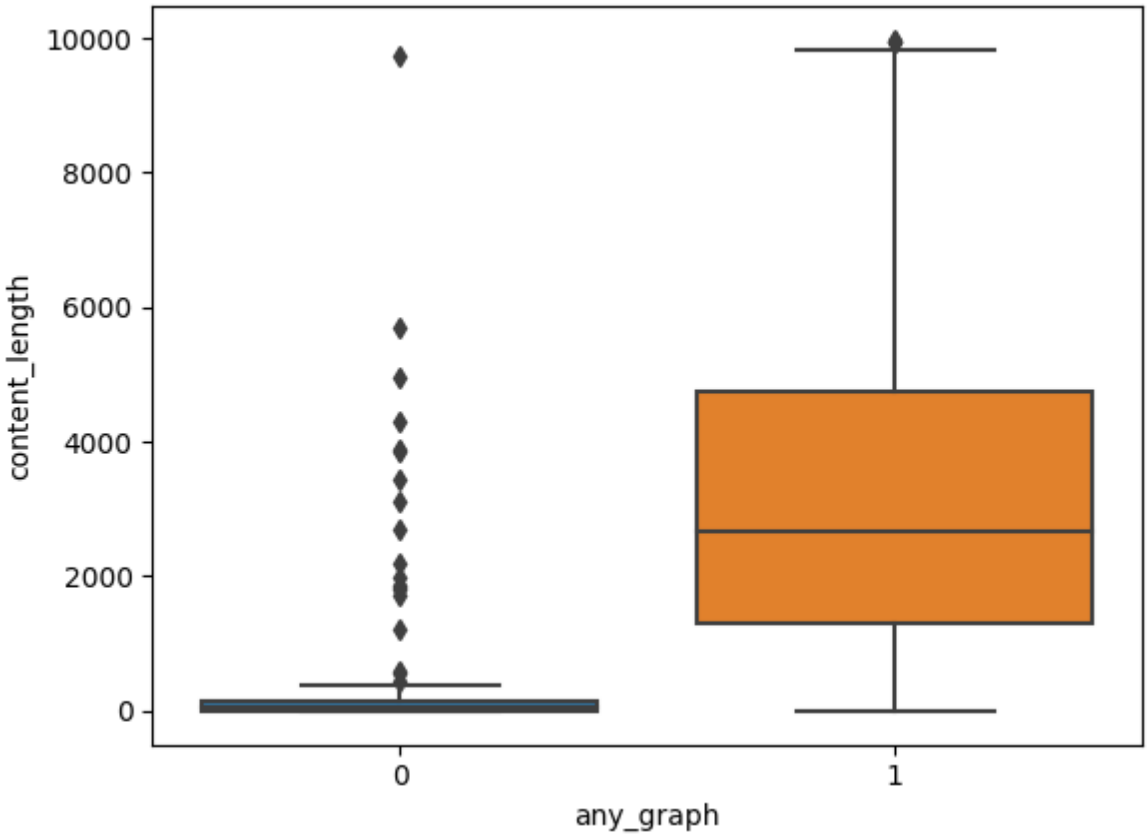
	url	content	graph	any_graph	graph_size	has_javascript
23	http://www.JARCO.com	\n ; Jarco | Setting the Propane Truck Standar...	()	0	0	0
36	http://www.ndxcards.com	\n ; \n	()	0	0	0
42	http://www.thevictorcloset.com	\n ; The House of Victor Official Global Bouti...	()	0	0	0
67	http://maistro.ru	\n	()	0	0	0
77	http://yoalearning.com	約克郡線上學院 YOA Learning ;	()	0	0	0

In [ ]:
```
pdf['content_length'] = pdf['content'].apply(lambda x: len(str(x)) if str(x).low
```

In [ ]:
```
# normalize the content length by the sum of the content length per any_graph gr
pdf['content_length'] = pdf.groupby('any_graph')['content_length'].apply(lambda
```

In [ ]:
```
sns.boxplot(data=pdf[pdf['content_length']<10_000], y='content_length', x='any_g
```

Out[ ]:
```
<Axes: xlabel='any_graph', ylabel='content_length'>
```



The unsuccessfull websegmenter seems to be for content of lower size.

Samples of of unsuccessful and not NaN or Javascript and long enough length

In [ ]:
```
pdf[(pdf['any_graph']== 0)&(pdf['has_nan_or_javascript']== 0)&(pdf['content_leng
```

Out[ ]:
```
(25, 10)
```

We have 25 samples to manually review/

In [ ]:
```
pdf[(pdf['any_graph']== 0)&(pdf['has_nan_or_javascript']== 0)&(pdf['content_leng
```

Out[ ]:

	url	content	graph	any_graph	grap
23	http://www.JARCO.com	\n ; Jarco | Setting the Propane Truck Standar...	()	0	
42	http://www.thevictorcloset.com	\n ; The House of Victor Official Global Bouti...	()	0	
107	http://www.synerbuild.com	\n ; Home | synerBuild | San Francisco Project...	()	0	
109	http://bishvilaych.org	; Bishvilech Center | Medical Center for Wom...	()	0	
156	http://blackstoreboutique.negocio.site	BlackStore.Boutique - Vestuário/marcas ; Black...	()	0	
191	http://adviesgroeppinson.be	\n ; Reconnect Your Domain | Wix.com ; top of ...	()	0	
222	http://kubernesys.com	\n ; Deprecated ; : Required parameter $id fo...	()	0	
233	http://illawarraophthalmology.com.au	\n ; Home | Illawarraoptho ; top of page ; 02 ...	()	0	
237	http://joingoldstarmortgage.com	\n ; Loan Officers | United States | Join Gold...	()	0	
255	http://www.intentinterior.com	\n ; Best Interior Designer & Decorator in Cha...	()	0	
270	http://45000feet.com	\n ; 色色资源站无码-色色资源站无码AV-色色资源站无码AV网址 ; 联系电话 ; 15...	()	0	
282	http://senderoverdepr.com	\n ; 大连金州盛源建筑材料厂 ; 服务热线： ; 020-123456789 ; 网站首...	()	0	
293	http://allgardenmachines.com	\n ; Tractor Mower Manufacturer,Lawn Tractor M...	()	0	
308	http://invoiceplatform.com	invoiceplatform.com ; invoiceplatform.com ; Th...	()	0	

	url	content	graph	any_graph	grap
371	http://mlwlf.com	\n ; 清新县米乐舞旅行之家 - mlwlf.com ; 清新县米乐舞旅行之家 - mlw...	()	0	
407	http://mtgravattsubaru.com.au	Zupps Mt Gravatt Subaru | Subaru Dealer Mt Gra...	()	0	
442	http://www.templelogic.com	\n ; Temple Logic - Custom Software Developmen...	()	0	
473	http://www.ajyaguru.com	\n ; 阿侠谷 - 首页 ; \r\n< ; !–[if IE]>\r\n ; \r\n...	()	0	
522	http://www.softsyshosting.com	\n ; SoftSys Hosting – Latest Generation Globa...	()	0	
666	http://www.sc-zxkj.com	\n ; Warning! ; sc-zxkj.com has expired.\r\n ...	()	0	
685	http://acampamentovaledasaguas.com.br	\n ; Acampamento Vale das Águas | Sítio Para L...	()	0	
712	http://wrecords.pt	\n ; Wrecords - Recording Studios ; Services ;...	()	0	
725	http://www.itwreddipac.com	\n ; \r\n\tRegistrant WHOIS contact informatio...	()	0	
888	http://www.elite.link	elite.link ; elite.link ; This domain is avail...	()	0	
993	http://www.soundcrete.us	\n ; Sound Crete Contractors Inc. –	()	0	

In [ ]: