موضوع ارائه:

پروژه کرالر سایت خبری عربی

محمد ندیمی

Abstract

This project is for crawl news articles of Alsabah website with python and Scrapy.

Web Crawler

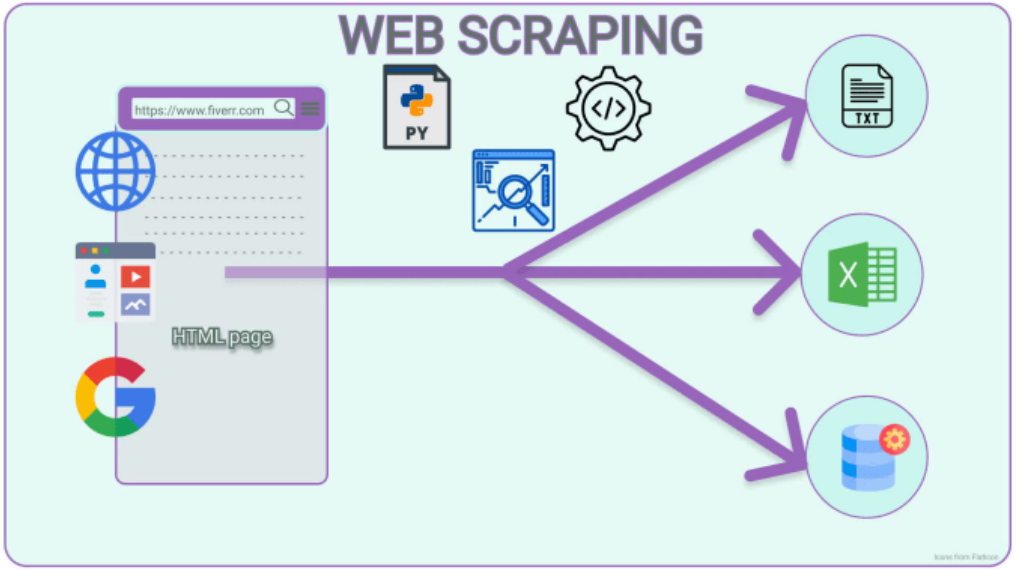
A web crawler can serve two functions:

- Systematically browsing the web to index content for search engines. Web crawlers copy pages for processing by a search engine, which indexes the downloaded pages for easier retrieval so that users can get search results faster.
This was the original meaning of web crawler.

- Automatically retrieving content from any web page. This is more commonly called web scraping. This meaning of web crawler came about as companies other than search engines started using web scrapers to retrieve web information.

scrapy

Scrapy is a Python-based open-source web scraping framework used for extracting data from websites. It has built-in tools and libraries for handling common web scraping tasks and uses a modular architecture with a powerful selector system for parsing HTML and XML documents.



MongoDb

What is MongoDB?

MongoDB is an open-source NoSQL document-oriented database system designed for handling large amounts of unstructured data. It uses a document model and provides advanced features like dynamic queries and real-time data aggregation.

Features

- Crawl all posts of the website.
- Crawl any number of pages you want.
- Crawl new posts of website immediately and realtime.
- Crawl (title,img_url,publish_date,link,description,text_without_line_breaks,text_with_line_breaks,tag) from each post.

Lastnews Page

Link

Every posts has a link to details page and crawler gets the link of everyposts on the page.

Pagenation

Crawler paginate the pages till the number of page we want to crawl.

Post item page

Title
Tag
Main_img,
Publish_date

Text

In this page, It gets the title, tag, img_url, publish_date and text.

Results in database

DATA ▾ | EXPORT COLLECTION

```
                    "…بدأت شركة (ميتا) اختبار تصميم جديد لتطبيق واتساب لجو
text_without_line_breaks: "…نيويورك: وكالاتبدأت شركة (ميتا) اختبار تصميم جديد لتطبيق واتساب لج "
title: "{واتساب} تختبر تصميماً خاصاً بالحواسيب اللوحيّة"

_id: ObjectId('6405789d65e63784c8641c65')
link: "https://alsabaah.iq/72723-.html"
img_url: "https://alsabaah.iq/uploads/posts/2023-03/1678039974_image001.jpg"
publish_date: 1678048200
tag: "منصة"
text_with_line_breaks: "بغداد: نجلاء الخالدي
                    …لا نعرف كيف ساقتنا الأقدار إلى "اضنة" المدينة اله"
text_without_line_breaks: "… بغداد: نجلاء الخالديلا نعرف كيف ساقتنا الأقدار إلى "اضنة" المدينة "
title: "عراقيون يروون قصصاً إنسانيّة عن زلزال تركيا المدمر"

_id: ObjectId('6405789e65e63784c8641c69')
link: "https://alsabaah.iq/72722-.html"
img_url: "https://alsabaah.iq/uploads/posts/2023-03/jmbxqywujakvf28p6ha5w6.jpg"
publish_date: 1678048200
tag: "ثقافة"
text_with_line_breaks: "دعد ديب
                    …يتعلق الإنسان عادة بما يخاف منه بشكل غير واع وغير مدرك، وهذا"
text_without_line_breaks: "…دعد ديب يتعلق الإنسان عادة بما يخاف منه بشكل غير واع وغير مدرك، وهذ "
title: "فيلم {العين الزرقاء الشاحبة} واستعارة أقنعة ادغار ألان بو"

_id: ObjectId('6405789f65e63784c8641c6d')
link: "https://alsabaah.iq/72721-.html"
img_url: "https://alsabaah.iq/uploads/posts/2023-03/1678039609.jpg"
publish_date: 1678048200
tag: "ثقافة"
text_with_line_breaks: "علي حمود الحسن
                    …تحرمُ صفحة "سينما" على التذكير بأفلام كلاسيكيَّة عراقيّ"
text_without_line_breaks: "…علي حمود الحسنتحرمُ صفحة "سينما" على التذكير بأفلام كلاسيكيَّة عراقيّ"
title: "إشكاليَّة الموت والخوف منه ..{السقا مات}"
```

Summery

The AlsabahmongoSpider is a web scraper built using the Python Scrapy library. It crawls the Alsabaah.iq website, a popular Iraqi news website, and extracts news articles' data. The data is then stored in a MongoDB database.

sources

www.alsabah.iq

Thank you for your attention