Source Catalog

This file contains information on websites that were used for crawling

Websites

- https://www.cdc.gov/flu/testing/index.html
- https://health.ucsd.edu/

Notes

- I crawled two websites: CDC and UCSD Health
- All data that was collected came from publicly available health information pages
- The pages do not require logins or authentications
- Crawler is respectful of rate limits
- Crawler retrieved content from HTML text
- The crawler extracts phone numbers, addresses, and facility names
- The crawler is ethical in the sense that it does not bypass restrictions

Refresh/update strategy

- Fetches a targeted page and generates a new JSON file each time. For example, if I input the CDC web page, it will grab content from that specific page and generate a JSON file. If I input the UCSD web page, it'll fetch content from that page and generate a JSON file specific to that page.
- This ensures that data reflects current information
- Output file is timestamped