Company Scraper

I am creating a custom company scraper to check to see if certain firm sites contain keywords.

1. **Load the Excel File and Focus on the “US Pre-Seed” Sheet**

- Focus on the rows *Company* and *Url*

2. **List of Keywords to Scrape**

- Private Equity
- Capital Markets
- Leverage Finance
- Investment Banking
- b2b saas
- pre-seed
- Southeast
- Latin
- Hispanic
- Florida

The scraper should check for the presenece of each keyword on the website and return a “Yes” or “No” for each keyword. Additionally, if the keyword is found, the specific location or context where found should be included in a field.

4. **Accessing the Excel File and Fetching each Company && Site**

- The scraper should access the specified “US Pre-Seed” sheet. And fetch each website individually.

E.g.,

Vine St. Ventures || www.vinestventures.com

The scraper should then fetch the website, and find a way to access each page and analyze all web content to check for the presence of the specified keywords.

5. **Parsing Each Page On Each Site**

- How can we allow the scraper to search through every available page or link on a specific site?
    - Intiial Page Load: Loading the landing page and get the HTML content
    - Parsing the Pages: Extract all internal links.
    - Crawling Internal Pages: Once all links are extracted, send additonal requests to each of the internal pages and extract all HTML content.
        - Recursive Crawling: Repeaat the process of extracting links from each of these pages. e.g.,This is necessary to ensure that nested pages (e.g., a "Team" page linked from the "About" page) are also crawled.

    - ## Searching For Keywords: On each page, including the internal pages, search for the spefied keywords. - Keywords: Private Equity, Capital Markets, Leverage Finance, Investment Banking,

"b2b saas", "pre-seed", Southeast, Latin, Hispanic, Florida.

-

.

6. **Creating the Output File**