Best performance and scalability

- Scrapy (does both; parse HTML and connect to a webpage. However can't scrape javascript)

For parsing HTML data

- Beautiful Soup (some performance concerns larger scale, can't scrape javascript websites

Connecting to webpage

- HTTPX

Getting data from javascript websites requires the use of a browser instance, this can be achieved with a library like Selenium. This should be used sparingly however because of performance considerations. Selenium is often used in tandem with another library like Beautiful Soup.

Another use of selenium is the ability to emulate user interaction with the webpage. For example, we can fill out forms and click buttons to access data that might not be immediately accessible. This aligns with our potential need to navigate a page to reach the PDFs containing our data.

For scraping PDFs:
PDFMiner can extract text from pdf to a text file, we can then parse it for keywords.
Tabula to extract tables within PDFs.