Ашық көздерден ақпарат алу. Парсер бағдарламалары. Көшірілетін мәліметтерді талдау

Ашық дереккөзді зерттеу – мәліметтерді зерттеуде қалыптасқан жаңа әдіс, яғни ашық дереккөздер арқылы зерттеу жүргізу. Яғни, компьютер мен ноутбуктың алдында отырып-ақ, ғаламтордағы анық, фейк-емес ақпараттарды пайдалана отырып, зерттеу жасау бағыты.

Мұндай зерттеуді жанжал орын алған немесе дау-дамай өршіп тұрған аймақты зерттеу, коррупцияны тексеру мен ашу үшін, адамдарды іздеу мен қылмыстың бетін ашу үшін, экологиялық және тарихи зерттеулерді жасау үшін қолдануға болады.

Зерттеуге пайдалануға болатын ресурстар өте көп. Бірақ олардың негізгілерін екі үлкен топқа бөліп қарастыруға болады. Біріншісі – кең таралған материалдар болса, екіншісі – күңгірт ақпараттар.

Кең таралған материалдарға мыналар жатады:

1. Карта және спутник сервер

2. Бұқаралық ақпарат құралдары

3. Қолданушы контент (әлеуметтік желі)

1. Картаны Видео және фотоматериалдарға геолокация ретінде қолдануға болады. Экологиялық зерттеулер (өзен-көлдің тартылуы, мұздықтың еруі) жасауға да ыңғайлы. Спутниктен түсірілген суреттердің кейбірі ақылы. Ақысы $2000 дейін жетуі мүмкін.

2. Халықаралық ақпарат агенттіктері мен медиакорпорациялар әлемде орын алған оқиғаға, жағдайға байланысты жеке журналисттік зерттеулер жүргізуі мүмкін. Онда зерттеу жүргізуге мұндай корпорациялардың қауқары мен ресурсы жетерлік. Өзіміздің зерттеу жұмыстарымызға аталмыш агеттіктер тапқан ақпараттарды, инсайдерлік деректерді пайдалануымызға болады. Алған мәліметтерге міндетте түрде сілтеме жасалуы тиіс.

3. Әлеуметтік желінің қолданушылары өте көп. Секундт сайын фотосурет пен видео, жазба жариялайды. Мұнда мол дерек жатыр. Соны ақтарып керегін табу алу қажет. Соның негізінде ауқымды зерттеулер жасауға болады. Мәселен, 2014 жылы Украина аспанында Малайзияның Боинг жолаушылар ұшағы апатқа ұшырап, 300-ге жуық адам қайтыс болады. Қызығы сол – украин-россия соғысы қызып жатқан тұста екі елдің бірі байқаусызда әскери ұшақты тоқтатамын деп, жолаушылар ұшағын атып түсіреді. Атқан – «Бук» зымыран кешені боп шықты. Алайда бұл зымыранды қай елдің әскерилері атқаны белгісіз күйде қалды.

Осы тұста Bellingcat медиаагенттігінің тілшілері әлеуметтік желі арқылы зерттеу жасап, «Бук» зымыранының Ресейдікі екенін дәлелдейді. Ондай қорытынды жасау үшін тілшілер ұшақ құлаған аумақтың геолокациясын бөліскен фотосурет, видеоның бәрін тексеріп, зымыран тасығышты өздейді. Мұндай әскери техниканы ел-жұрт көп көре бермейтіндіктен, оны суретке түсірушілер де көп болғаны белгілі. Соны тексере келе, журналисттер «Бук» зымыранының Ресей аумағында жүргенін әрі Ресейдің әскери базасына әкелінгенін фотосуреттер арқылы дәлелдейді. Міне бұл – әлеуметтік желі арқылы журналисттік зерттеу жасаудың айқын көрінісі.

Companies House

LexisNexis®

INTERNET ARCHIVE

4. Деректер базасы немесе мұрағат материалдары. Google ebooks түрлі құжат пен кітаптың ішінен мәтін іздеуге, Companies House тіркелген заңды тұлғаларды табуға болады. **Companies House**-тың тағы бір артықшылығы – оффшорға тіркелген компанияларды да көрсете алады. Lexis Nexis – жекеменшік деректер қоры.

Қазақстанда да заңды және жеке тұлғалардың деректерін қарауға мүмкіндік беретін Egov, Учет.кз, tenderplus.kz, goszakup.gov.kz сынды ресурстар бар.

Келесі интернет ресурстарды «күңгірт ақпарат көздері» деуге болады. Мұндағы ақпараттар зерттеу жасауға жарамды. Бірақ этикалық жағы күмәнді. Өйткені кей басылымдар мұндайды этика мен мәдениетке жат деп қолданбайды. Ал кейбір басылымдар қолдана береді. Өйткені ол да ғаламтордағы ақпарат болғандықтан, зерттеуге көмегін тигізеді деп санайды.

Тарап кеткен ақпарат

Контакт кітапшасы

Қорғалмаған деректер базасы

Ендеше «күңгірт ақпарат көздерін» қарастырып көрейік.

Біріншісі, ақпарат жүйесін **хакерлер** бұзып, содан тарап кеткен деректер. Torrent немесе Telegram арқылы жарияланған, Vikileaks, Dehashed платформаларына шыққан материалдар.

Екіншісі, **конакт-кітапша.** Getcontact белгілі бір телефон нөмірдің иесі кім екенін анықтауға көмектеседі. Бұл қосымшаны ұялы телефоныңызға жүктесеңіз, сіздің де контакт-кітапшаңыздағы телефон нөмірлері жалпы базаға енгізіледі. Сондықтан журналист ретінде оны өзіңіздің телефоныңызға жүктемей, жаңа сатып алынған смартфонға немесе контактісі жоқ телефондарға жүктегеніңіз ыңғайлы.

Үшіншісі, а**шық дереккөздер.** Азаматтар білместікпен немесе ұмытып кетіп ашық қалдырған дереккөздер жатады.

Парсинг әдетте үлкен көлемде деректерді тез жинау қажет болған кезде қолданылады. Ол арнайы парсер қызметтерінің көмегімен жүзеге асырылады.

Парсинг - бұл деректерді автоматты түрде жинау және құрылымдау процесі.

Арнайы бағдарламалар немесе парсер қызметтері сайтқа кіріп, берілген шартқа сәйкес келетін деректерді жинайды.

Парсинг жасау-процесті автоматтандыратын арнайы бағдарламалардың көмегімен белгілі бір сайттарда орналастырылған ақпаратты жинау және жүйелеу.

Қарапайым мысал-әлеуетті серіктестердің байланыстарын жинау керек делік. Мұны қолмен жасауға болады. Әр сайтқа кіріп, "контактілер" бөлімін іздеу керек, телефонды жеке кестеге көшіру керек және т.б. сондықтан әр сайтқа бес-жеті минут кетеді. Бірақ бұл процесті автоматтандыруға болады. Парсинг шарттарын орнатсаңыз біраз уақыттан кейін Сіз сайттар мен телефондардың тізімі бар дайын кесте аласыз.

Парсингтың артықшылығы айқын — егер сіз оны қолмен жинау және деректерді сұрыптаумен салыстырсаңыз:

- сіз деректерді тез аласыз;
- таңдама жасау үшін ондаған параметрлерді орнатуға болады;
- есепте қателер болмайды;
- парсингты белгілі бір жиілікпен орнатуға болады-мысалы, әр дүйсенбі сайын деректерді жинау;
- көптеген парсерлер деректерді жинап қана қоймай, сайттағы қателерді қалай түзетуге кеңес береді.

Желіде парсинг үшін көптеген шешімдер бар. Олар "бұлтта" немесе "қорапта"болуы мүмкін:

- бұлтты нұсқа — SaaS, Сізге тіркелу керек және қызметпен тікелей браузерде жұмыс істеу керек;
- қорап нұсқасы-бұл компьютерге орнатылып, онымен бағдарлама терезесінде жұмыс істеу керек шешім.

Екі жағдайда да сіз парсерге кіру үшін біраз уақыт қолжетімділік үшін төлейсіз. Мысалы, ай, жыл немесе бірнеше жылға жазылуға болады.

Парсингті қолдану әдістері

Парсингтың қолданылуын қысқаша екі мақсат түрінде көрсетуге болады:

- бәсекелестердің қалай жұмыс істейтінін жақсы түсіну және олардан қандай да бір тәсілдерді алу үшін парсинг жасау;
- қателерді жою, өзгерістерді тез енгізу және т. б. үшін жеке алаңды талдау.

Парсингты қолдану заңды ма?

Парсинг дегеніміз не екенін анықтағаннан кейін, бұл қолданыстағы заңнама нормаларына сәйкес келмейтін нәрсе сияқты көрінуі мүмкін. Шын мәнінде, бұл олай емес. Парсинг заңмен қудаланбайды. Бірақ келесі амалдарға тыйым салынған:

- сайтты бұзу (яғни пайдаланушылардың жеке кабинеттерінің деректерін алу және т. б.);
- DDOS-шабуылдар (Егер деректерді талдау нәтижесінде сайтқа тым жоғары жүктеме түссе);
- авторлық контентті қарызға алу (копирайты бар фотосуреттер, түпнұсқалығы нотариуспен расталған бірегей мәтіндер және т.б. олардың заңды орнында қалдырған дұрыс).
- Егер көпшілікке қол жетімді ақпарат жинауға қатысты болса, онда парсинг жасау заңды. Яғни, қолмен жинауға болатын барлық нәрсеге парсинг жасауға болады.

Парсинг келесі мақсаттар үшін қолданылады:

- Баға саясатын талдау. Нарықтағы белгілі бір тауарлардың орташа құнын түсіну үшін бәсекелестер туралы мәліметтерді пайдалану ыңғайлы. Алайда, егер бұл жүздеген және мыңдаған позициялар болса, оларды қолмен жинау мүмкін емес.
- Өзгерістерді бақылау. Талдау тұрақты негізде жүзеге асырылуы мүмкін, мысалы, апта сайын, нарықтағы орташа бағаның не өскенін және бәсекелестерде қандай жаңалықтар пайда болғанын анықтайды.
- Өз сайтында тәртіп орнату. Ия, бұл да мүмкін. Интернет-дүкенде бірнеше мың өнім болса, тіпті қажет. Жоқ беттерді, дубльдерді, толық емес сипаттаманы, белгілі бір сипаттамалардың болмауын немесе сайтта көрсетілгенге сәйкес келмейтін деректерді табыңыз. Парсермен аталған амалдарды тезірек орындауға болады.
- Әлеуетті клиенттер базасын алу. Мысалы, белгілі бір салада және қалада шешім қабылдаушылардың тізімін жасауға байланысты талдау бар. Ол үшін өзекті және мұрағаттық түйіндемелерге қол жеткізумен жұмыс іздеу сайттарында жеке кабинет қолданылуы мүмкін. Мұндай базаны одан әрі пайдалану этикасын әр компания өзі анықтайды.

Қандай ақпаратқа парсинг жасауға болады?

Сіз веб-сайттағы барлық нәрсені ашық қол жетімді ете аласыз. Көбінесе:

- тауарлардың атаулары мен санаттары;
- негізгі сипаттамалары;
- бағасы;
- акциялар мен жаңалықтар туралы ақпарат;
- тауарларды кейіннен "өздері үшін" өзгерту үшін сипаттау мәтіндері және т. б.
- Сайттардағы суреттерді техникалық жағынан да парсинг жасауға болады, бірақ жоғарыда айтылғандай, егер олар авторлық құқықпен қорғалған болса, қажет емес. Бөтен сайттардан олардың жеке кабинеттеріне енгізген пайдаланушыларының жеке деректерін жинауға болмайды.

Деректерге қалай парсинг жасалады?

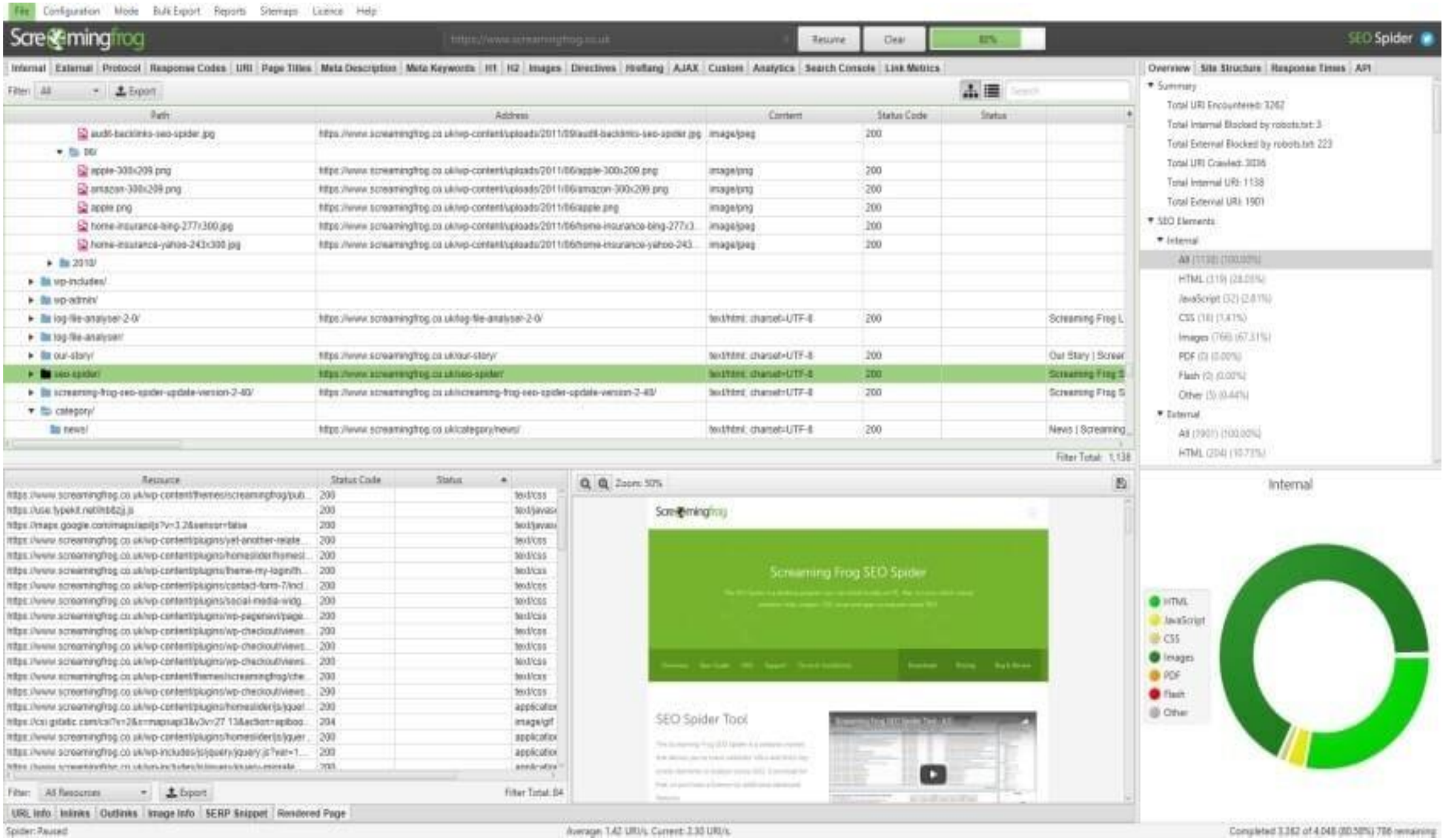
Деректерге парсинг жасау үшін екі форматтың біреуін таңдауға болады:

- нарықта көп кездесетін арнайы бағдарламаларды пайдаланыңыз;
- оларды өзіңіз жазыңыз. Ол үшін кез-келген бағдарламалау тілін қолдануға болады, мысалы, PHP, C++, Python/

Егер парақтағы барлық ақпарат қажет болмаса, бірақ тек белгілі бір нәрсе (тауарлардың атаулары, сипаттамалары, бағасы) XPath қолданылады.

Screaming Frog SEO Spider

Screaming Frog SEO Spider - SEO деректерімен жұмыс істеуге мамандандырылған танымал бағдарлама, сайттарды тексеруге арналған кең функционалдылыққа ие, бағдарлама мүмкіндіктерінің толық тізімі бірнеше бетті алады. Алғашқы танысу кезінде бағдарламаның интерфейсі көптеген қойындылар мен терезелерге байланысты қиын болып көрінуі мүмкін, бірақ біраз уақыт жұмыс істегеннен кейін оның ыңғайлы екендігі белгілі болады, қойындылар есептерге жылдам қол жеткізуге мүмкіндік береді, терезелер талдау нәтижелерін ыңғайлы түрде құрылымдайды.


File Configuration Mode Bulk Export Reports Sitemaps Licence Help
Screaming Frog
Resume
Clear
SEO Spider
Internal External Protocol Response Codes URI Page Titles Meta Description Meta Keywords H1 H2 Images Directives Hreflang AJAX Custom Analytics Search Console Link Metrics
Overview Site Structure Response Times API
Path Address Content Status Code Status
audit-backlinks-seo-spider.jpg image/jpeg 200
/06/
apple-300x209.png image/png 200
amazon-300x209.png image/png 200
apple.png image/png 200
home-insurance-bing-277x300.jpg image/jpeg 200
home-insurance-yahoo-243x300.jpg image/jpeg 200
2010/
wp-includes/
wp-admin/
log-file-analyser-2-0/ text/html; charset=UTF-8 200 Screaming Frog L
log-file-analyser/
our-story/ text/html; charset=UTF-8 200 Our Story | Screar
seo-spider/ text/html; charset=UTF-8 200 Screaming Frog S
screaming-frog-seo-spider-update-version-2-40/ text/html; charset=UTF-8 200 Screaming Frog S
category/
news/ text/html; charset=UTF-8 200 News | Screaming
Filter Total: 1,138
Summary
Total URI Encountered: 3262
Total Internal Blocked by robots.txt: 3
Total External Blocked by robots.txt: 223
Total URI Crawled: 3036
Total Internal URI: 1138
Total External URI: 1901
SEO Elements
Internal
All (1138) (100.00%)
HTML (319) (28.05%)
JavaScript (32) (2.81%)
CSS (16) (1.41%)
Images (766) (67.31%)
PDF (2) (0.00%)
Flash (2) (0.00%)
Other (5) (0.44%)
External
All (1901) (100.00%)
HTML (204) (10.73%)
Resource Status Code Status
https://www.screamingfrog.co.uk/wp-content/themes/screamingfrog/pub... 200 text/css
https://use.typekit.net/nbt6zjj.js 200 text/javas
https://maps.google.com/maps/api/js?v=3.2&sensor=false 200 text/javas
https://www.screamingfrog.co.uk/wp-content/plugins/yet-another-relate... 200 text/css
https://www.screamingfrog.co.uk/wp-content/plugins/homeslider/homesl... 200 text/css
https://www.screamingfrog.co.uk/wp-content/plugins/theme-my-login/th... 200 text/css
https://www.screamingfrog.co.uk/wp-content/plugins/contact-form-7/incl... 200 text/css
https://www.screamingfrog.co.uk/wp-content/plugins/social-media-widg... 200 text/css
https://www.screamingfrog.co.uk/wp-content/plugins/wp-pagenavi/page... 200 text/css
https://www.screamingfrog.co.uk/wp-content/plugins/wp-checkout/views... 200 text/css
https://www.screamingfrog.co.uk/wp-content/plugins/wp-checkout/views... 200 text/css
https://www.screamingfrog.co.uk/wp-content/plugins/wp-checkout/views... 200 text/css
https://www.screamingfrog.co.uk/wp-content/themes/screamingfrog/che... 200 text/css
https://www.screamingfrog.co.uk/wp-content/plugins/wp-checkout/views... 200 text/css
https://www.screamingfrog.co.uk/wp-content/plugins/homeslider/js/jquer... 200 applicatio
https://csi.gstatic.com/csi?v=2&s=mapsapi3&v3v=27.13&action=apiloa... 204 image/gif
https://www.screamingfrog.co.uk/wp-content/plugins/homeslider/js/jquer... 200 applicatio
https://www.screamingfrog.co.uk/wp-includes/js/jquery/jquery.js?ver=1... 200 applicatio
Filter: All Resources
Export
Filter Total: 84
Zoom: 50%
Screaming Frog SEO Spider
SEO Spider Tool
Internal
HTML
JavaScript
CSS
Images
PDF
Flash
Other
URL Info Inlinks Outlinks Image Info SERP Snippet Rendered Page
Spider: Paused
Average: 1.42 URI/s. Current: 2.30 URI/s.
Completed 3,262 of 4,048 (80.58%) 786 remaining

Easy Web Extract

Easy Web Extract қарапайым және күрделі сайттардан деректерді жинауға мүмкіндік беретін көптеген мүмкіндіктерді ұсынады. Бағдарлама деректерді жинауды конфигурациялау үшін терең бағдарламалау білімін қажет етпейді, арнайы шебер сізге талдау үлгісін орнату қадамдарын басшылыққа алады және орнатуды тез шешу үшін бейне сабақтар бар. Бір ерекшелігі-сіз белгілі бір өнімдерді автоматты түрде іздеуді бағдарламалай аласыз және тек қажетті деректерді жинай аласыз. Бағдарламаның тағы бір ерекшелігі – бірнеше ағындарды жинау, 24 түрлі веб-беттерге дейін, бұл сіздің талдау уақытыңызды үнемдейді. Жылдам талдаудың кері жағы-күдікті әрекетке байланысты ір-ді веб-сайттан бұғаттау, сақ болыңыз.

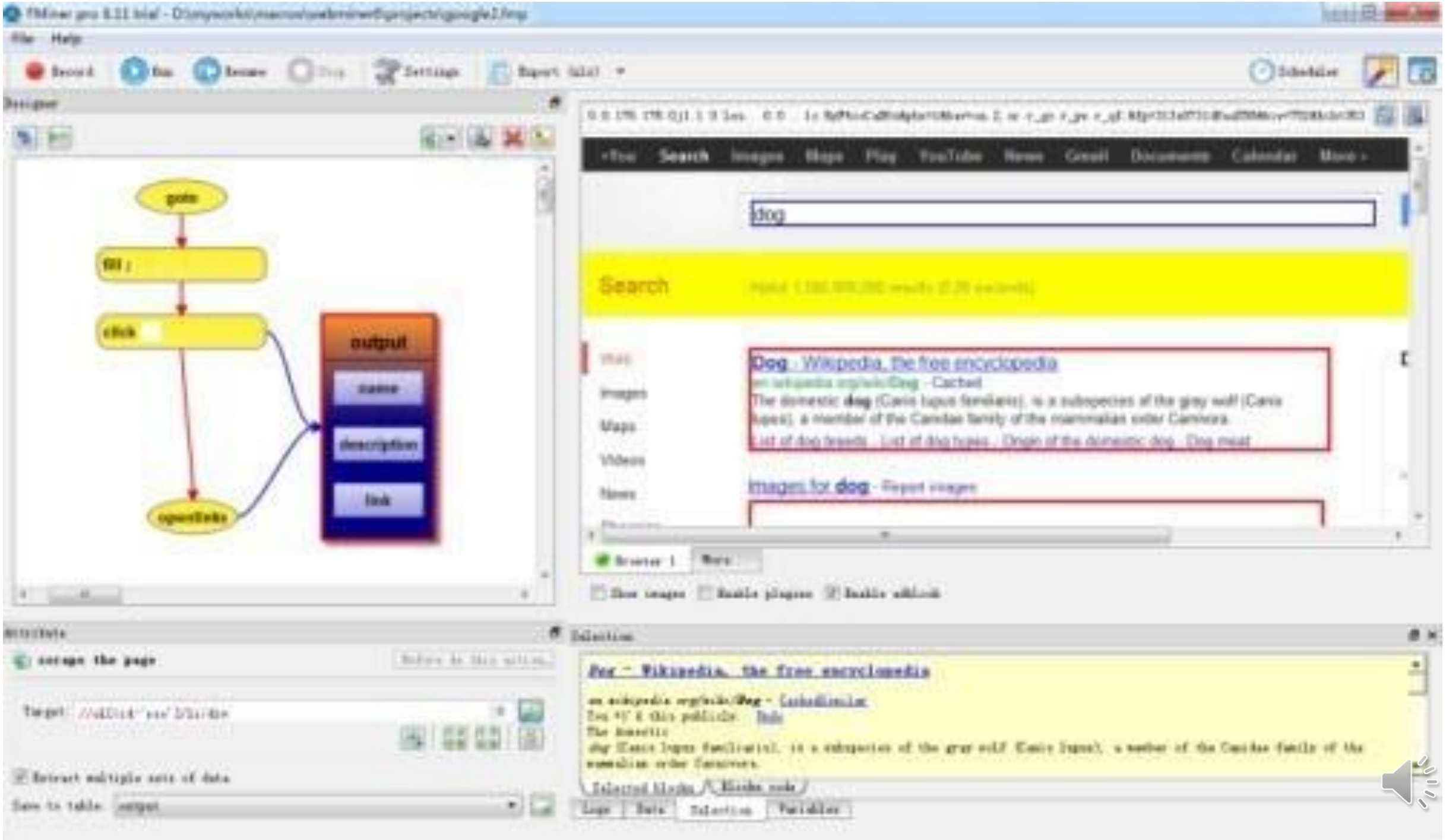
Кейбір сайттар асинхронды сұрауларды жасау үшін клиент тарапынан деректерді динамикалық жүктеу әдістерін қолданады. Мұндай деректер қарапайым талдаушылар үшін проблема болып табылады, өйткені веб-мазмұн бастапқы HTML-ге енбейді. Easy Web Extract мұндай деректерді жинау мүмкіндігі туралы мәлімдейді, тестілеу кезінде бағдарлама барлық сайттарды басқара алмады, сізге қажет сайттарда осы мүмкіндікті тексеру керек.


Project properties - Specify Extracting Pattern - Data Columns
Project
Extracting pattern
Start URL
Data
First Data Record
Data columns
Crawl rules
Startup Urls
Save file
HTML DOM
Extracting Pattern - computer - Google Product Search
Web Images Videos Maps News Shopping Gmail more
Google
computer
About 1,472,266 results (0.47 seconds)
Everything
Shopping
Computer Sale at Dell.com
www.Dell.com Find Specials on High Performance Computers w/ 2010 Intel Co
Buy Computers
www.Amazon.com/Bamboo $150 Amazon Gift Card when You Buy an ASUS B
Show only:
Google Checkout
Free shipping
New items
Any category
Desktop Computers
Servers
System & Power Cables
Computer Books
Security Cameras
More »
Any price
Under $600
$600 - $1,000
$1,000 - $2,000
Over $2,000
Enter location for tax and shipping:
ZIP or city, state
OK
Apple iMac - 4 GB RAM - 3.06 GHz - 500 GB HDD
All-in-one, Apple Mac OS X 10.6, NVIDIA GeForce 9400M, 3 MB cache - Monitor: LCD display - 21.5" - TFT active matrix - 1920 x 1080
The all-new, all-in-one iMac packs a complete, high-performance computer into a beautifully thin design. It includes built-in wireless, Mac OS X, and the iLife '09. So within ...
4.5 out of 5 401 reviews - Add to Shopping List
Dell Vostro - 230 - 3 GB RAM - 2.93 GHz - 250 GB HDD
Mini tower, Win 7 Professional, Intel GMA X4500HD, English, 3 MB cache - Monitor
From its scalable design to its ability to integrate with the Dell business hardware and services, the Vostro 230 desktop provides
Data columns
HTML Data columns
Column Name
Type
Link2Detail
URL
Name
TEXT
Price New
TEXT
Price Used
TEXT
Short Descript
TEXT
Data column properties
Name
Link2Detail
Type
URL
Link to details
Max. length
Required?
Transformation Script
HTML information
Relative DOM Path
Select another
DIV\DIV2\H3\A
Absolute DOM Path
HTML\BODY\DIV5\DIV4\DIV1\OL\LI1\DIV\DIV2\H3\A
HTML Code
Close
Preview
Back
Next
Finish
Click to select Web information as Extracted Data Column
Cancel

FMiner

FMiner-бұл сіздің әрекеттеріңізді жазуға және жазылған сценарийлерді кейіннен ойнатуға негізделген сайттарды талдау құралы. Осылайша жасалған әрекеттер тізбегі (макростар) визуалды форматта өңделеді, бұл құралды бағдарламалау тілдерін білместен пайдалануға мүмкіндік береді.

Бағдарлама динамикалық жүктелген деректермен (AJAX) жұмыс істейді, бірнеше ағынмен жұмыс істеуді қолдайды, іздеу нәтижелерімен және бірнеше шығыс форматтарымен жұмыс істеуге мүмкіндік береді. Бағдарламада жылдам бастауға арналған бейне сабақтар бар, бірақ нұсқаулықтары бар беттер жұмыс істемейді және сайттағы Соңғы жаңартулар 2015 жылдан басталады, бұл әзірлеуші өнімді бақыламайды, бірақ орнату файлдары қол жетімді және сіз екі апта ішінде бағдарламаның толық нұсқасын тегін пайдалану үшін ғимарат жүктей аласыз.

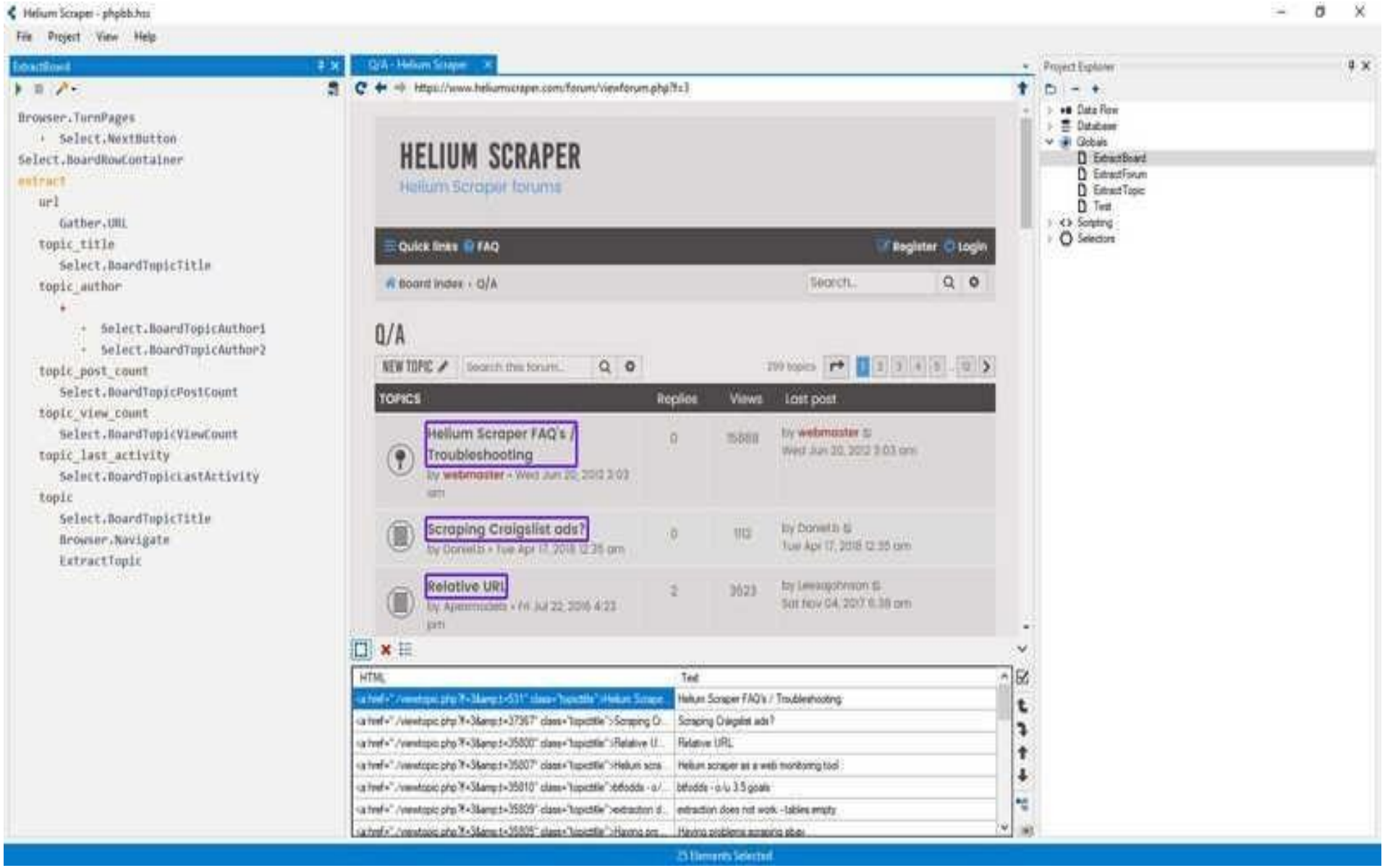

File Help
Record
Run
Browse
Stop
Settings
Export (xls)
Schedule
Designer
goto
fill :
click
output
name
description
link
Search
dog
Search
Dog - Wikipedia, the free encyclopedia
Images
Maps
Videos
News
Images for dog - Report images
Show images
Enable plugins
Enable adblock
Selection
Logs
Data
Selection
Variables

Helium scraper

Helium scraper-бұл сайттардан деректерді талдауға арналған тағы бір бағдарлама. Бағдарламамен жұмыс істеу принципі FMiner-мен жұмыс істеуге ұқсас, тек Жоспарланған әрекеттерді визуалды түрде көрсетудің орнына, бағдарлама кодты шығарады. Жалпы, алғашқы танысу кезіндегі интерфейс алдыңғы бағдарламалар сияқты түсінікті емес, бірақ бағдарлама жұмыс процесінің негіздерін тез түсінуге көмектесетін бейне сабақтар мен білім базасын ұсынады.

Функционалдығы бойынша Бағдарлама жоғарыда қарастырылғанға ұқсас, бірақ бірқатар ерекшеліктері бар. Әрине, бұл басқа бағдарламалар үлкен дерекқорлармен жұмыс істеуді игермейді дегенді білдірмейді, бірақ егер сіз көптеген деректерді жин

Парсинг жасауды жоспарласаңыз, Helium scraper-ге мұқият қарау керек. Тағы бір ерекшелігі — API-мен жұмыс істеу мүмкіндігі, сіз сұраныстарды жобаңызға біріктіре аласыз.

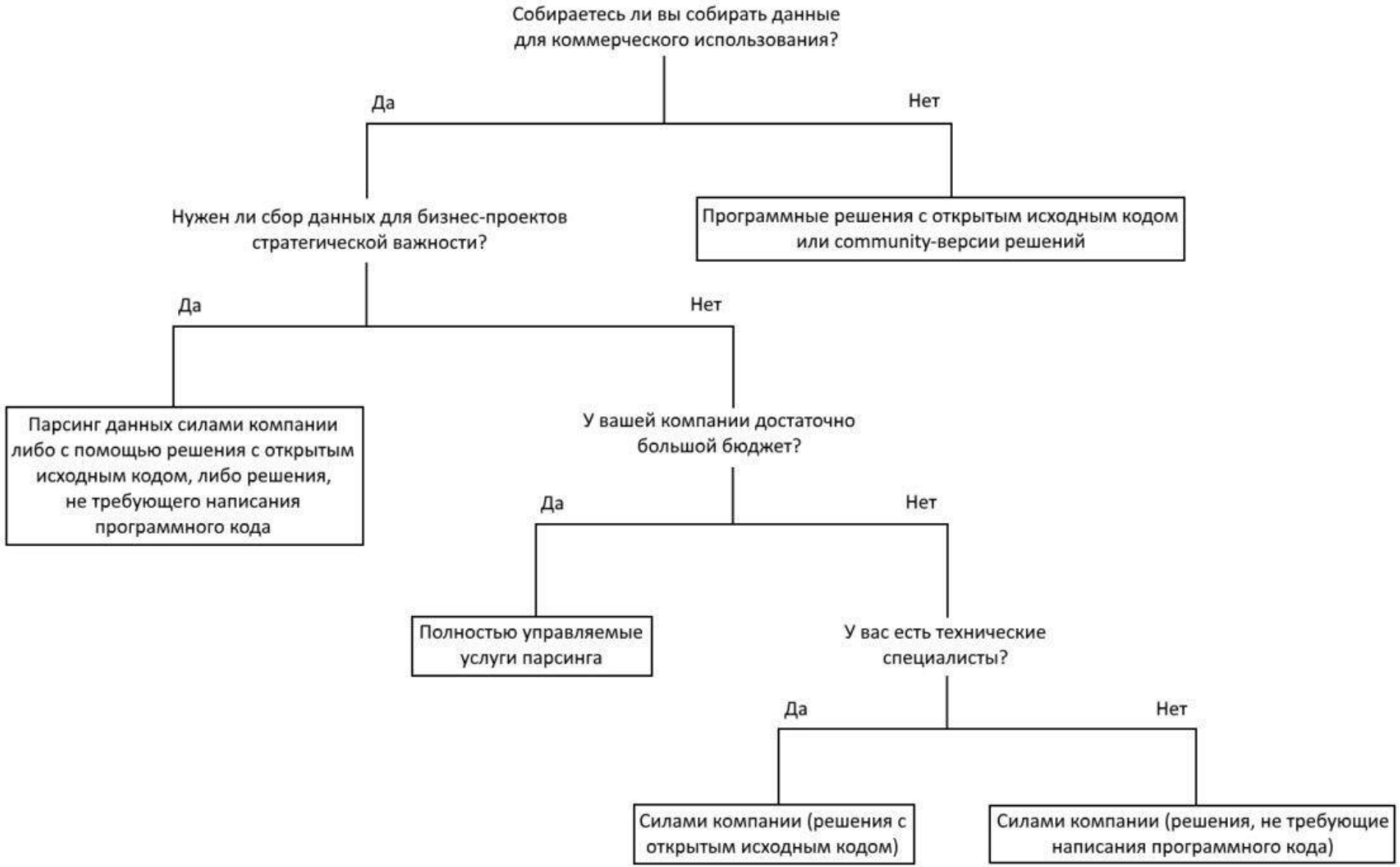

Helium Scraper - phpbb.hs
File Project View Help
ExtractBoard
Browser.TurnPages
Select.NextButton
Select.BoardRowContainer
extract
url
Gather.URL
topic_title
Select.BoardTopicTitle
topic_author
Select.BoardTopicAuthor1
Select.BoardTopicAuthor2
topic_post_count
Select.BoardTopicPostCount
topic_view_count
Select.BoardTopicViewCount
topic_last_activity
Select.BoardTopicLastActivity
topic
Select.BoardTopicTitle
Browser.Navigate
ExtractTopic
Q/A - Helium Scraper
https://www.heliumscraper.com/forum/viewforum.php?f=3
HELIUM SCRAPER
Helium Scraper forums
Quick links FAQ
Register Login
Board index › Q/A
Search...
Q/A
NEW TOPIC
Search this forum...
TOPICS
Replies
Views
Last post
Helium Scraper FAQ's / Troubleshooting
Scraping Craigslist ads?
Relative URL
HTML
Text
Helium Scraper FAQ's / Troubleshooting
Scraping Craigslist ads?
Relative URL
Helium scraper as a web monitoring tool
bfodds - o/u 3.5 goals
extraction does not work - tables empty
Project Explorer
Data Flow
Database
Globals
ExtractBoard
ExtractForum
ExtractTopic
Test
Scripting
Selectors
25 Elements Selected

Сайтты парсингтен қалай қорғауға болады?

- Жоғарыда айтқанымыздай, парсинг әрқашан қалыпты мақсаттарда қолданыла бермейді. Егер сіз бәсекелестердің шабуылынан қорқатын болсаңыз, сайтты қорғауға болады. Мұны істеудің бірнеше жолы бар.
- Белгілі бір уақыт ішінде сіздің сайтыңызда жасалуы мүмкін әрекеттер санын шектеңіз. Мысалы, бір IP-мекен-жайдан бір минут ішінде үш сұраныс жасауға рұқсат етіңіз.
- Күдікті әрекетті қадағалаңыз. Егер сіз бір мекен-жайдан көптеген сұраныстарды байқасаңыз, оған кіруге тыйым салыңыз. Немесе reCAPTCHA-ны көрсетіңіз, сонда пайдаланушы бот немесе талдаушы емес, адам екенін растайды.
- Тіркелген келуші сайттағы әрекеттерді орындай алатындай есептік жазба жасаңыз.
-

- Алаңға кірген әр адамға идентификация жасаңыз. Мысалы, пішінді толтыру жылдамдығы немесе түймені басу орны. Пайдаланушының орналасқан жері, экранның ажыратымдылығы туралы ақпарат жинауға мүмкіндік беретін сценарийлер бар.
- Сайттың құрылымы туралы ақпаратты жасырыңыз. Оған тек әкімші қол жеткізе алады.
- Бір уақытта әртүрлі IP мекенжайларынан келетін ұқсас немесе бірдей сұрауларға назар аударыңыз. Парсинг үлестірілген болуы мүмкін. Мысалы, прокси-серверлер арқылы.
- Қалай болғанда да, бағдарламаны емес, нақты пайдаланушыны блоктау қаупі бар екенін ұмытпаңыз. Сондықтан, ең бастысы-сайттың қауіпсіздігі немесе әлеуетті Клиентті жоғалту қаупі сізге байланысты.

ВЫБОР ИНСТРУМЕНТОВ И ПАРТНЕРОВ ДЛЯ СБОРА ДАННЫХ В ЗАВИСИМОСТИ ОТ ВАШИХ УСЛОВИЙ



Веб-сайттарды талдау кезінде қандай қиындықтар туындауы мүмкін?

- Күрделі құрылымы бар Веб-сайттар: көптеген веб-беттер HTML-ді қолдануға негізделген және бір веб-парақтың құрылымы басқасының құрылымынан мүлдем өзгеше болуы мүмкін. Сондықтан бірнеше веб-сайтты спарсинг жасау керек болған кезде, олардың әрқайсысы өз талдаушысын жасауы керек.
- Парсинг қолдауы қымбат болуы мүмкін: веб-сайттар әрқашан веб-беттің дизайнын өзгертеді. Егер жиналған деректердің орналасқан жері өзгерсе, онда деректерді жинаушылардың бағдарламалық коды қайтадан өзгертілуі керек.

- Веб-сайттар қолданатын парсингке қарсы құралдар: мұндай құралдар веб-әзірлеушілерге роботтар мен адамдарға көрсетілетін мазмұнды басқаруға, сонымен қатар роботтарға веб-сайтта деректерді жинау мүмкіндігін шектеуге мүмкіндік береді.
- Парсингтен қорғаудың кейбір әдістері: IP-мекен-жайларды бұғаттау, captcha (толық автоматтандырылған қоғамдық Turing test to tell Computers and Humans Apart — компьютерлер мен адамдарды ажыратуға арналған толықтай автоматты Тьюринг тесті) және парсерлерге арналған тұзақтар.

- Авторизацияның қажеттілігі: Бүкіләлемдік ғаламторда белгілі бір ақпаратты жинау үшін алдымен авторизациядан өту қажет болуы мүмкін. Сондықтан, веб-сайт кіруді талап еткен кезде, талдаушы сұрау салумен бірге жіберілген cookie файлдарын сақтайтынына көз жеткізу керек, осылайша веб-сайт талдаушыны бұрын кірген адам ретінде қабылдайды.
- Жүктеу жылдамдығы баяу немесе тұрақсыз: веб-сайттар мазмұнды баяу жүктегенде немесе сұрауларға жауап бермесе, бетті жаңарту көмектесе алады, дегенмен талдаушы мұндай жағдайда не істеу керектігін білмеуі мүмкін.

Назарларыңызға рақмет!