WEB SCRAPING

TAYARA.TN

OUMAYMA ABAYED
ESLEM SEBRI

OUR PROJECT

01

Web Scraping

02

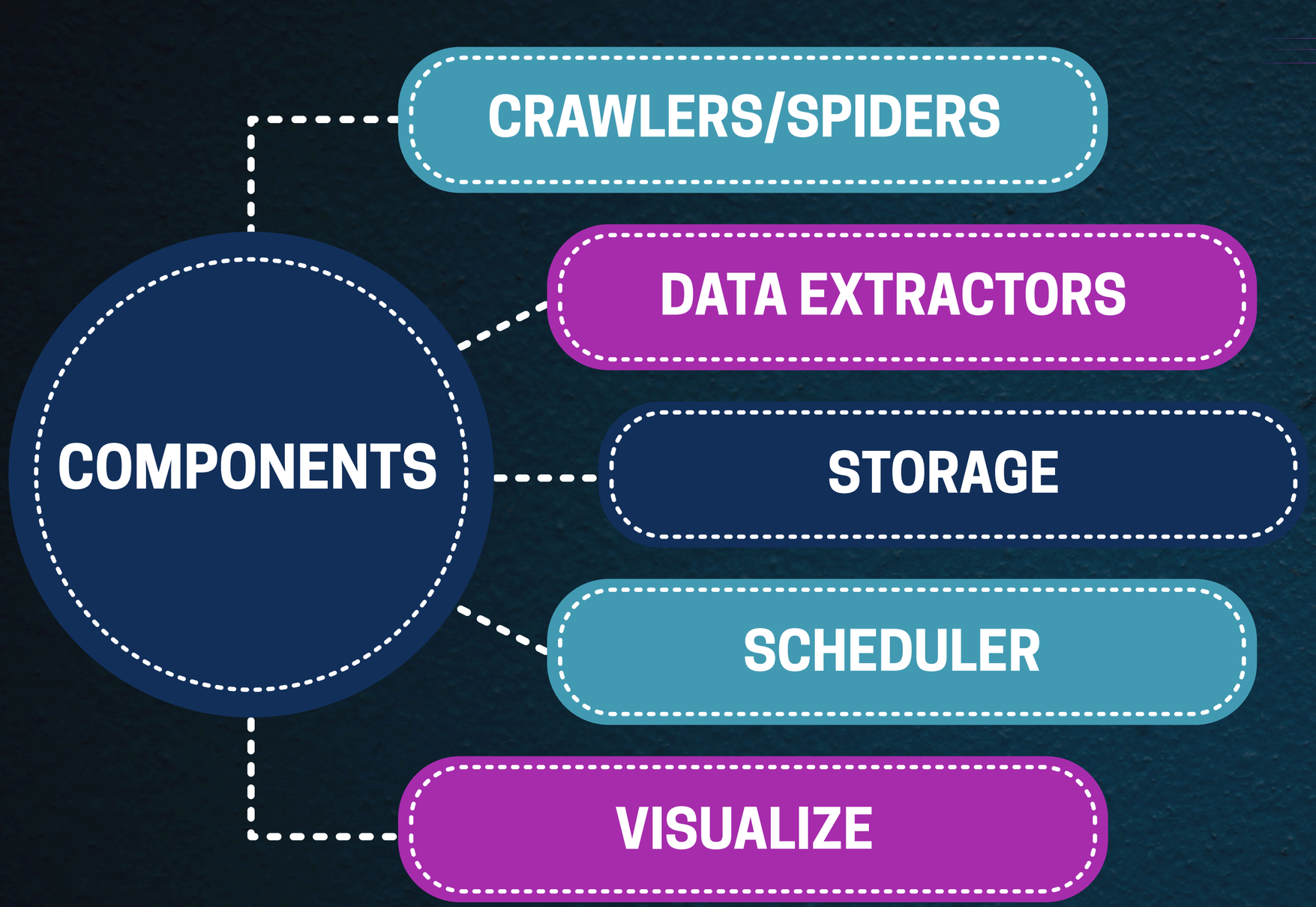
Main Concept

03

Solution

ABOUT WEB SCRAPING

Web scraping is the automated process of extracting data from websites. It involves using a program known as a web scraper, which visits web pages, interprets their content, and collects the needed information. This extracted data is then organized into structured formats like Excel files, JSON, or XML, making it easy to use in spreadsheets, applications, or databases. A web scraper typically includes several core components, each designed to handle different parts of the data extraction workflow.


COMPONENTS
CRAWLERS/SPIDERS
DATA EXTRACTORS
STORAGE
SCHEDULER
VISUALIZE

COMPONENTS

VISIONCRAWLERS/SPIDERS

Crawlers, or spiders, are the first step in the web scraping process. These automated bots are programmed to surf the internet to find and retrieve web pages.

DATA EXTRACTORS

Once a crawler retrieves a web page, the data extractor takes over. This component is responsible for parsing the HTML, XML, or JSON content of the page to identify and extract the information of interest

STORAGE

The extracted data needs to be stored for further analysis or use. This component of the web scraper determines how and where data is saved.

SCHEDULER

The scheduler is the orchestrator of the scraping process. It manages the queue of URLs to be visited, prioritizing them based on predefined criteria to ensure an efficient scraping process.

VISUALIZE

The results of the analysis are visualized using charts, graphs to help understand the data.

FUNCTIONAL FLOW OF WEB SCRAPING

The process of web scraping follows a logical sequence of steps, ensuring that data is collected systematically and efficiently.

1- URL Identification

The first step involves identifying the URLs of the pages to scrape. This list can be generated dynamically or based on a predefined set of targets.

2- Request Sending

The scraper sends HTTP requests to the web servers hosting the target pages. Here, techniques like rotating user agents or using proxy servers might be employed to mimic human behavior and avoid detection.

3- Data Parsing & Extraction

Upon receiving the web page content, the scraper parses the HTML or other formats to locate the data of interest. Libraries like Beautiful Soup for Python allows for easy navigation of the DOM tree.

4- Data Storage & Analysis

The extracted information is then saved in the chosen format. Subsequent analysis can transform this data into valuable insights, such as trend analysis in social media posts or price monitoring across e-commerce platforms.

KEY ALGORITHMS & SOLUTIONS

HTTP/HTTPS Protocol

Governs the request/ response mechanism between the scraper and the web server.

Parsing Algorithms

DOM Parsing (Document Object Model traversal)

Tree Traversal (depth-first or breadth-first to explore the HTML tree)

Robots Exclusion Protocol

Standard that tells web scrapers which pages or sections they are allowed or forbidden to scrape.

Rate Limiting Algorithms

Techniques like token bucket algorithms can control request rates to avoid overloading the server or getting blocked.

OUR SOLUTION

It’s Time to
Break Now!

SENDING HTTP REQUESTS TO THE WEBSITE

Before extracting any data, our scraper must first connect to the website by sending an HTTP request to the server that hosts the website. Once the request is sent, the server processes it and returns an HTML response. This is the content we will parse to extract the data we need.

PARSING THE HTML RESPONSE

After receiving the HTML content from the server, the next step is to parse it. This is the process where we read and understand the structure of HTML so that specific pieces of data can be located. By parsing the HTML, we can extract data from elements that are typically used for listing vehicles, such as the name, price, location, color and other car characteristics.

DATA EXTRACTION AND ORGANIZATION

Once we have identified the correct elements in the HTML structure, the next step is to extract the data and organize it in a meaningful way.

- **Data Points:** For the "Véhicules" section, typical data points include: Vehicule Name, Price, Location, Gearbox, state, Year, Brand, Model, Fuel .
- **Data Storage Format:** We will store extracted data in a structured format, specifically in pandas DataFrame (since we are using Python), which is a table-like structure. This makes it easy to manage, clean, and process the data.

HANDLING MULTIPLE PAGES

tayara.tn has listings spread across multiple pages. To get all the relevant data, our scraper needs to handle multiple pages.

- **Pagination Logic:** The URL for each page is usually structured with a query parameter like ?page=1, ?page=2, etc. we will iterate over these page numbers and collect data from each one.
- **Data Aggregation:** As data is collected from multiple pages, it is aggregated into a single dataset (e.g., a large list of vehicle listings) that can be stored and processed together.

LEGAL CONSIDERATIONS FOR WEB SCRAPING

1. **Intellectual Property:** All content on Tayara is protected by intellectual property rights. Scraping or using content without permission may violate these rights.
2. **Data Privacy:** Personal data collected through scraping must comply with privacy laws. Unauthorized scraping of personal data is prohibited.
3. **Security:** Scraping tools must not interfere with the website’s normal operation. Actions that bypass security measures or violate user privacy are not allowed.
4. **Cookies and Tracking:** Web scraping that interacts with cookies or tracking technologies must be in compliance with privacy policies and user consent.

STORING DATA IN AN SQL DATABASE

Once the data in a structured format, the next step is to store it in a SQL database. This allows for more powerful querying, searching, and manipulation of the data later. We chose MySQL/PostgreSQL since they are more robust relational databases for larger projects.

- **Table Creation**: Before inserting the data, we will create a table in the database to hold it. The table schema should match the structure of the data we are scraping.
  - Name (string)
  - Price (string or numeric)
  - Location (string)
  - Gearbox (string)
  - Vehicle state (string)
  - Year (numeric)
  - Brand (string)
  - Model (string)
  - Fuel (string)

- **Data Insertion**: Once the table is created, we will insert the data using SQL queries. In Python, this can be done using libraries like:
  - **SQLAlchemy** is an Object Relational Mapper (ORM) that allows you to interact with the database using Python objects.
  - **Pandas** has a built-in method .to_sql() that can directly upload a DataFrame into the SQL database.

ADVANTAGES

OF WEB SCRAPING ON TAYARA

- **Automated Data Collection**: Web scraping allows for the automatic extraction of large amounts of data, saving significant time and effort compared to manual collection.

- **Market Analysis**: By scraping data on prices, and product availability, businesses can gain insights into market trends, competitive pricing, and customer behavior on Tayara.

- R**eal-Time Data**: Web scraping provides access to up-to-date listings and offers on Tayara, allowing users to access fresh information for analysis or decision-making.

- L**ead Generation**: Scraping user and product information can help businesses gather leads for potential customers or vendors interested in Tayara’s marketplace.

LIMITATIONS

OF WEB SCRAPING ON TAYARA

- **Legal and Ethical Issues:** As mentioned in Tayara's Terms of Use, scraping without permission may violate intellectual property rights and user agreements, leading to legal consequences.

- **Website Stability:** Frequent scraping can put a load on Tayara’s servers, potentially causing slowdowns, downtime, or disruptions to the service for other users.

- **Data Quality and Structure:** The data collected via scraping may be incomplete, poorly structured, or inconsistent, requiring additional work to clean and organize it.

- **Security Risks:** Scraping tools can unintentionally breach security protocols or access sensitive data, violating privacy regulations or leading to unauthorized access to personal information.

- **Ban or Blocking:** Websites like Tayara may implement measures to detect and block scraping activities, limiting or completely cutting off access to scraped data.

THANK YOU

STAY TUNED FOR THE NEXT STEP !