// ROKASGRĀMATA

SERPCTRL // Crawler

Vietnes pārmeklētājs. Jūsu datorā. Maksā vienreiz.

Screaming Frog stila vietņu pārmeklētājs. Pārmeklējiet jebkuru vietni vai pārbaudiet URL sarakstu un eksportējiet visus datus vienā XLSX failā. Vietējā darbvirsmas lietotne macOS un Windows operētājsistēmām. Cena: 50 €, vienreizējs maksājums. Bez abonementa, atjaunošanas maksas vai lietotāja konta.

// SATURS

// KAS TAS IR UN KĀ PALAIST

Kas tas ir un kā palaist

Divi darbības režīmi: "Zirnekļa" (Spider) pārmeklēšana no URL, analizējot katru iekšējo saiti, vai URL saraksta pārbaude no CSV, TSV vai TXT faila bez tālākas pārmeklēšanas. Abos gadījumos tiek ģenerēts viens XLSX fails ar vienu rindu katrai lapai. Tiks analizēts: statuss, indeksējamība, tituls, meta dati, H1 un H2 virsraksti, canonical un robots tagi, X-Robots-Tag, vārdu skaits, atbildes laiks, dziļums, ienākošās un izejošās saites, kā arī veikts shēmas (schema) audits.

Screaming Frog ir nozares standarts, bet maksā 245 € gadā un virs 500 URL aizver aiz maksas licences. SERPCTRL // Crawler nodrošina 90% no galvenajām funkcijām par 50 €, kas ir vienreizējs maksājums par ērti lietojamu darbvirsmas lietotni.

// CENA

SERPCTRL // Crawler	**50 €, vienreizējs maksājums.** Bez atjaunošanas maksas, licences faila vai lietotāja konta.
Screaming Frog	245 € gadā, vienai vietai. Bezmaksas līmenis ir ierobežots līdz 500 URL.
Sitebulb	~1330 € gadā (~110 €/mēn). Mākonī.
Ahrefs Site Audit	~2148 € gadā (179 €/mēn). Iekļauts plašākā plānā.

// UZSTĀDĪŠANA

- **macOS.** Lejupielādējiet .app failu. Pārvelciet uz Applications. Divreiz noklikšķiniet. Pirmajā reizē var vajadzēt labo klikšķi un Open, lai apietu Gatekeeper.
- **Windows.** Lejupielādējiet .exe. Divreiz noklikšķiniet. SmartScreen var brīdināt pirmajā reizē. Noklikšķiniet uz *More info*, tad *Run anyway*.
- **No avota.** pip install -r requirements.txt, tad python main.py. Vai izmantojiet ./run-dev.sh uz macOS.

// KUR NONĀK IZVADE

macOS app	~/Documents/serpctrl-crawler/
Pārējie	./crawls/ blakus izpildāmajam failam
faila nosaukums	crawl_<domēns>_<laiks>.xlsx spider režīmā, urlcheck_<domēns>_<laiks>.xlsx saraksta režīmā

// PIRMĀ PALAIŠANA 30 SEKUNDĒS

1. Atveriet lietotni.
2. Mode = Spider from URL.
3. Ielīmējiet URL.
4. Noklikšķiniet Start Crawl.
5. Pagaidiet. Noklikšķiniet Open Output Folder, kad pabeigts.

// REŽĪMS_01

Spider režīms

Noklusējuma režīms. Izvēlieties **"Spider from URL"**, ielīmējiet sākuma URL un noklikšķiniet uz **"Start Crawl"**. Crawler apseko katru iekšējo saiti, seko tām dziļumā un pārtrauc darbu, sasniedzot lapu skaita robežu (*Max Pages*).

// KAS TIEK PĀRMEKLĒTS

- Saites no <a href> katrā lapā, atrisinātas pret lapas URL.
- Tikai tā paša reģistrētā domēna saites. blog.example.com un example.com tiek uzskatīti par vienu vietni (Public Suffix List). cdn.example.org nē.
- URL tiek normalizēti. Fragmenti noņemti, host mazos burtos, beigu slīpsvītra noņemta, query saglabāts.
- robots.txt fails tiek ielādēts vienu reizi katram domēnam un saglabāts kešatmiņā. Aizliegtie ceļi tiek izlaisti un marķēti kā *"Blocked by robots.txt"* ar statusu *"0"*.

// KAD LIETOT

- Pilnas vietnes auditi, kad vēlaties iegūt katru sasniedzamo lapu.
- Crawl-budžeta pārbaudes. Uzstādiet *Max Pages* zemu, lai novērtētu apjomu pirms īstā skrējiena.
- Atklāšana. Jums nav URL saraksta, un Jūs vēlaties tādu iegūt.

// DZIĻUMS

Crawl dziļums tiek pierakstīts katrai lapai. Sākuma URL ir dziļums 0, saites no tā ir dziļums 1, un tā tālāk. XLSX tiek sakārtots pēc dziļuma, tāpēc sākumlapa un galvenās navigācijas saites atrodas pašā augšā.

// REŽĪMS_02

URL saraksta režīms

Pārslēdzieties uz režīmu **"Check URL list (CSV)"**, noklikšķiniet uz **"Choose file..."** un izvēlieties vajadzīgo failu. Crawler auditē tieši šos URL bez tālākas pārmeklēšanas vai jaunu saišu atklāšanas. Tas ir noderīgi, ja Jums jau ir zināms konkrēto lapu saraksts.

// FAILU FORMĀTI

`.csv / .tsv`	Atdalītājs tiek noteikts automātiski no , \t ; |. Galveni arī nosaka automātiski.
`.txt`	Viens URL rindā. Viss, kas nesākas ar http:// vai https://, tiek ignorēts.
`URL kolonna`	Ja ir galvene, tiek izmantota kolonna ar nosaukumu url, urls, loc, address, link, page vai līdzīgu. Citādi uzvar pirmā kolonna, kas izskatās pēc URL.
`kodējums`	UTF-8 ar vai bez BOM. Sliktie baiti tiek aizstāti, nevis fails kļūdo.
`dublikāti`	Vienādie URL tiek apvienoti, oriģinālā secība saglabāta.

// KO SAŅEM ATPAKAĻ

Tādu pašu XLSX kā spider režīmā, ar tām pašām 28 kolonnām. Vienīgā atšķirība ir tāda, ka *Inlinks* rāda nulli katrai lapai (saišu grafs nav uzbūvēts), un *Crawl Depth* vienmēr ir 0.

Kad ielādē sarakstu, *Max Pages* automātiski palielinās līdz saraksta garumam, ja tas ir uzstādīts zemāk. Noklusētie 500 nepatraukst 2000 URL sarakstu, ja vien Jūs to skaidri nesamaziniet.

// KAD LIETOT

- Zināmu URL kopas pārbaude pēc deploy.
- Lapu auditēšana no sitemap eksporta, GSC eksporta vai paša izveidota saraksta.
- URL migrācijas pārbaude. Vecie URL iekšā, jaunais XLSX ārā, skatieties statusa kodus.

// IESTATĪJUMI

Iestatījumi

Četri rokturīši virs Start pogas. Noklusējumi der lielākajai daļai vietņu. Divi, ko reāli pieskaņosiet, ir **Max Pages** un **Threads**.

Max Pages	Spider apstājas pie šī skaitļa. Saraksta režīmā automātiski tiek palielināts līdz saraksta garumam. Noklusējums **500**. Pilnam vietnes auditam uzstādiet to labi virs paredzamā lapu skaita.
Threads	Paralēlie HTTP procesi (darbinieki). Noklusējuma vērtība **5**. Palieliniet uzmanīgi; vairums serveru stabili apstrādā līdz 10 pavedieniem, mazi VPS sāk dusmoties virs tā.
Delay (ms)	Pauze starp pieprasījumiem *katram darbiniekam*. Noklusējums **0**. Uzstādiet 100 līdz 250ms vietnēm, kas Jums nepieder. 500ms+ pieklājīgajai versijai.
Respect robots.txt	Ieslēgts pēc noklusējuma. Izslēgts nozīmē, ka crawler ignorē Disallow noteikumus. Izslēdziet šo iestatījumu tikai Jums piederošām vietnēm.

// MATEMĀTIKA, KO VĒRTS ZINĀT

Efektīvais pieprasījumu ātrums = **threads ÷ delay**. Pieci darbinieki ar 200ms aizturi = 25 pieprasījumi sekundē. Desmit darbinieki bez aiztures = tik ātri, cik Jūsu CPU un mērķa serveris spēj sekot.

// ROBOTS.TXT UZVEDĪBA

- Ielādēts vienreiz uz domēnu, kešēts uz visu skrējienu.
- Aizliegtie URL tiek izlaisti un marķēti ar statusu 0 un marķieri *Blocked by robots.txt*. Tie joprojām parādās XLSX, lai redzētu, kas tika izslēgts.
- Ja robots.txt nepastāv vai atgriež 4xx/5xx, viss tiek atļauts.

// XLSX IZVADE

XLSX izvade

Rezultāts ir viena darblapa ar fiksētu galvenes rindu un ieslēgtu automātisko filtru, sakārtota pēc pārmeklēšanas dziļuma un URL. Kopā ir 28 kolonnas. Atveriet Excel, Google Sheets vai Numbers un filtrējiet.

// KOLONNU KARTE

Identitāte	URL, Status Code, Status, Redirect URL, Content Type.
Indeksējamība	Indexability (Indexable / Non-Indexable), Indexability Status (Noindex, Canonicalised, Client Error, Blocked by robots.txt utt.).
On-page	Title 1 + length, Meta Description + length, H1-1, H1-2, H2-1, H2-2, Canonical, Meta Robots, X-Robots-Tag, Word Count.
Veiktspēja	Size (bytes), Response Time (ms), Crawl Depth, Inlinks, Outlinks.
Schema	Schema Count, Schema Types, Schema Status, Schema Issues. Pilna sadaļa nākamajā lpp.

// INDEKSĒJAMĪBAS LOĢIKA

Aprēķināta no tiem pašiem signāliem, ko izmanto Screaming Frog. Statusa kods, meta robots, X-Robots-Tag un canonical pret pašu URL. Lapa ir **Non-Indexable**, kad kāds no šiem to saka, un iemesls ir *Indexability Status* kolonnā.

// INLINKS / OUTLINKS

Outlinks norāda unikālo iekšējo saišu skaitu lapā. **Inlinks** norāda to lapu skaitu, kuras satur saiti uz konkrēto URL. Inlinks tiek aprēķināts pēc crawl beigām, tāpēc kolonna aizpildās eksportā, nevis dzīvajā logā.

// SCHEMA AUDITS

Schema audits

Katrs lapas <script type="application/ld+json"> elements tiek parsēts, @graph struktūras tiek sadalītas, un katrs bloks tiek auditēts atsevišķi. XLSX nonāk četras kolonnas.

`Schema Count`	Cik JSON-LD bloku ir lapā.
`Schema Types`	Ar komatiem atdalītas @type vērtības dokumenta secībā.
`Schema Status`	**NONE** (nav JSON-LD), **OK**, **WARNINGS**, **CRITICAL** vai **PARSE ERROR**. Uzvar sliktākais līmenis no visiem blokiem.
`Schema Issues`	Saīsināts [C] kritisko un [W] brīdinājumu kopsavilkums, viens uz katru atradumu. Filtrējiet šo kolonnu, lai ātri atrastu lapas ar salauztu schema.

// KAS TIEK PĀRBAUDĪTS

Tipam specifiskas pārbaudes rich-result tipiem, ko Google reāli rāda. Article, Product, FAQPage, BreadcrumbList, Recipe, Event, LocalBusiness, VideoObject, JobPosting, Organization. Katrs saņem savus noteikumus. Obligātos laukus, ISO datuma un ilguma formātus, TitleCase pārbaudi @type laukam, HTML virkņu īpašības, relatīvus URL un tukšus failu nosaukumus.

// KO APZINĀTI NEREDZ

Audits lasa **servera renderēto** HTML, ko crawler ielādēja. JSON-LD, ko JavaScript ievieto darbības laikā, šeit neparādās. Tā ir doma. AI crawler (GPTBot, ClaudeBot, PerplexityBot, CCBot) neizpilda JavaScript, tādēļ šis audits uzrāda tieši to informāciju, kas ir pieejama šiem botiem. Ja Jūsu schema jācitē AI Overview, tai jābūt SSR dokumentā.

// VADĪBA UN PADOMI

Vadība un padomi

// POGAS

Start Crawl	Zaļā poga. Atslēdzas, kad crawl skrien.
Pause / Resume	Aptur darbiniekus, tos nenogalinot. Drošs garām crawliem, kad vajag atbrīvot bandwidth.
Stop	Sarkanā poga, kas pieprasa korektu darbības pārtraukšanu. XLSX failā tiks saglabāti visi līdz tam brīdim iegūtie dati.
Open Output Folder	Saīsne uz eksporta mapi. Strādā pēc crawl beigām.

// DZĪVAIS STATUSS

- Progress bar pildās kā *processed / max pages* (spider) vai *processed / list size* (saraksta režīms).
- Pašreizējais URL parādās zem statusa etiķetes.
- Logs ir krāsots. Zaļš 2xx, dzeltens 3xx, sarkans 4xx/5xx un kļūdas.

// PADOMI, KAS IETAUPA LAIKU

- Lielām vietnēm vispirms palaidiet ar zemu *Max Pages*, lai novērtētu apjomu. Tad palieliniet īstajam skrējienam.
- Nelaidiet threads pārāk augstu uz maziem serveriem. 5 līdz 10 parasti der. Virs tā uzmaniet 5xx kļūdas logā.
- Ja neizdodas automātiski noteikt CSV struktūru, saglabājiet URL sarakstu kā vienkāršu .txt failu (viens URL katrā rindā). Šī metode ir visdrošākā.
- Filtrējiet XLSX pēc *Schema Status* = CRITICAL, lai vienā klikšķī atrastu salauztu schema visā vietnē.
- Filtrējiet pēc *Indexability* = Non-Indexable, lai redzētu katru lapu, ko Google neindeksēs, un iemeslu kāpēc.

// ROKASGRĀMATAS BEIGAS

50 € vienreizējs maksājums. Vietēja instalācija, bez kvotām un atjaunošanas maksas. Pārmeklēšana notiek Jūsu datorā. XLSX dati pieder Jums.