Web scraping: an accessible data acquisition opportunity for every comapny – even SMEs

Explaining the basics of web scraping and why ut could be valuable for you

These days, as the economy is digitizing at an impressive pace, data-driven decision-making is getting more and more important for every business to maintain its competitive advantage. The key ingredient that enables data-driven decision-making is, of course, data. This refers to both data on the internal organizational processes as well as data external to the organization.

Organizing the internal data processes have been documented elsewhere, in this article we focus on the external acquisition of data. Acquiring data can be a costly endeavor, both in terms of time and money, especially for small to medium enterprises (SMEs). In this article, we will zoom in on **web scraping** as a means to gather valuable external data.

The web as a source of data

Web scraping is the systematic and automated extraction of data from websites. By means of web scraping, companies can tap into one of the most significant data sources in existence: the internet. On the internet, customers and competitors leave data on their behaviour, insights and interactions which can be useful for particular decisions a company is facing. It is a perfect data source for SMEs as it allows for efficient and cost-effective data acquisition.

Especially for those SMEs which are interested in working with data or want to explore how data science can help their business, web scraping – when applied to specific organizational challenges and decisions – allows for a low barrier to start gathering external data and develop the internal data science skills.

Interreg North Sea Region FBD
European Regional Development Fund
EUROPEAN UNION
FUTURES BY DESIGN
futuresbydesign.net

Web scraping – the basics

Virtually any piece of text, set of numbers or displayed image that you see on any web page can be scraped and thus be classified as data to be collected, analyzed, and turned into valuable insights that can guide decision making. The only condition is that the data should be consistently presented across multiple pages of a particular website. For instance, the price of an Amazon.com product is always displayed in the same location for different products and can therefore be scraped.

Practically, web scraping consists of 2 main processes: 1) crawling – meaning "going through the different pages of interest"; and 2) scraping – meaning "selecting and storing the data you are interested in". These processes are often confused, but they are quite distinct: web crawling takes care of loading the different pages from which the data needs to be scraped. For instance, all the different board game product pages on Amazon.com. Subsequently, web scraping extracts the desired information from all of these pages, such as prices, names, reviews, etc. and collects them into a more useful format (e.g. Excel workbooks). Once that's done, the data is ready for analysis.

Web scraping in practice

As web scraping can be used to collect almost any data that can be found online, the most important aspect is to brainstorm about how you are able to leverage a specific type of data. The main takeaway is: anything you see online could be scraped. However, whether it is relevant to gather anything is hardly ever the case. Make sure you gather data which aid you in making particular decisions which currently are less- or ill-informed and taken on the basis of anecdotes or gut-feeling.

One example of a company leveraging web scraping in their core business is HiQ. HiQ scrapes publicly available employee data from LinkedIn (a court has ruled that this is legal in 2019). They use this data to help the HR departments of their corporate clients to manage their workforce. They develop employee skill maps and build attrition models to predict when employees are likely to leave their jobs. As a result, HiQ can predict whether skills shortages might emerge and estimate turnover risks months in advance, saving their clients a lot of money and time.

Another interesting example is Proven, a skincare company that scrapes customer reviews to personalize their product offerings. As such, web scraping helps their insight on what they actually know about the experience their clients

have with their products. Since this is a lot of data, they in turn feed the scraped reviews into a machine-learning algorithm to discover any interesting correlations which in turn help boost their customer experience and profitability.

What can I do with web scraping?

Several examples of scrapable data points together with their potential applications are:

- **Stock prices:** stock prices can be collected from financial websites such as finance.yahoo.com in order to make better investment decisions.
- **Email addresses/phone numbers:** email addresses and phone numbers can be collected from yellow page websites such as Yellowpages.com in order to generate better sales leads.
- **Addresses:** business addresses can be scraped from business registers such as e-justice.europa.eu to create a list of business locations to determine the geographical landscape of the competition.
- **Product data:** product data such as prices, names, reviews and delivery times can be scraped from e-commerce websites such as Amazon.com to perform competitor analysis.
- **Social media posts:** social media posts can be collected from websites such as Twitter.com to perform sentiment analysis and understand the public opinion on certain topics.
- **Menu information:** restaurant menu information can be scraped from websites such as Takeaway.com to create dynamic prices for your own dishes.

This is a non-exhaustive list of examples and the web scraping possibilities are almost endless. What data points you need to scrape depends on your specific business goals and applications.

How can I start with web scraping?

The above practical examples and suggestions illustrate that web scraping can be used on a wide variety of domains of interest for a business. Ultimately, web scraping is an important tool, which belongs in any business's toolbox, to facilitate the acquisition of valuable, publicly available data that could take the business to the next level. So where to start?

- **Step 1:** Identify a specific point of focus – what is a specific challenge within the organization for which you argue "Having these data from site ABC would aid us in making a better informed decision".

- **Step 2:** Create an overview of which data you specifically need.

- **Step 3:** Check whether it is allowed by the website to scrape the particular data – some sites ensure there is copy- or database right on the data they display.

- **Step 4:** Use a particular scraping tool to gather the data as described in Step 2. Examples of user-friendly tools are ParseHub and ScrapingBee. You can also check out our web scraping Python tutorial here, where we use the Beautiful Soup package.

- **Step 5:** Need any help? Contact info@jadsmkbdatalab.nl!

Interreg North Sea Region FBD European Regional Development Fund
EUROPEAN UNION
FUTURES BY DESIGN
futuresbydesign.net