Module 10 Career Connection

Web Scraping Skills Are in Demand

Nice job this week. You learned how to web scrape libraries, which is an important set of skills to showcase when talking with prospective employers. This week we really want to emphasize why web scraping is a valuable skill to add to your resume and the various ways in which web scraping may show up in the technical interview.

In this section, you will:

- Learn how web scraping is used on the job.
- Learn how to showcase this new skill to be a more competitive candidate in the job market.
- Answer common interview questions about web scraping.

On-the-job web scraping involves using specialized web crawling tools to extract the desired data from other websites-usually, as you've learned in this module, this data will then be stored in a database and used for analysis. Companies that might use this technique of data extraction may use the data for competitor analysis, tracking market trends, price research on competitors, or simply to be a data-driven company to improve.

As a data analyst and engineer, you can use your web scraping skills to collect information for future employers-so don't be shy about showcasing that you know how to do this.

Employer Competitive Advantage

Adding web scraping to your resume's skills list will help you pass through the applicant screening filters. It also tells a potential employer that you can capably harvest the data they need. For more resume support, see Milestone: Develop Your Resume.

See Milestone: Develop Your Resume by accessing your career services content using the link in the left-hand navigation menu.

Consider the following possible scenarios, for which you might find yourself working with web scraping in the professional world.

After reading each case study, you will see some common technical interview questions around the scenario you were presented. During a technical interview, you may receive one of two types of questions-you might even get both. Typically, you can expect the technical interview to fall into two different categories:

- **Technical questions with short answers**
- **Broader, less-technical questions that require reflective thinking**

Let's get started.

Technical Interview Preparation

Case Study No. 1: Price Comparison You've just been offered that job at NASA you've been dreaming about, and it's time for that final technical interview. As part of the interview, they give you the following prompt to read and consider.

You recently started working at a major online clothing retailer (Company A) whose focus is on selling high volume to increase its profit margins. Because this e-retailer is not a boutique marketplace selling to a niche market, it depends on its product, and its competitive pricing, to reach the largest audience possible.

However, a few months ago, a major competitor (Company B) entered the online market, and Company A wants to keep an eye on its competitor's pricing so that it can offer the best possible prices to its own customers.

You've been hired to do the data analysis, but because Company B isn't going to just go ahead and release all of its pricing in a well-documented API for you (you wish!), you'll have to regularly scrape the data off its pages and maintain a database of the products, prices, and

any price changes.

After reading this prompt, the NASA technical interviewers ask the following questions:

1. **Would you consider it legal to scrape data from your competitors?**
    - *This is a relatively gray area. The short answer is yes, unless there's some sort of privacy/legal agreement on its website that specifically prohibits web scraping. However, even if it's legal to scrape data from the page, do consider how you then use that data may have legal implications of its own. In other words, you couldn't scrape data and then republish and represent it as your own.*

2. **Can you scrape data behind a login page?**
    - *Yes, you could. But it is significantly more difficult. You would need to provide the web application with valid credentials, and then navigate to the authenticated portion of the site and scrape-to do this, you could use some sort of browser automation tool like Selenium Web Driver. This process, though, is not readily recommended.*

Case Study No. 2: Airline Tickets

FlyCheap is a locally owned tech company with a big idea. Using its browser extension, customers can book flights to travel all over the world on essentially any airline. However, it's facing a very real problem-its major competitors, Kayak and Google, change their flight prices multiple times per day. So how do FlyCheap customers know when the best time is to book their tickets?

Well, that's where you come in. You've just been hired to improve the functionality of the browser extension by allowing it to pop up with an alert, for those who have it installed, when the price of a flight has dropped. But because the Google Flights API was recently deprecated, and you can no longer just make an API request for it-you're going to have to get the information yourself.

Your Task

Of course, you can't sit there all day monitoring the prices of flights manually-there are just too many. So your first task on the job with FlyCheap is to write an application that scrapes data from Kayak, Google Flights, and other companies and monitors price variations. When a flight drops or increases in price, that information will get fed to the browser extension and then on to the end-user.

1. **Can you extract data from sites not written in English?**
    - *Of course. You can extract data in any language, even if it's not in a Roman-style alphabet (i.e., Chinese, Japanese, Korean), but obviously the material you scrape remains in the language you scrape it in.*

2. **Can you republish data and/or information that you scraped from the web?**
    - *Another gray area-maybe! Watch out for policies that explicitly forbid redistribution of material and/or the citation guidelines. You might be able to freely republish, not republish at all, or republish*

with limitations and credits to the original authors. If you're unsure, get in touch with the owner of the site you're scraping from.

Case Study No. 3: Customer Review Sentiment Analysis

Companies and their products live and die on customer reviews-all hail the mighty five gold stars. Your current company sells its products on Amazon, but it needs access to comprehensive data analytics on the customer reviews.

Consider the ethical implications of scraping data:

- how much data are you requesting from the web site and are you going to overload their server load?
- is the data protected by law?
- can you keep the data once it has been scraped?
- can you take data for free and commercialize it?
- do you provide a way for the owner of the data to contact you if necessary?

There are many different things to consider, and some of these ethical considerations blur the lines into what is legal and what is not. For what it's worth, we recommend you always respect the site owner's wishes, include a transparent User Agent string, and seriously consider the purposes to which you are scraping data. Oh, and if there is a public API available, use that instead!

Your Task

Imagine you were asked to use web scraping to crawl the Amazon product pages for your company's products and extract the review text and numerical value, then using a sentiment analysis library like **VADER Sentiment Analysis (https://github.com/cjhutto/vaderSentiment)** to analyze the text to provide useful and actionable feedback for your company.

Note: You do not actually need to complete the web scraping, nor do you have to know how to use the VADER Sentiment Analysis tool.

The hiring manager poses the following question based on the above scenario, how would you respond?

Would you say that web scraping is ethical?

Continue to Hone Your Skills

If you're interested in learning more about the technical interviewing process and practicing algorithms in a mock interview setting, check out our **upcoming workshops. (https://careernetwork.2u.com/?utm_medium=Academics&utm_source=boot_camp)**

© 2020 - 2022 Trilogy Education Services, a 2U, Inc. brand. All Rights Reserved.