Portfolio.

Dimas Adi Saputra

Data Engineer

dimadisaputra@gmail.com

linkedin.com/in/dimadisaputra

instagram.com/dimadisaputra

github.com/dimadisaputra

dimadisaputra.github.io

Caridulu.

Final Assignment at Madiun State Polytechnic
February - May 2024

Description

This project aims to develop Caridulu, a product search website designed to compare products from various marketplaces in Indonesia. Currently supporting Shopee Indonesia, Tokopedia, and Lazada Indonesia, Caridulu provides users with the ability to compare product offerings across multiple platforms.

Project Steps

- Web Scraping for Product Data: Implement Python and BeautifulSoup for scraping product search results from multiple marketplace platforms to retrieve product name, price, rating, units sold, seller location, product image, and URL.
- Frontend Development: Utilize React Vite to develop a responsive and interactive interface that displays search results fetched from web scraping across various marketplace platforms.
- Backend Development: Develop backend functionalities using FastAPI to facilitate communication between the frontend website and the web scraping processes.

Technologies Used

- BeautifulSoup
- FastAPI
- JavaScript
- PostgreSQL
- Python
- React + Vite

Newspaper Broad Crawl.

Data Engineer at Nolimit Indonesia
November 2023

Description

The objective of this project is to scrape and parse articles from various online media portal links or domains provided in a database table. Examples of inputs for this scraper include links such as detik.com, antara.com, and others. The expected output from scraping articles includes link, domain, title, content, date and author. The scraped data will be converted into JSON format and produced to Kafka for further processing.

Project Steps

- Input Source: Fetch a list of links or domains from a designated database table.
- Scraping and Parsing: Perform scraping and parsing of articles from each provided link or domain to extract the specified data fields: Link, Domain, Title, Content, Date, and Author.
- Data Conversion: Convert the scraped data into JSON format.
- Data Streaming: Produce the JSON formatted data to Kafka for further processing.

Technologies Used

- Apache Kafka
- Newspaper3k
- PostgreSQL
- Python

Google Maps Reviews Scraper.

Data Engineer at Nolimit Indonesia
September 2023

Description

The objective of this project is to scrape location and review data from Google Maps, including location name, rating, comments (reviewer name, review date, review content), and location metadata such as coordinates, link to image, and other relevant details.

Project Steps

- Input Location Link: The program will use the provided link to scrape data from Google Maps.
- Data Scraping: Collecting information such as location name, rating, comments (reviewer's name, review date, review content), and location metadata (coordinates, link to image, and others).
- Data Storage: Successfully scraped data can be stored in JSON, CSV, or XLSX format (JSON as default) in a data buffer for further processing.

Technologies Used

- Beautiful Soup
- Python
- Selenium

TikTok Comments Scraper.

Data Engineer at Nolimit Indonesia
August - September 2023

Description

This project aims to scrape comments from specific TikTok videos. The data collected includes comments, username, date, video ID, author username, and comment count.

Project Steps

- Checking Task Status: The program checks the database for tasks with a pending status.
- Performing Scraping: If pending tasks are found, the program scrapes comments from the TikTok videos.
- Updating Task Status: The task status is changed to running during scraping. After scraping, the task status is updated to success or error in the database.
- Collecting Data: The collected data includes comments, username, date, video ID, author username, and comment count.
- Converting Data to JSONL: The collected data is converted into JSONL format.
- Streaming Data to Kafka: The JSONL formatted data is then produced/streamed to Kafka for further processing.

Technologies Used

- Apache Kafka
- Playwright
- PostgreSQL
- Python

Online Media Portal Scraper.

Data Engineer at Nolimit Indonesia
June - July 2023

Description

The objective of this project is to extract article links from over 1000 online media portals, such as CNN Indonesia, Kompas.com, Detik.com, and others. This project utilizes the Scrapy Framework for web scraping. The collected data is converted into JSON, JSONL, CSV, or XLSX formats and stored in a data buffer for further processing.

Project Steps

- Data Collection: Using Scrapy to crawl and scrape news article links from over 1000 online media portals.
- Data Processing: Converting the collected data into the desired formats (JSON, JSONL, CSV, XLSX).
- Data Storage: Storing the data in a buffer for further processing.

Technologies Used

- Python
- Scrapy Framework