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 alocate 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.

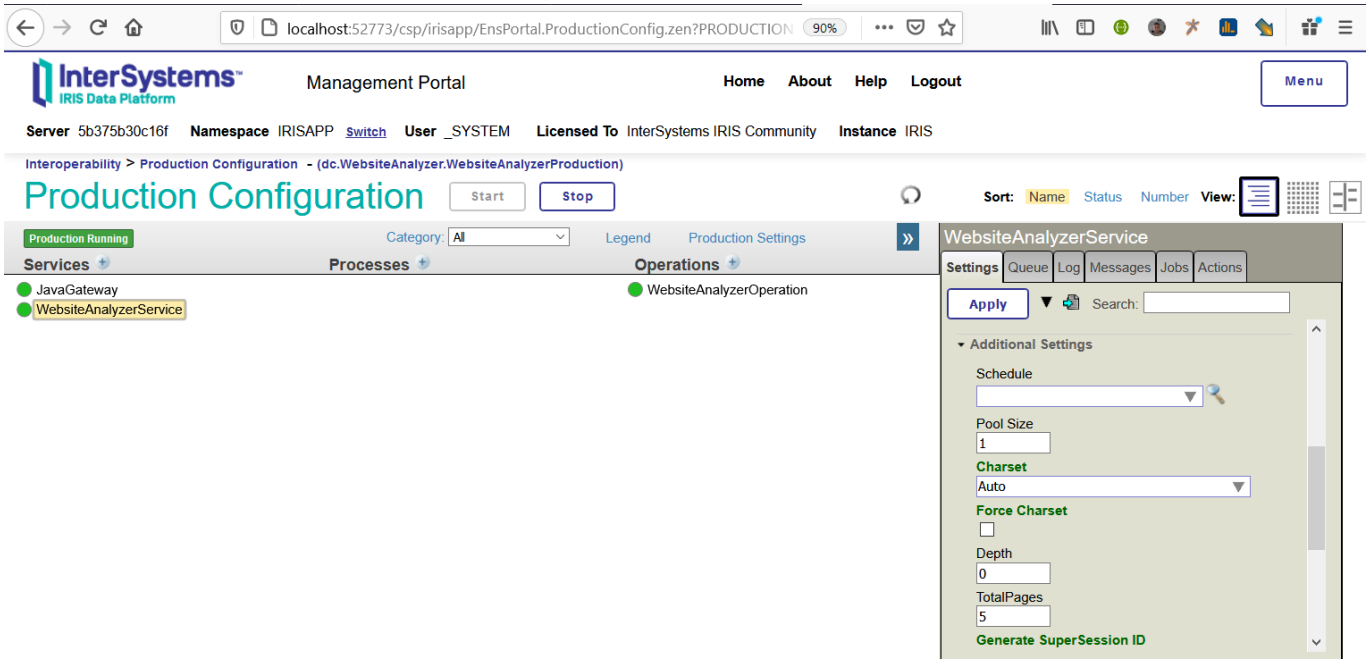
3 - Go to the project directory: cd website-analyzer.

4 - Execute: docker-compose build (wait some minutes)

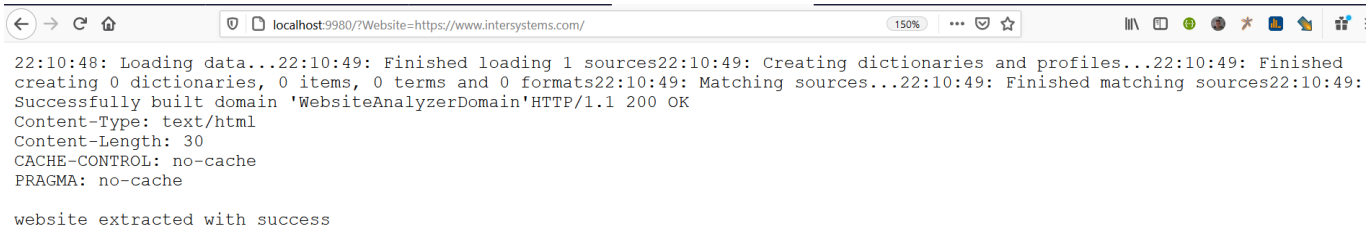
5 - Execute: docker-compose up -d

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...

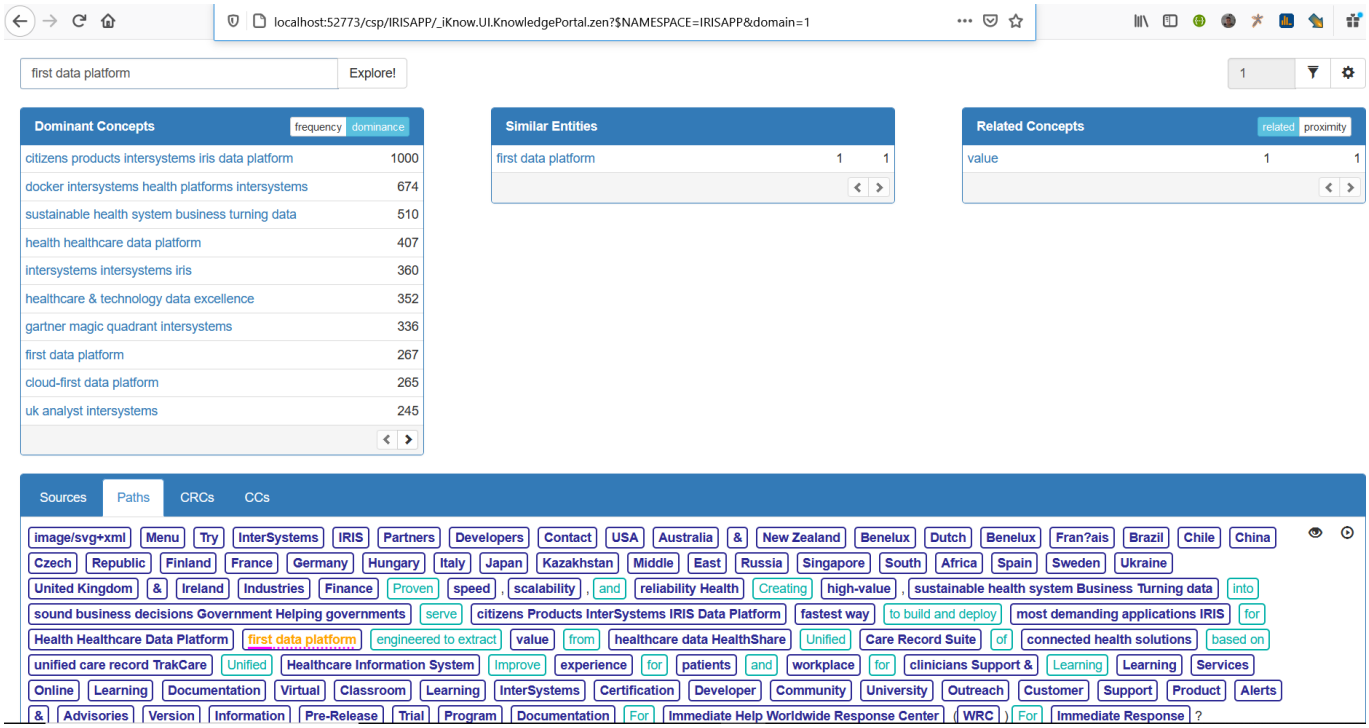


8 - Now, go to your browser to initiate a crawler: http://localhost:9980?Website=https://www.intersystems.com/ (to analyze intersystems site, any URL can be used)



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/jKnow.UI.KnowledgePortal.zen ?$NAMESPACE=IRISAPP&domain=1



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