webcrawler

October 4, 2024

[14]:
```
import requests
from bs4 import BeautifulSoup
from urllib.parse import urljoin, urlparse
import time


crawled_data = {
    "pages": [],
    "links": {},
}

visited = set()

def crawl(url, depth):
    if depth == 0 or url in visited:
        return

    visited.add(url)
    headers = {
        'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/
 ↪537.36 (KHTML, like Gecko) Chrome/58.0.3029.110 Safari/537.3'
    }

    try:
        response = requests.get(url, headers=headers)
        if response.status_code != 200:
            print(f"Failed to retrieve {url} (status code: {response.
 ↪status_code})")
            return

        soup = BeautifulSoup(response.text, 'lxml')
        print(f"Crawling: {url}")

        crawled_data["pages"].append(url)
        crawled_data["links"][url] = []

        for link in soup.find_all('a', href=True):
```

```
            next_url = urljoin(url, link['href'])

            if urlparse(url).netloc == urlparse(next_url).netloc:
                crawled_data["links"][url].append(next_url)
                crawl(next_url, depth - 1)

        time.sleep(1)

    except Exception as e:
        print(f"Error crawling {url}: {e}")


def print_stats():
    print("\n--- Crawling Stats ---")
    total_pages = len(crawled_data["pages"])
    total_links = sum(len(links) for links in crawled_data["links"].values())

    print(f"Total pages crawled: {total_pages}")
    print(f"Total links found: {total_links}")

if __name__ == "__main__":
    start_url = input("Enter the URL to crawl: ")
    depth_limit = int(input("Enter the crawling depth limit (e.g., 2): "))


    crawl(start_url, depth_limit)
    print_stats()
```

```
Enter the URL to crawl:  https://books.toscrape.com/
Enter the crawling depth limit (e.g., 2):  2

Crawling: https://books.toscrape.com/
Crawling: https://books.toscrape.com/index.html
Crawling: https://books.toscrape.com/catalogue/category/books_1/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/travel_2/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/mystery_3/index.html
Crawling: https://books.toscrape.com/catalogue/category/books/historical-
fiction_4/index.html
Crawling: https://books.toscrape.com/catalogue/category/books/sequential-
art_5/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/classics_6/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/philosophy_7/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/romance_8/index.html
```

```
Crawling: https://books.toscrape.com/catalogue/category/books/womens-
fiction_9/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/fiction_10/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/childrens_11/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/religion_12/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/nonfiction_13/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/music_14/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/default_15/index.html
Crawling: https://books.toscrape.com/catalogue/category/books/science-
fiction_16/index.html
Crawling: https://books.toscrape.com/catalogue/category/books/sports-and-
games_17/index.html
Crawling: https://books.toscrape.com/catalogue/category/books/add-a-
comment_18/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/fantasy_19/index.html
Crawling: https://books.toscrape.com/catalogue/category/books/new-
adult_20/index.html
Crawling: https://books.toscrape.com/catalogue/category/books/young-
adult_21/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/science_22/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/poetry_23/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/paranormal_24/index.html
Crawling: https://books.toscrape.com/catalogue/category/books/art_25/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/psychology_26/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/autobiography_27/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/parenting_28/index.html
Crawling: https://books.toscrape.com/catalogue/category/books/adult-
fiction_29/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/humor_30/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/horror_31/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/history_32/index.html
Crawling: https://books.toscrape.com/catalogue/category/books/food-and-
```

```
drink_33/index.html
Crawling: https://books.toscrape.com/catalogue/category/books/christian-
fiction_34/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/business_35/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/biography_36/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/thriller_37/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/contemporary_38/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/spirituality_39/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/academic_40/index.html
Crawling: https://books.toscrape.com/catalogue/category/books/self-
help_41/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/historical_42/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/christian_43/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/suspense_44/index.html
Crawling: https://books.toscrape.com/catalogue/category/books/short-
stories_45/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/novels_46/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/health_47/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/politics_48/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/cultural_49/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/erotica_50/index.html
Crawling:
https://books.toscrape.com/catalogue/category/books/crime_51/index.html
Crawling: https://books.toscrape.com/catalogue/a-light-in-the-
attic_1000/index.html
Crawling: https://books.toscrape.com/catalogue/tipping-the-velvet_999/index.html
Crawling: https://books.toscrape.com/catalogue/soumission_998/index.html
Crawling: https://books.toscrape.com/catalogue/sharp-objects_997/index.html
Crawling: https://books.toscrape.com/catalogue/sapiens-a-brief-history-of-
humankind_996/index.html
Crawling: https://books.toscrape.com/catalogue/the-requiem-red_995/index.html
Crawling: https://books.toscrape.com/catalogue/the-dirty-little-secrets-of-
getting-your-dream-job_994/index.html
Crawling: https://books.toscrape.com/catalogue/the-coming-woman-a-novel-based-
```

```
on-the-life-of-the-infamous-feminist-victoria-woodhull_993/index.html
Crawling: https://books.toscrape.com/catalogue/the-boys-in-the-boat-nine-
americans-and-their-epic-quest-for-gold-at-the-1936-berlin-
olympics_992/index.html
Crawling: https://books.toscrape.com/catalogue/the-black-maria_991/index.html
Crawling: https://books.toscrape.com/catalogue/starving-hearts-triangular-trade-
trilogy-1_990/index.html
Crawling: https://books.toscrape.com/catalogue/shakespeares-
sonnets_989/index.html
Crawling: https://books.toscrape.com/catalogue/set-me-free_988/index.html
Crawling: https://books.toscrape.com/catalogue/scott-pilgrims-precious-little-
life-scott-pilgrim-1_987/index.html
Crawling: https://books.toscrape.com/catalogue/rip-it-up-and-start-
again_986/index.html
Crawling: https://books.toscrape.com/catalogue/our-band-could-be-your-life-
scenes-from-the-american-indie-underground-1981-1991_985/index.html
Crawling: https://books.toscrape.com/catalogue/olio_984/index.html
Crawling: https://books.toscrape.com/catalogue/mesaerion-the-best-science-
fiction-stories-1800-1849_983/index.html
Crawling: https://books.toscrape.com/catalogue/libertarianism-for-
beginners_982/index.html
Crawling: https://books.toscrape.com/catalogue/its-only-the-
himalayas_981/index.html
Crawling: https://books.toscrape.com/catalogue/page-2.html

--- Crawling Stats ---
Total pages crawled: 74
Total links found: 4401
```