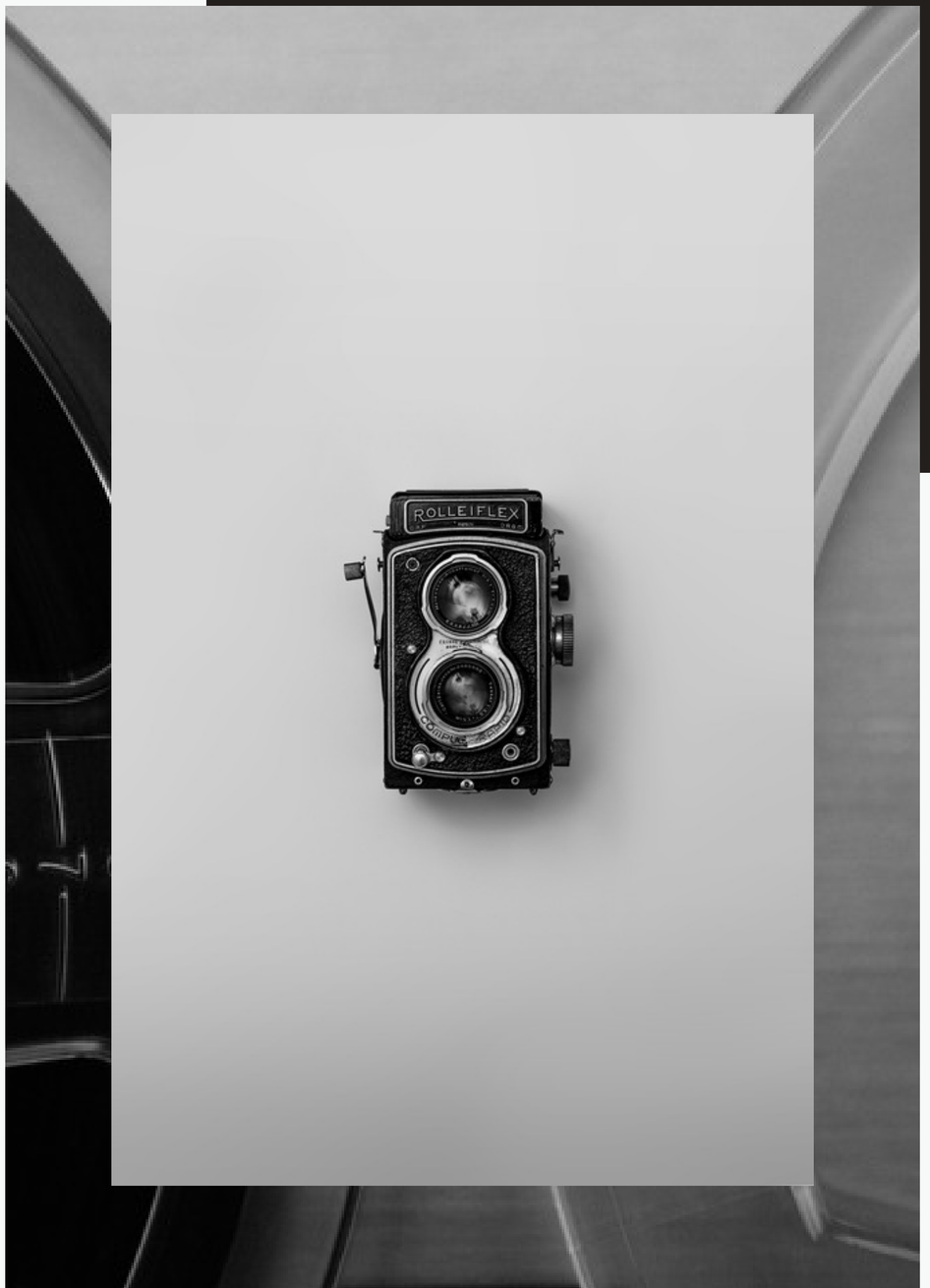
By James Komo

The Web in Seconds

The Presentation

- Problem Statement
- Proposed Solution
- Features & Tools
- Challenges
- Demo

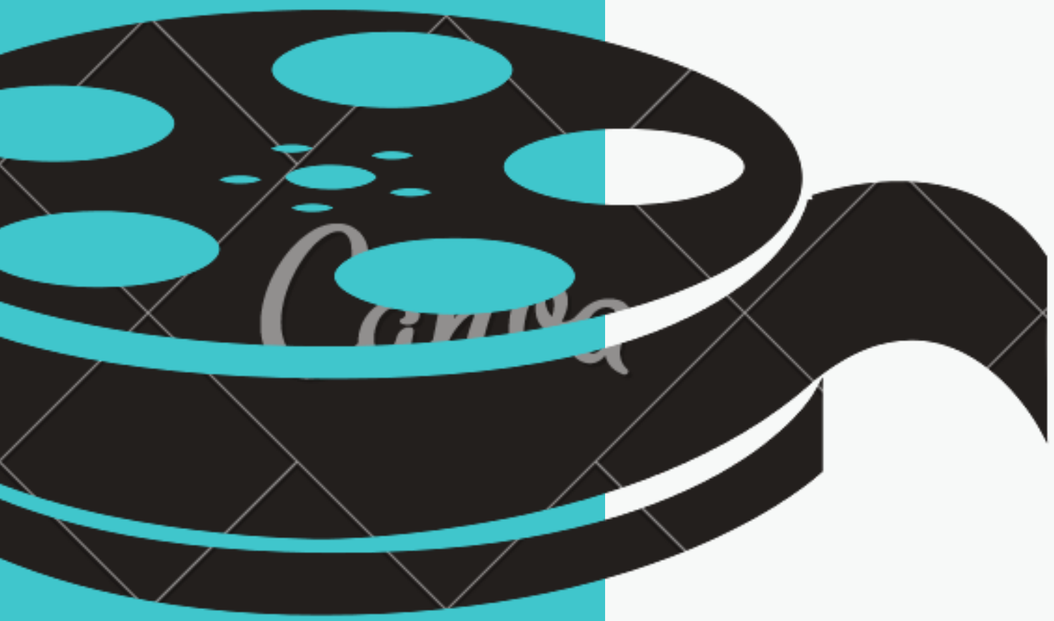


PROBLEM STATEMENT

An overwhelming proportion of data that businesses use for the purpose of developing business insights in decision making is derived from the internet, and the tendency to depend on data-informed insights is expected to become a more conspicuous mainstream practice with the expansion of Internet of Things

The question is, why spend tens of hours seeking data instead of focus on data manipulation and predeictive analysis? This web crawler answers that question in seconds

The Solution



This Web crawler scans the web and produces an index of the web pages (URLs) for post-processing and download the web contents to CSV. In this process, it also locates and extracts the texts and catalogs the hyperlinks and tags.

Tools

- BeautifulSoup API
- Requests Library
- LXML Parser
- Python Time Module
- Python CSV Module
- Codebeautify.org

Features

- **CSV**
  Allows writing to CSV
- **MultiPage Manipulation**
  Allows scraping across multiple pages
- **Delay Management**
  To avoid server overload
- **Parse & User Agent**
  To convert HTML to Python Objects and avoid bot blocking

Challenges

AntiScraping Technologies

Some sites have dynamic coding algorithms to disallow bot access and implement IP blocking mechanisms

DOM Manipulation

There we challenges in traversing DOM trees to identify child/parent/sibling relationships

Dynamic Sites

Managing dynamic content for constantly changing UI iterfaces

The Future

DATABASE

Use Databases instead of CSV

PANDAS

Use Python Pandas for better data analysis, predictive analysis and forecasting

UI/UX

Allow users to pass links to sites they want scraped through a UI interface

Demo

Q & A

Thank You