Web Scraping

William Trang

What is data?

In convenient terms, data is anything that can be stored.

Unstructured vs Structured Data

Structured Data

Often numbers or labels, stored in a structured framework of columns and rows relating to pre-set parameters.

- ID CODES IN DATABASES
- NUMERICAL DATA GOOGLE SHEETS
- STAR RATINGS

Semi-unstructured Data

Loosely organized into categories using meta tags

- EMAILS BY INBOX, SENT, DRAFT
- TWEETS ORGANIZED BY HASHTAGS
- FOLDERS ORGANIZED BY TOPIC

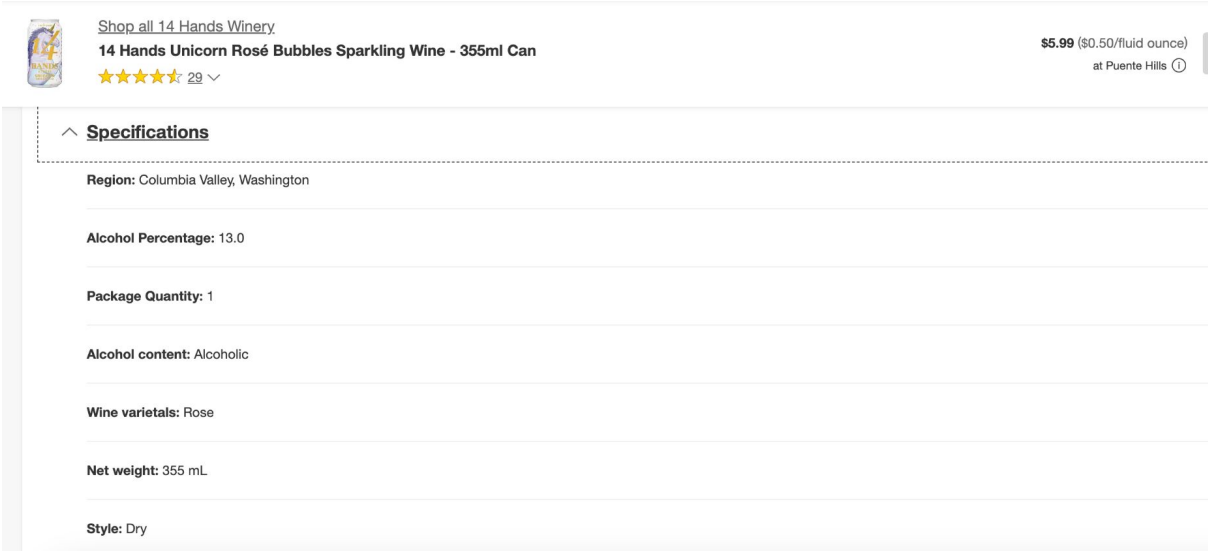
Unstructured Data

Text-heavy information that's not organized in a clearly defined framework or model.

- MEDIA POSTS, EMAILS, ONLINE REVIEWS
- VIDEOS, IMAGES
- SPEECH, SOUNDS

What is web scraping?

Web Scraping is an automatic way to retrieve unstructured data from a website and store them in a structured format.



Reasons for web scraping

Web scraping can significantly speed up the process of data collection. Instead of manually searching for, parsing through, and manually entering the data, you can have it done manually.

In our context, the Target website has thousands of bottles of wine that we don’t want to manually go through!

Challenges of web scraping

There are two main challenges of web scraping

- Every website is different, so there is no “one size fits all” solution
- Websites are constantly changing: take PairAnything for example, so the way they have to be scraped is different too

Can you scrape every website?

The simple answer is no. Every website has its own set of rules on who can and who cannot scrape. This is always stored in a robots.txt file. To access this, just take the name of a website and type /robots.txt after it. For example, target.com/robots.txt will pull up the scraping rules.

```
Sitemap: https://www.target.com/sitemap_keywords-index.xml.gz
Sitemap: https://www.target.com/sitemap_stores-index.xml.gz
Sitemap: https://www.target.com/sitemap_taxonomy-categories-index.xml.gz
Sitemap: https://www.target.com/sitemap_pdp-index.xml.gz
Sitemap: https://www.target.com/sitemap_taxonomy-brand-index.xml.gz
Sitemap: https://www.target.com/sitemap_facet-categories-index.xml.gz

User-agent: *
Disallow: /*/Ntk
Disallow: /*/Ntt
Disallow: /*/Ntx
Disallow: /*%7Cd_
Disallow: /*/schoollist/
Disallow: /*BTWN
Disallow: /[path]/
Disallow: /7078046/
Disallow: /7079046/
Disallow: /AddToList
Disallow: /AddToRegistry
Disallow: /admin
Disallow: /advancedGiftRegistrySearchView
Disallow: /AjaxSearchNavigationView
Disallow: /Allons_voter
Disallow: /bp/c/
Disallow: /bp/guest_mfg_brand
Disallow: /bp/p/
Disallow: /CallToActionModalView
Disallow: /cgi-bin
Disallow: /cgi-local
Disallow: /Checkout
Disallow: /CheckoutEditItemsDisplayView
Disallow: /CheckoutOrderBillingView
Disallow: /CheckoutOrderShippingView
Disallow: /CheckoutSignInView
Disallow: /co-
Disallow: /common
Disallow: /coupons.
Disallow: /custom-reviews/
Disallow: /data
Disallow: /database/philboard.mdb
Disallow: /dir_on_server/
Disallow: /EmailCartView
Disallow: /EnlargedImageView
Disallow: /ESPDisplayOptionsViewCmd
```

Alternatives to web scraping: APIs

Sometimes, websites will completely disallow web scraping from all agents. This is where APIs come in. APIs are ways built by companies that allow users to interact with their data in a predefined way. APIs are resistant to front-end changes, as they are built by the companies, but can be everchanging as well.