

Coding Exercise

Problem

We'd like you to write a simple web crawler in Javascript or Python

Given a starting URL, the crawler should visit each URL it finds on the same domain. It should print each URL visited, and a list of links found on that page. The crawler should be limited to one subdomain - so when you start with <https://www.overstory.com/>, it would crawl all pages on the overstory.com website, but not follow external links, i.e. links on a non overstory domain.

We would like to see your own implementation of a web crawler. Please do not use frameworks like scrapy or crawlee which handle all the crawling behind the scenes or someone else's code.

You are welcome to use libraries to handle things like HTML parsing.

Please write your solution as you would a production piece of code. This exercise is intended to allow you to demonstrate how you design software and write high quality code. We are interested in how your program is structured: the trade-offs you've made (which ones? why?), what behaviour the program exhibits, and appropriate use of concurrency, test, and so on. With this in mind please focus on writing a command line application and include a readme file that outlines how to run your app and documents key design decisions made with justification.

We will be testing it on websites of different size and complexity.

We'd love it if you could do this within the next week or so, but let me know if you'll need more time. Once you have submitted your task we will then schedule a 1 hour call with our engineers during which you'll share your screen and work on an extension exercise.

When you're ready, please submit your solution as a ZIP file in the following format: NAME_SURNAME.zip.