ABBOTTABAD UNIVERSITY OF SCIENCE AND TECHNOLOGY

Project

Submitted To:	Sir Jamal Abdul Ahad
Submitted By:	Ayesha Jadoon
Roll no:	14638
Subject:	Data Structures
Semester:	BSCS (3rd A)
Date:	1/2/2024

Project Name: Web Crawler

Description:

This Web Crawler is a Python-based tool that automatically visits web pages, extracts useful information, and saves it in a structured format. It follows links on a website up to a specified depth and ensures responsible crawling by following robots.txt rules.

Features:

- **Fetches web pages** and extracts titles, headings, paragraphs, and images.
- **Follows links** within the same domain up to a given depth.
- **Avoids duplicate pages** using content hashing.
- **Checks robots.txt** to ensure ethical crawling.
- **Saves extracted data** in a JSON file.

DSA Concepts Used:

1. **Queue**
    - Used deque to store and process URLs level by level (FIFO).
2. **Set (For Fast Lookup of Visited URLs)**
    - Used set() to avoid revisiting the same URLs.
3. **Hashing**
    - Used hashing to detect duplicate content and avoid storing it again.
4. **Dictionary (For Data Storage)**
    - Used dictionary to store extracted content (title, paragraphs, images) in a structured format.

Output:

- **Enter URL**: The program asks you to input the website URL to crawl.
- **Crawling Process**: It starts crawling the website, showing messages like "[Crawling] URL" for each page and skipping blocked or duplicate pages.

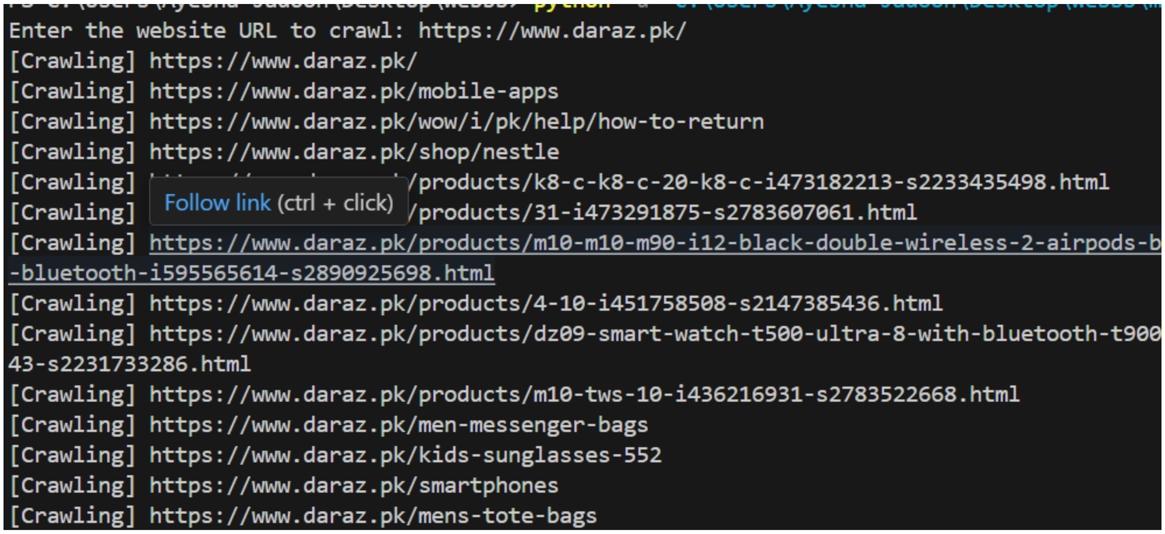
- **Completion:** Once done, it displays the total number of URLs crawled and links extracted, and saves the data in a JSON file.
- **SON Output:** The crawled data, including URLs, titles, headings, and images, is saved in a JSON file.

Tested on the following websites:

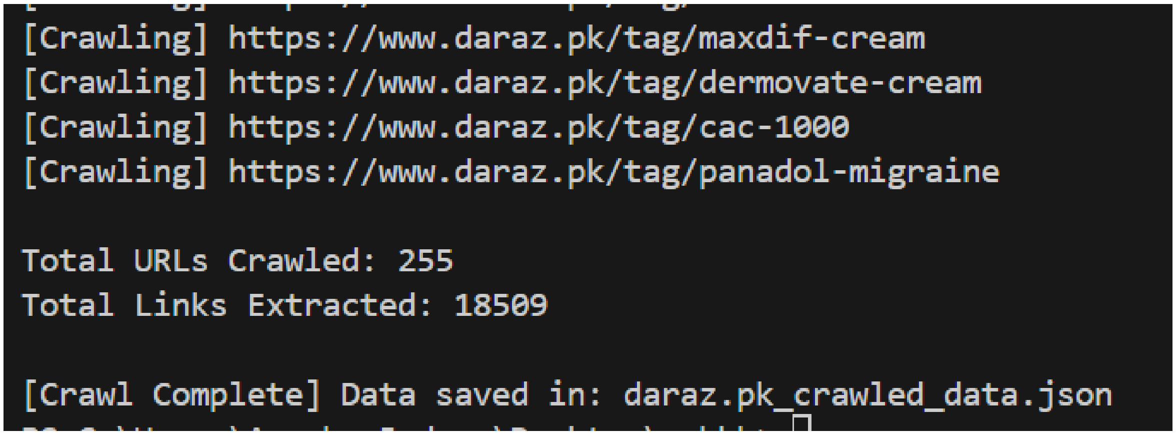
- https://www.daraz.pk/
- http://quotes.toscrape.com/
- http://books.toscrape.com/
- http://commoncrawl.org.com/
- https://www.google.com

Output Screenshots:

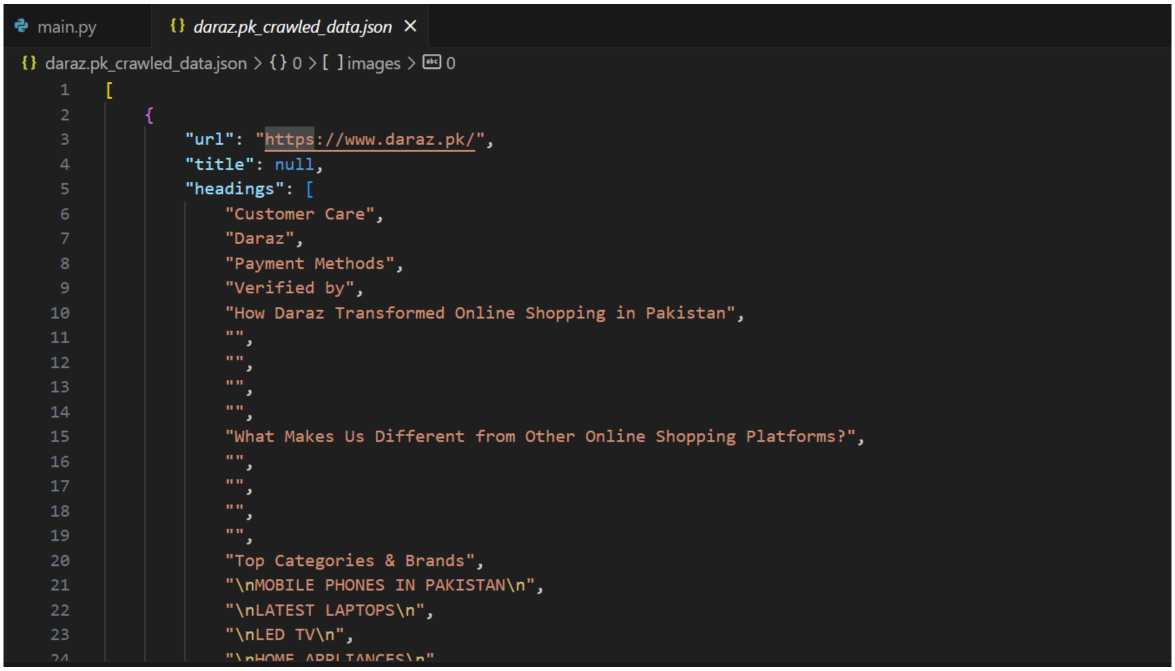
- **For** https://www.daraz.pk/



```
Enter the website URL to crawl: https://www.daraz.pk/
[Crawling] https://www.daraz.pk/
[Crawling] https://www.daraz.pk/mobile-apps
[Crawling] https://www.daraz.pk/wow/i/pk/help/how-to-return
[Crawling] https://www.daraz.pk/shop/nestle
[Crawling]                     /products/k8-c-k8-c-20-k8-c-i473182213-s2233435498.html
[Crawling]                     /products/31-i473291875-s2783607061.html
[Crawling] https://www.daraz.pk/products/m10-m10-m90-i12-black-double-wireless-2-airpods-b
-bluetooth-i595565614-s2890925698.html
[Crawling] https://www.daraz.pk/products/4-10-i451758508-s2147385436.html
[Crawling] https://www.daraz.pk/products/dz09-smart-watch-t500-ultra-8-with-bluetooth-t900
43-s2231733286.html
[Crawling] https://www.daraz.pk/products/m10-tws-10-i436216931-s2783522668.html
[Crawling] https://www.daraz.pk/men-messenger-bags
[Crawling] https://www.daraz.pk/kids-sunglasses-552
[Crawling] https://www.daraz.pk/smartphones
[Crawling] https://www.daraz.pk/mens-tote-bags
```

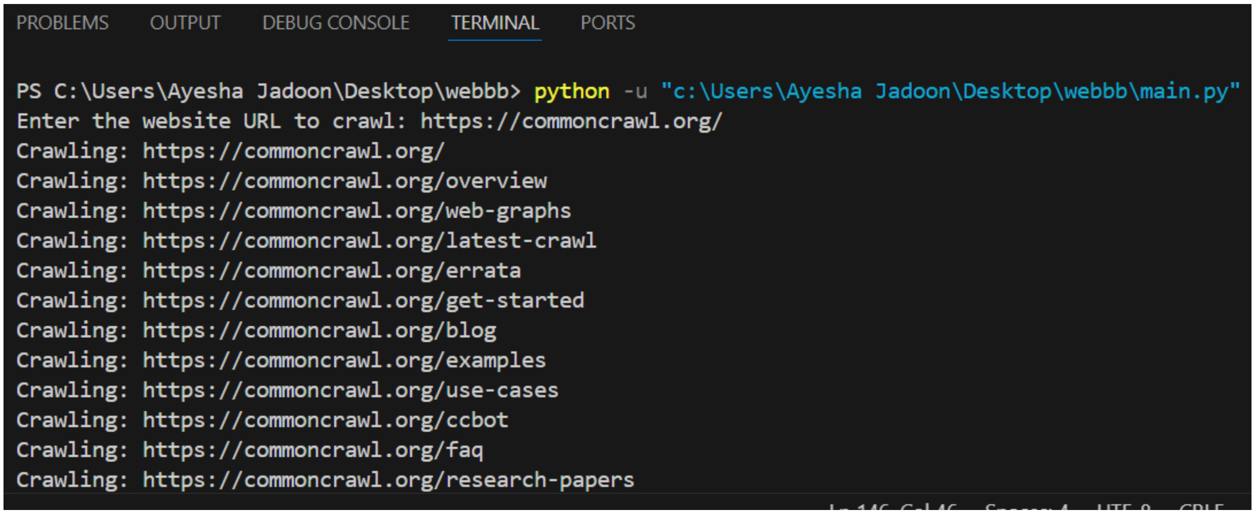


```
[Crawling] https://www.daraz.pk/tag/maxdif-cream
[Crawling] https://www.daraz.pk/tag/dermovate-cream
[Crawling] https://www.daraz.pk/tag/cac-1000
[Crawling] https://www.daraz.pk/tag/panadol-migraine

Total URLs Crawled: 255
Total Links Extracted: 18509

[Crawl Complete] Data saved in: daraz.pk_crawled_data.json
```

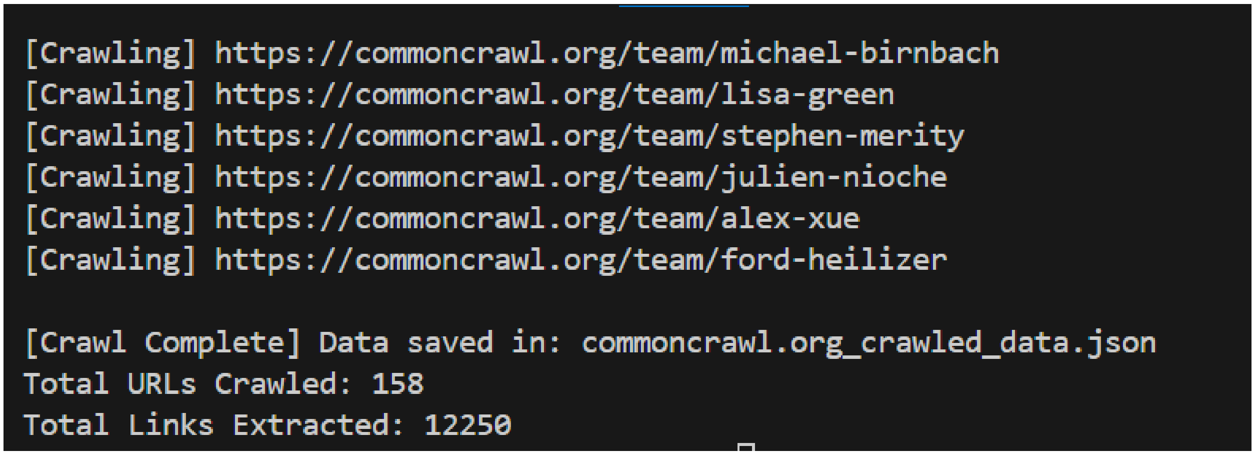

main.py
daraz.pk_crawled_data.json
daraz.pk_crawled_data.json > {} 0 > [ ] images > 0

```
[
    {
        "url": "https://www.daraz.pk/",
        "title": null,
        "headings": [
            "Customer Care",
            "Daraz",
            "Payment Methods",
            "Verified by",
            "How Daraz Transformed Online Shopping in Pakistan",
            "",
            "",
            "",
            "",
            "What Makes Us Different from Other Online Shopping Platforms?",
            "",
            "",
            "",
            "",
            "Top Categories & Brands",
            "\nMOBILE PHONES IN PAKISTAN\n",
            "\nLATEST LAPTOPS\n",
            "\nLED TV\n",
            "\nHOME APPLIANCES\n"
```

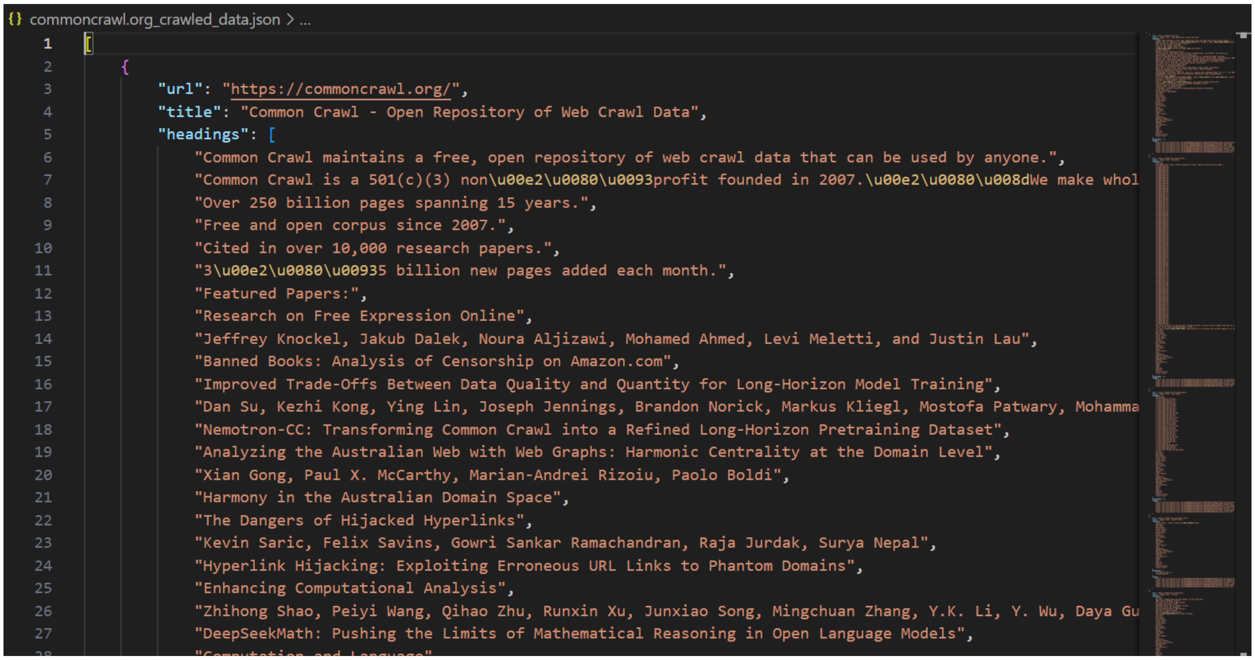
- **For** http://commoncrawl.org.com/



```
PS C:\Users\Ayesha Jadoon\Desktop\webbb> python -u "c:\Users\Ayesha Jadoon\Desktop\webbb\main.py"
Enter the website URL to crawl: https://commoncrawl.org/
Crawling: https://commoncrawl.org/
Crawling: https://commoncrawl.org/overview
Crawling: https://commoncrawl.org/web-graphs
Crawling: https://commoncrawl.org/latest-crawl
Crawling: https://commoncrawl.org/errata
Crawling: https://commoncrawl.org/get-started
Crawling: https://commoncrawl.org/blog
Crawling: https://commoncrawl.org/examples
Crawling: https://commoncrawl.org/use-cases
Crawling: https://commoncrawl.org/ccbot
Crawling: https://commoncrawl.org/faq
Crawling: https://commoncrawl.org/research-papers
```

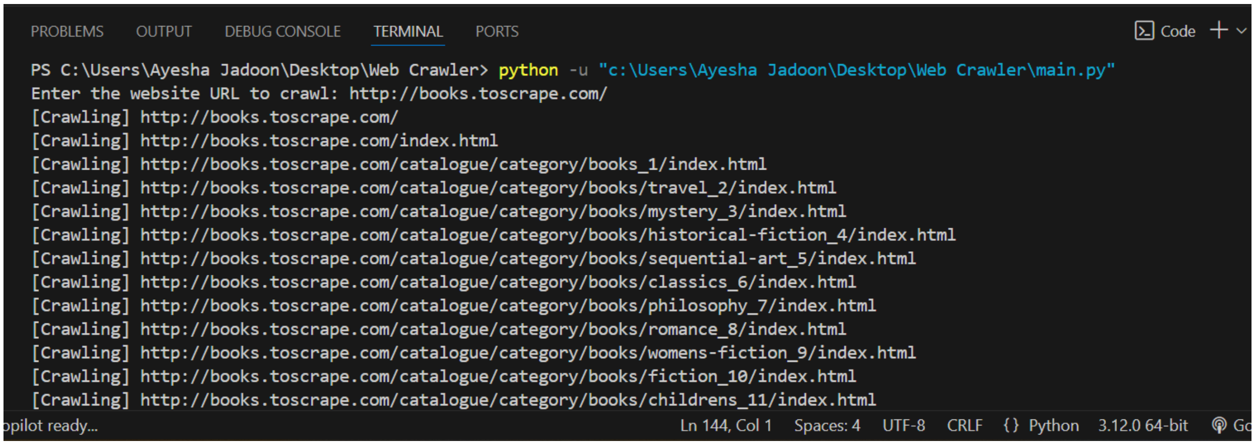


```
[Crawling] https://commoncrawl.org/team/michael-birnbach
[Crawling] https://commoncrawl.org/team/lisa-green
[Crawling] https://commoncrawl.org/team/stephen-merity
[Crawling] https://commoncrawl.org/team/julien-nioche
[Crawling] https://commoncrawl.org/team/alex-xue
[Crawling] https://commoncrawl.org/team/ford-heilizer

[Crawl Complete] Data saved in: commoncrawl.org_crawled_data.json
Total URLs Crawled: 158
Total Links Extracted: 12250
```

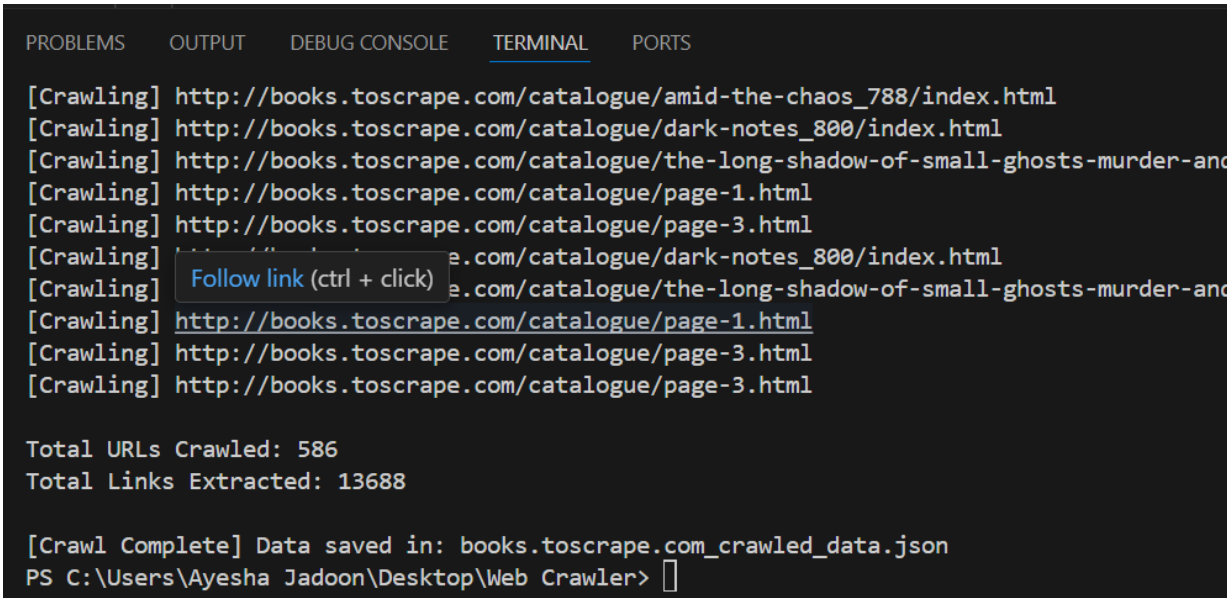


```
[
    {
        "url": "https://commoncrawl.org/",
        "title": "Common Crawl - Open Repository of Web Crawl Data",
        "headings": [
            "Common Crawl maintains a free, open repository of web crawl data that can be used by anyone.",
            "Common Crawl is a 501(c)(3) non\u00e2\u0080\u0093profit founded in 2007.\u00e2\u0080\u008dWe make whol
            "Over 250 billion pages spanning 15 years.",
            "Free and open corpus since 2007.",
            "Cited in over 10,000 research papers.",
            "3\u00e2\u0080\u00935 billion new pages added each month.",
            "Featured Papers:",
            "Research on Free Expression Online",
            "Jeffrey Knockel, Jakub Dalek, Noura Aljizawi, Mohamed Ahmed, Levi Meletti, and Justin Lau",
            "Banned Books: Analysis of Censorship on Amazon.com",
            "Improved Trade-Offs Between Data Quality and Quantity for Long-Horizon Model Training",
            "Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohamma
            "Nemotron-CC: Transforming Common Crawl into a Refined Long-Horizon Pretraining Dataset",
            "Analyzing the Australian Web with Web Graphs: Harmonic Centrality at the Domain Level",
            "Xian Gong, Paul X. McCarthy, Marian-Andrei Rizoiu, Paolo Boldi",
            "Harmony in the Australian Domain Space",
            "The Dangers of Hijacked Hyperlinks",
            "Kevin Saric, Felix Savins, Gowri Sankar Ramachandran, Raja Jurdak, Surya Nepal",
            "Hyperlink Hijacking: Exploiting Erroneous URL Links to Phantom Domains",
            "Enhancing Computational Analysis",
            "Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y.K. Li, Y. Wu, Daya Gu
            "DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models",
            "Computation and Language",
```

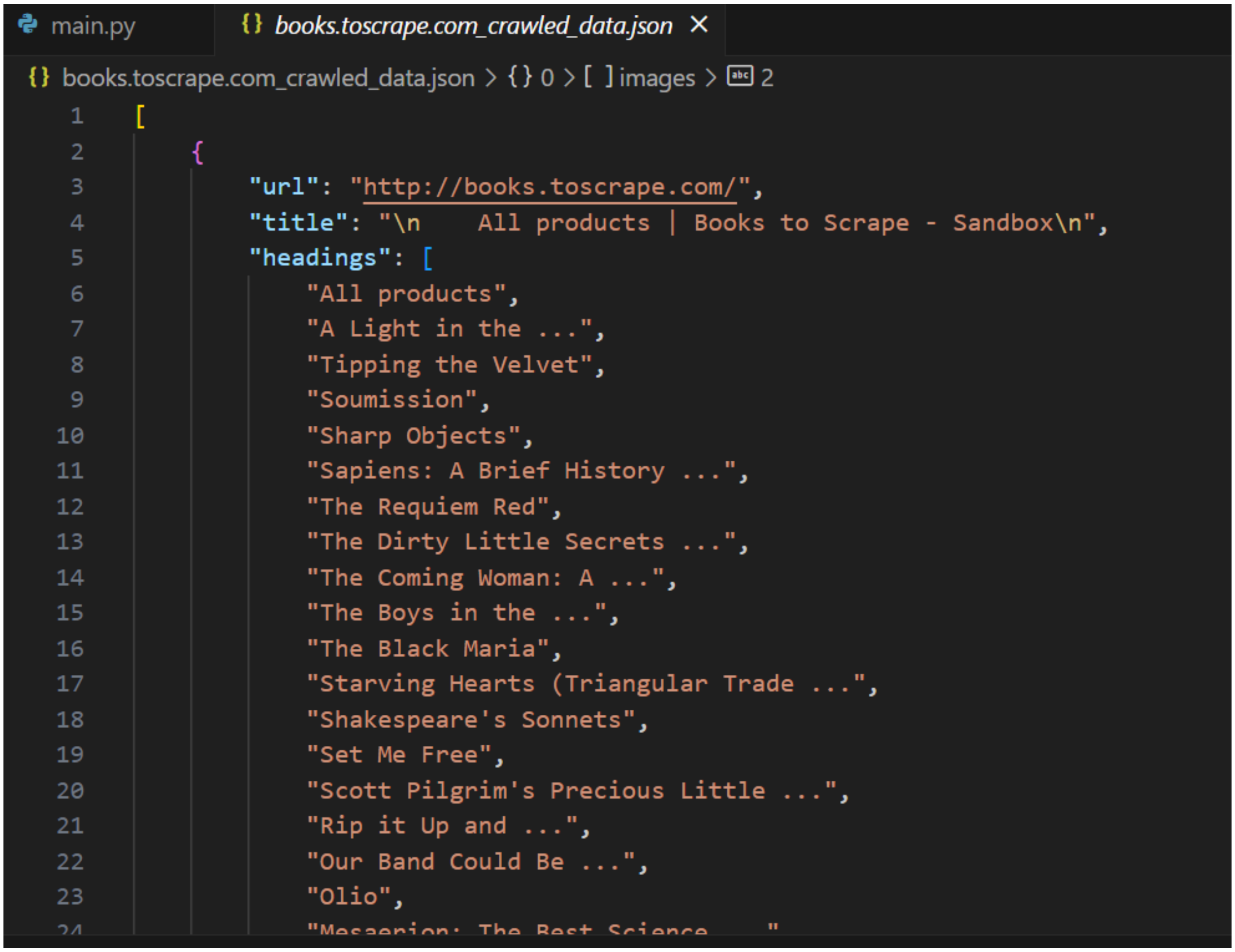
- **For** http://books.toscrape.com/



```
PS C:\Users\Ayesha Jadoon\Desktop\Web Crawler> python -u "c:\Users\Ayesha Jadoon\Desktop\Web Crawler\main.py"
Enter the website URL to crawl: http://books.toscrape.com/
[Crawling] http://books.toscrape.com/
[Crawling] http://books.toscrape.com/index.html
[Crawling] http://books.toscrape.com/catalogue/category/books_1/index.html
[Crawling] http://books.toscrape.com/catalogue/category/books/travel_2/index.html
[Crawling] http://books.toscrape.com/catalogue/category/books/mystery_3/index.html
[Crawling] http://books.toscrape.com/catalogue/category/books/historical-fiction_4/index.html
[Crawling] http://books.toscrape.com/catalogue/category/books/sequential-art_5/index.html
[Crawling] http://books.toscrape.com/catalogue/category/books/classics_6/index.html
[Crawling] http://books.toscrape.com/catalogue/category/books/philosophy_7/index.html
[Crawling] http://books.toscrape.com/catalogue/category/books/romance_8/index.html
[Crawling] http://books.toscrape.com/catalogue/category/books/womens-fiction_9/index.html
[Crawling] http://books.toscrape.com/catalogue/category/books/fiction_10/index.html
[Crawling] http://books.toscrape.com/catalogue/category/books/childrens_11/index.html
```

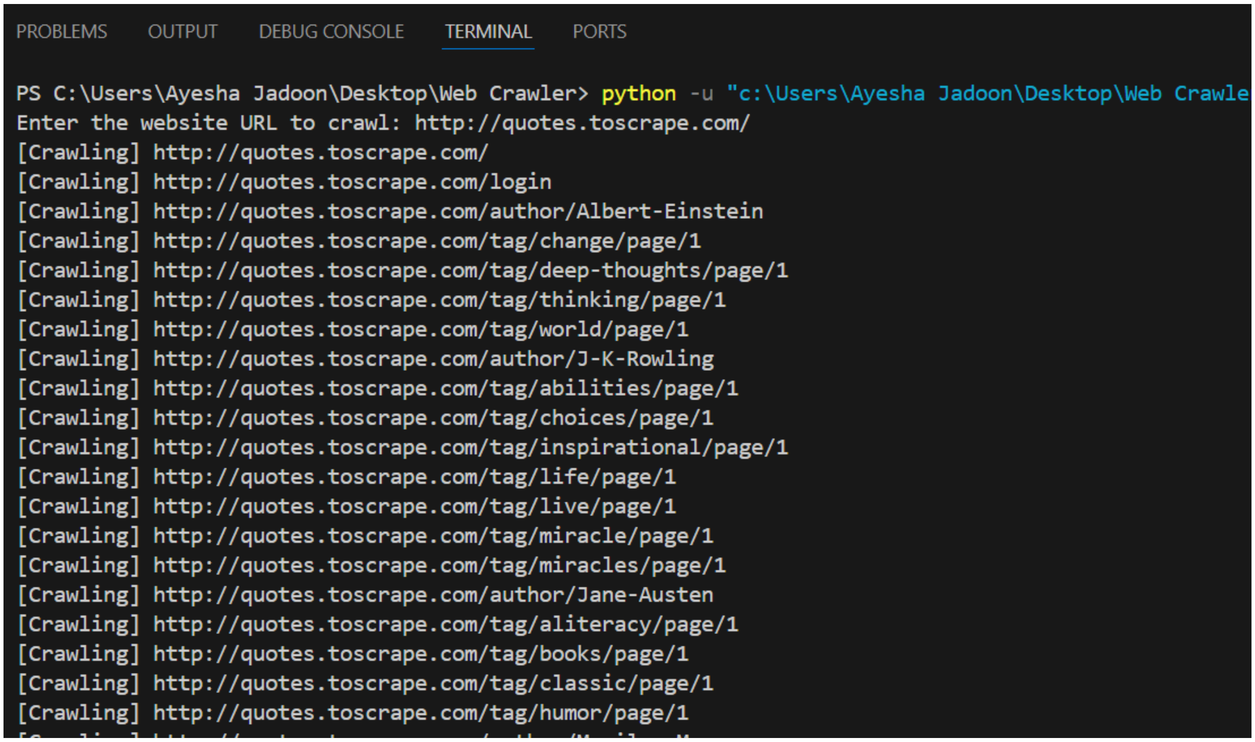


```
[Crawling] http://books.toscrape.com/catalogue/amid-the-chaos_788/index.html
[Crawling] http://books.toscrape.com/catalogue/dark-notes_800/index.html
[Crawling] http://books.toscrape.com/catalogue/the-long-shadow-of-small-ghosts-murder-and
[Crawling] http://books.toscrape.com/catalogue/page-1.html
[Crawling] http://books.toscrape.com/catalogue/page-3.html
[Crawling]                     e.com/catalogue/dark-notes_800/index.html
[Crawling]                     e.com/catalogue/the-long-shadow-of-small-ghosts-murder-and
[Crawling] http://books.toscrape.com/catalogue/page-1.html
[Crawling] http://books.toscrape.com/catalogue/page-3.html
[Crawling] http://books.toscrape.com/catalogue/page-3.html

Total URLs Crawled: 586
Total Links Extracted: 13688

[Crawl Complete] Data saved in: books.toscrape.com_crawled_data.json
PS C:\Users\Ayesha Jadoon\Desktop\Web Crawler>
```

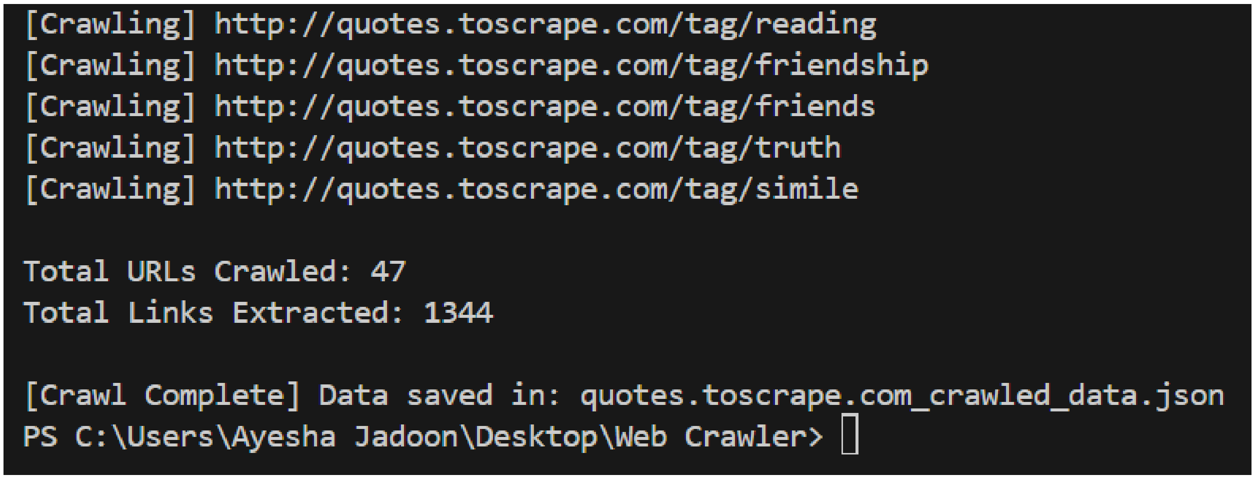

main.py
books.toscrape.com_crawled_data.json
books.toscrape.com_crawled_data.json > {} 0 > [ ] images > 2

```
[
    {
        "url": "http://books.toscrape.com/",
        "title": "\n    All products | Books to Scrape - Sandbox\n",
        "headings": [
            "All products",
            "A Light in the ...",
            "Tipping the Velvet",
            "Soumission",
            "Sharp Objects",
            "Sapiens: A Brief History ...",
            "The Requiem Red",
            "The Dirty Little Secrets ...",
            "The Coming Woman: A ...",
            "The Boys in the ...",
            "The Black Maria",
            "Starving Hearts (Triangular Trade ...",
            "Shakespeare's Sonnets",
            "Set Me Free",
            "Scott Pilgrim's Precious Little ...",
            "Rip it Up and ...",
            "Our Band Could Be ...",
            "Olio",
            "Mesaerion: The Best Science ...",
```

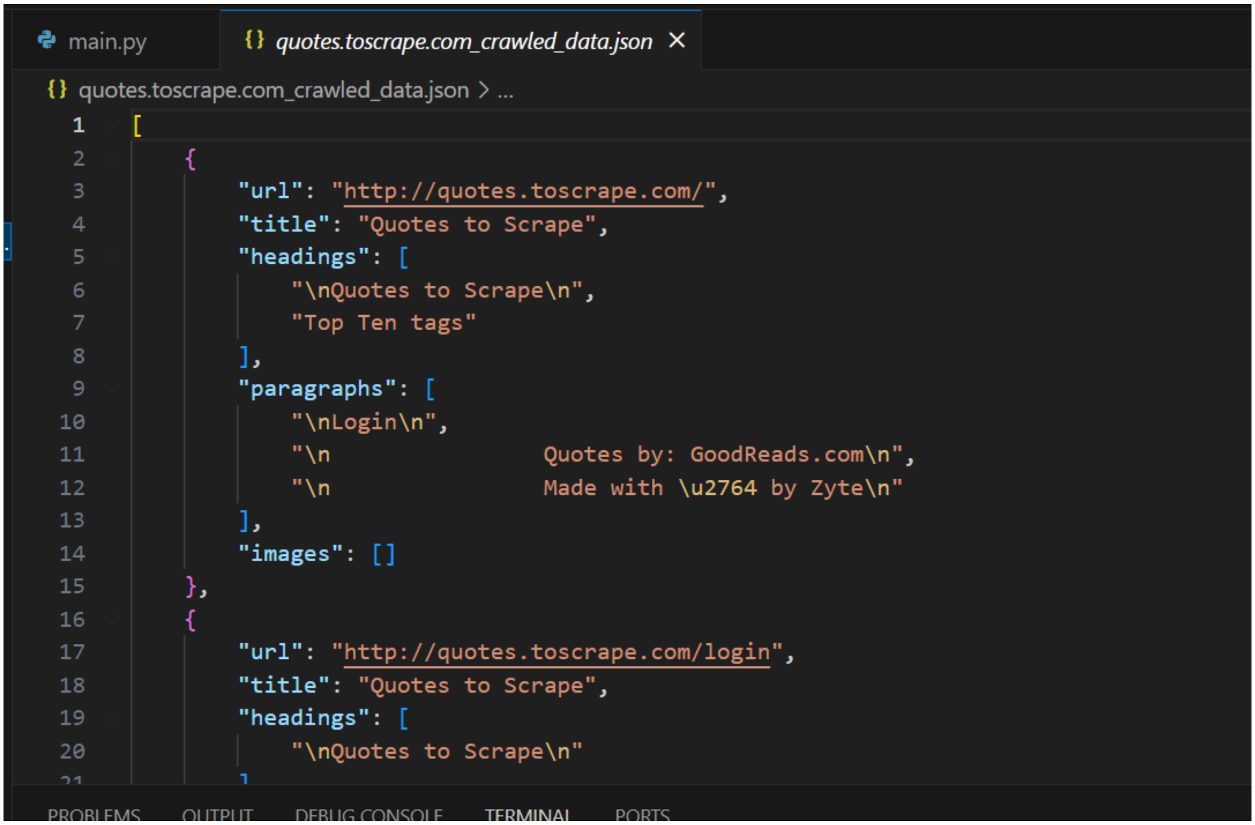
- **For** http://quotes.toscrape.com/



```
PS C:\Users\Ayesha Jadoon\Desktop\Web Crawler> python -u "c:\Users\Ayesha Jadoon\Desktop\Web Crawle
Enter the website URL to crawl: http://quotes.toscrape.com/
[Crawling] http://quotes.toscrape.com/
[Crawling] http://quotes.toscrape.com/login
[Crawling] http://quotes.toscrape.com/author/Albert-Einstein
[Crawling] http://quotes.toscrape.com/tag/change/page/1
[Crawling] http://quotes.toscrape.com/tag/deep-thoughts/page/1
[Crawling] http://quotes.toscrape.com/tag/thinking/page/1
[Crawling] http://quotes.toscrape.com/tag/world/page/1
[Crawling] http://quotes.toscrape.com/author/J-K-Rowling
[Crawling] http://quotes.toscrape.com/tag/abilities/page/1
[Crawling] http://quotes.toscrape.com/tag/choices/page/1
[Crawling] http://quotes.toscrape.com/tag/inspirational/page/1
[Crawling] http://quotes.toscrape.com/tag/life/page/1
[Crawling] http://quotes.toscrape.com/tag/live/page/1
[Crawling] http://quotes.toscrape.com/tag/miracle/page/1
[Crawling] http://quotes.toscrape.com/tag/miracles/page/1
[Crawling] http://quotes.toscrape.com/author/Jane-Austen
[Crawling] http://quotes.toscrape.com/tag/aliteracy/page/1
[Crawling] http://quotes.toscrape.com/tag/books/page/1
[Crawling] http://quotes.toscrape.com/tag/classic/page/1
[Crawling] http://quotes.toscrape.com/tag/humor/page/1
```



```
[Crawling] http://quotes.toscrape.com/tag/reading
[Crawling] http://quotes.toscrape.com/tag/friendship
[Crawling] http://quotes.toscrape.com/tag/friends
[Crawling] http://quotes.toscrape.com/tag/truth
[Crawling] http://quotes.toscrape.com/tag/simile

Total URLs Crawled: 47
Total Links Extracted: 1344

[Crawl Complete] Data saved in: quotes.toscrape.com_crawled_data.json
PS C:\Users\Ayesha Jadoon\Desktop\Web Crawler> 
```

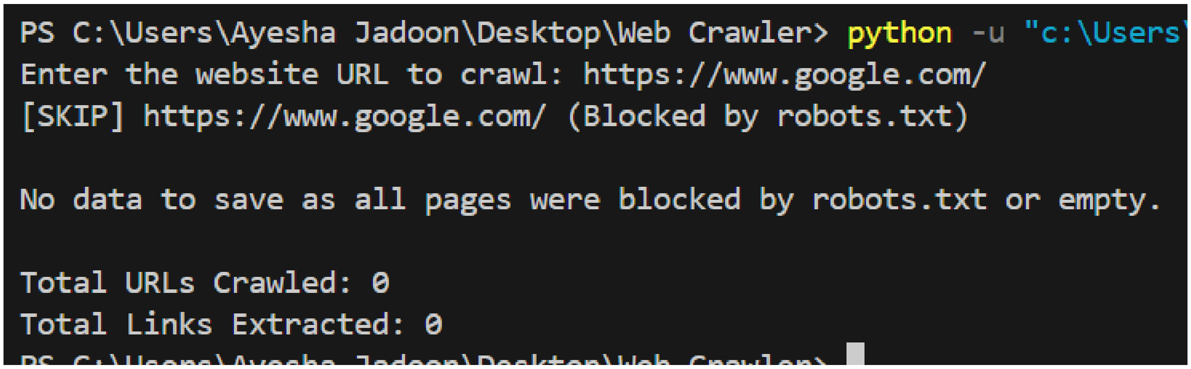

main.py
quotes.toscrape.com_crawled_data.json
quotes.toscrape.com_crawled_data.json > ...

```
[
    {
        "url": "http://quotes.toscrape.com/",
        "title": "Quotes to Scrape",
        "headings": [
            "\nQuotes to Scrape\n",
            "Top Ten tags"
        ],
        "paragraphs": [
            "\nLogin\n",
            "\n                  Quotes by: GoodReads.com\n",
            "\n                  Made with \u2764 by Zyte\n"
        ],
        "images": []
    },
    {
        "url": "http://quotes.toscrape.com/login",
        "title": "Quotes to Scrape",
        "headings": [
            "\nQuotes to Scrape\n"
        ]
```

PROBLEMS
OUTPUT
DEBUG CONSOLE
TERMINAL
PORTS

- **For** https://www.google.com

  the web crawler will attempt to crawl and extract content from Google's homepage. However, because Google has strict rules in place to block web crawlers from scraping their content (due to robots.txt and other security measures), the program will likely encounter a block and skip crawling this page.



```
PS C:\Users\Ayesha Jadoon\Desktop\Web Crawler> python -u "c:\Users
Enter the website URL to crawl: https://www.google.com/
[SKIP] https://www.google.com/ (Blocked by robots.txt)

No data to save as all pages were blocked by robots.txt or empty.

Total URLs Crawled: 0
Total Links Extracted: 0
PS C:\Users\Ayesha Jadoon\Desktop\Web Crawler>
```