Y Combinator Company Data Collection Report

Task: Collect structured data from Y Combinator company pages.

Data Source: https://www.ycombinator.com/companies

Web Scraping Methodology

Web scraping involves extracting data from websites using code. Python is a popular choice for data analysis, and I leveraged its rich ecosystem of libraries for this project.

Data Extraction Process

1. **Batch Identification and URL Collection:**
    - Identified various batches on the YC website, each containing hundreds of companies.
    - Extracted batch names to facilitate grouping company URLs.
    - Used BeautifulSoup for efficiently collecting URLs from static content.
2. **URL Batching and Processing:**
    - Due to the large number of companies (4600+), scraping all URLs at once was deemed inefficient.
    - Batched URLs into groups of 1147 for manageable processing.
    - Sequentially iterated through each batch for data extraction.
3. **Dynamic Content Handling:**
    - Playwright library was employed for handling dynamic content encountered while collecting individual company URLs.
    - This approach ensured adaptability to changing website structures.
4. **Data Extraction and Storage:**
    - Extracted data point-by-point from each company webpage.
    - Utilized pandas DataFrames to structure the collected data.
    - Saved data for each batch as separate CSV files for organization.
    - Subsequently, these CSVs were combined into a single, comprehensive CSV file.
5. **Data Transformation (CSV to JSON):**
    - Leveraged Python's capabilities to convert the final CSV file into a JSON format, as per project requirements.

Challenges and Solutions

- **Timeout Errors:** Frequent timeout issues occurred during scraping with Playwright, likely due to the high volume of requests.
    - Implemented a distributed scraping approach by processing data in batches. This slowed down the process but ensured successful data collection.
    - Increased the default timeout value for webpages, mitigating some timeout occurrences.
    - Identified and separately scraped URLs that still encountered timeouts.

Future Considerations

- Explore more robust scraping frameworks or techniques to optimize scraping speed and handle potential website changes in a commercial setting.

Overall, this project successfully collected structured data from Y Combinator company pages, demonstrating proficiency in web scraping methodologies and data manipulation techniques.