
Article

[Yuri Marx Perei...](#) · Dec 21, 2020 2m read

[Open Exchange](#)

Do NLP in any website with InterSystems IRIS and Crawler4J

Today, is important analyze the content into portals and websites to get informed, analyze the concorrents, analyze trends, the richness and scope of content of websites. To do this, you can allocate people to read thousand of pages and spend much money or use a crawler to extract website content and execute NLP on it. You will get all necessary insights to analyze and make precise decisions in a few minutes.

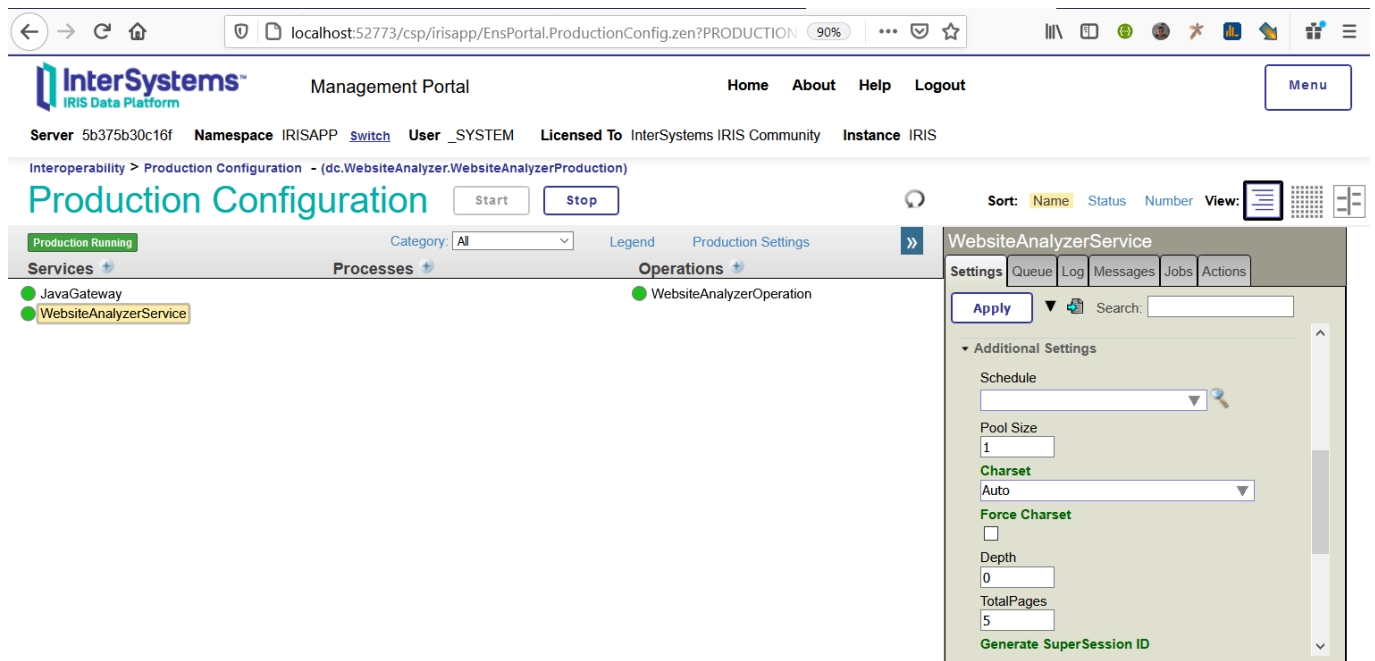
Gartner defines web crawler as: "A piece of software (also called a spider) designed to follow hyperlinks to their completion and to return to previously visited Internet addresses".

There are many web crawlers to extract all relevant website content. In this article I present to you Crawler4J. It is the most used software to extract website content and has MIT license. Crawler4J needs only the root URL, the depth (how many child sites will be visited) and total pages (if you want limit the pages extracted). By default only textual content will be extracted, but you config the engine to extract all website files!

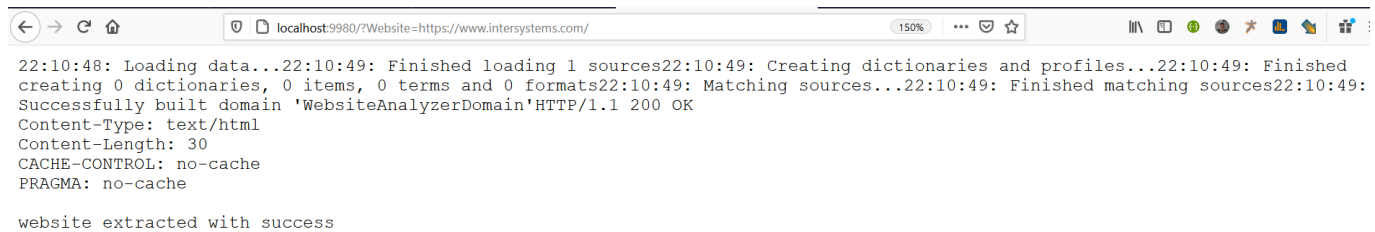
I created a PEX Java service to allows you using an IRIS production to extract the textual content to any website. the content is stored into a local folder and the IRIS NLP reads these files and show to you all text analytics insights!

To see it in action follow these procedures:

- 1 - Go to <https://openexchange.intersystems.com/package/website-analyzer> and click Download button to see app github repository.
- 2 - Create a local folder in your machine and execute: <https://github.com/yurimarx/website-analyzer.git>.
- 3 - Go to the project directory: `cd website-analyzer`.
- 4 - Execute: `docker-compose build` (wait some minutes)
- 5 - Execute: `docker-compose up -d`
- 6 - Open your local InterSystems IRIS: <http://localhost:52773/csp/sys/UtilHome.csp> (user `SYSTEM` and password `SYS`)
- 7 - Open the production and start it: <http://localhost:52773/csp/irisapp/EnsPortal.ProductionConfig.zen?PRODUC...>



8 - Now, go to your browser to initiate a crawler: <http://localhost:9980?Website=https://www.intersystems.com/> (to analyze intersystems site, any URL can be used)



```
22:10:48: Loading data...22:10:49: Finished loading 1 sources22:10:49: Creating dictionaries and profiles...22:10:49: Finished creating 0 dictionaries, 0 items, 0 terms and 0 formats22:10:49: Matching sources...22:10:49: Finished matching sources22:10:49: Successfully built domain 'WebsiteAnalyzerDomain'HTTP/1.1 200 OK
Content-Type: text/html
Content-Length: 30
CACHE-CONTROL: no-cache
PRAGMA: no-cache

website extracted with success
```

9 - Wait between 40 and 60 seconds. A message you be returned (extracted with success). See above sample.

10 - Now go to Text Analytics to analyze the content extracted:
http://localhost:52773/csp/IRISAPP/iKnow.UI.KnowledgePortal.zen_?NAMESPACE=IRISAPP&domain=1

Do NLP in any website with InterSystems IRIS and Crawler4J

Published on InterSystems Developer Community (<https://community.intersystems.com>)

The screenshot displays the InterSystems IRIS NLP application interface. At the top, a search bar contains the text 'first data platform' and an 'Explore!' button. Below the search bar, there are three main panels: 'Dominant Concepts', 'Similar Entities', and 'Related Concepts'. The 'Dominant Concepts' panel shows a list of concepts with their frequency and dominance. The 'Similar Entities' panel shows a list of entities with their frequency and proximity. The 'Related Concepts' panel shows a list of related concepts with their frequency and proximity. Below these panels, there is a 'Sources' tab with a list of sources including 'image/svg+xml', 'Menu', 'Try', 'InterSystems', 'IRIS', 'Partners', 'Developers', 'Contact', 'USA', 'Australia', '&', 'New Zealand', 'Benelux', 'Dutch', 'Benelux', 'Fran?ais', 'Brazil', 'Chile', 'China', 'Czech', 'Republic', 'Finland', 'France', 'Germany', 'Hungary', 'Italy', 'Japan', 'Kazakhstan', 'Middle', 'East', 'Russia', 'Singapore', 'South', 'Africa', 'Spain', 'Sweden', 'Ukraine', 'United Kingdom', '&', 'Ireland', 'Industries', 'Finance', 'Proven', 'speed', 'scalability', 'and', 'reliability Health', 'Creating', 'high-value', 'sustainable health system Business Turning data', 'into', 'sound business decisions Government Helping governments', 'serve', 'citizens Products InterSystems IRIS Data Platform', 'fastest way', 'to build and deploy', 'most demanding applications IRIS', 'for', 'Health Healthcare Data Platform', 'first data platform', 'engineered to extract', 'value', 'from', 'healthcare data HealthShare', 'Unified', 'Care Record Suite', 'or', 'connected health solutions', 'based on', 'unified care record TrakCare', 'Unified', 'Healthcare Information System', 'Improve', 'experience', 'for', 'patients', 'and', 'workplace', 'for', 'clinicians Support &', 'Learning', 'Learning', 'Services', 'Online', 'Learning', 'Documentation', 'Virtual', 'Classroom', 'Learning', 'InterSystems', 'Certification', 'Developer', 'Community', 'University', 'Outreach', 'Customer', 'Support', 'Product', 'Alerts', '&', 'Advisories', 'Version', 'Information', 'Pre-Release', 'Trial', 'Program', 'Documentation', 'For', 'Immediate Help Worldwide Response Center', 'WRC', 'For', 'Immediate Response', '?

11 - Return to the production and see Depth and TotalPages parameters, increase the values if you want extract more content. Change Depth to analyze sub links and change TotalPages to analyze more pages.

12 - Enjoy! And if you liked, vote (<https://openexchange.intersystems.com/contest/current>) in my app: [website-analyzer](#)

I will write a part 2 with implementations details, but all source code is available in Github.

[#Analytics](#) [#InterSystems IRIS](#)
[Check the related application on InterSystems Open Exchange](#)

Source URL: <https://community.intersystems.com/post/do-nlp-any-website-intersystems-iris-and-crawler4j>