Name: Hima Varshini Parasa
GTID: 903945136

CS 6675 HW1

Problem 1. Hand-on Experience with a Web Crawler
Option 1.1: Experience with an Open-Source Crawlers

Table of Contents

1. INTRODUCTION

For this assignment, I implemented a web crawler using Scrapy to extract detailed information about books from books.toscrape.com. The crawler navigates through the entire website, fetching information from all available book pages and generating structured data for further analysis.

- Crawls ~1100 pages with pagination handling.
- Extracts detailed book information for all the books:

  - Title
  - Price
  - Availability
  - Category
  - Description
  - Star Rating

- Saves data in JSON format (web_archive.json) for easy storage and retrieval.
- Created crawl statistics and visualizations.

2. SOURCE CODE

https://github.com/himap2569/WebScraping/blob/main/books/spiders/bookspider.py

3. OUTPUT

a) Truncated output screenshots from Terminal

```
himavarshiniparasa@Himas-MacBook-Air desktop % cd books
himavarshiniparasa@Himas-MacBook-Air books % scrapy startproject books
New Scrapy project 'books', using template directory '/Library/Frameworks/Python.framework/Versions/3.11/lib/python3.11/site-packages/scrapy/templates/project', created in:
    /Users/himavarshiniparasa/Desktop/books/books

You can start your first spider with:
    cd books
    scrapy genspider example example.com
himavarshiniparasa@Himas-MacBook-Air books % cd books
himavarshiniparasa@Himas-MacBook-Air books % scrapy genspider bookspider books.toscrape.com

Created spider 'bookspider' using template 'basic' in module:
  books.spiders.bookspider
```

```
himavarshiniparasa@Himas-MacBook-Air books % scrapy crawl bookspider

2025-01-21 07:31:45 [scrapy.utils.log] INFO: Scrapy 2.12.0 started (bot: books)
2025-01-21 07:31:45 [scrapy.utils.log] INFO: Versions: lxml 5.3.0.0, libxml2 2.12.9, cssselect 1.2.0, parsel 1.10.0, w3lib 2.2.1, Twisted 24.11.0, Python 3.11.9 (v3.11.9:de54cf5be3, Apr  2 2024, 07:12:50) [Cl
ang 13.0.0 (clang-1300.0.29.30)], pyOpenSSL 25.0.0 (OpenSSL 3.4.0 22 Oct 2024), cryptography 44.0.0, Platform macOS-15.1.1-arm64-arm-64bit
2025-01-21 07:31:45 [scrapy.addons] INFO: Enabled addons:
[]
2025-01-21 07:31:45 [asyncio] DEBUG: Using selector: KqueueSelector
2025-01-21 07:31:45 [scrapy.utils.log] DEBUG: Using reactor: twisted.internet.asyncioreactor.AsyncioSelectorReactor
2025-01-21 07:31:45 [scrapy.utils.log] DEBUG: Using asyncio event loop: asyncio.unix_events._UnixSelectorEventLoop
2025-01-21 07:31:45 [scrapy.utils.log] DEBUG: Using reactor: twisted.internet.asyncioreactor.AsyncioSelectorReactor
2025-01-21 07:31:45 [scrapy.utils.log] DEBUG: Using asyncio event loop: asyncio.unix_events._UnixSelectorEventLoop
2025-01-21 07:31:45 [scrapy.extensions.telnet] INFO: Telnet Password: f28cc7483dca39d9
2025-01-21 07:31:45 [scrapy.middleware] INFO: Enabled extensions:
['scrapy.extensions.corestats.CoreStats',
 'scrapy.extensions.telnet.TelnetConsole',
 'scrapy.extensions.memusage.MemoryUsage',
 'scrapy.extensions.closespider.CloseSpider',
 'scrapy.extensions.logstats.LogStats']
2025-01-21 07:31:45 [scrapy.crawler] INFO: Overridden settings:
{'BOT_NAME': 'books',
 'CLOSESPIDER_PAGECOUNT': 1000,
 'DOWNLOAD_DELAY': 1,
 'FEED_EXPORT_ENCODING': 'utf-8',
 'NEWSPIDER_MODULE': 'books.spiders',
 'SPIDER_MODULES': ['books.spiders'],
 'TWISTED_REACTOR': 'twisted.internet.asyncioreactor.AsyncioSelectorReactor',
 'USER_AGENT': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64)'}
2025-01-21 07:31:45 [scrapy.middleware] INFO: Enabled downloader middlewares:
['scrapy.downloadermiddlewares.offsite.OffsiteMiddleware',
 'scrapy.downloadermiddlewares.httpauth.HttpAuthMiddleware',
 'scrapy.downloadermiddlewares.downloadtimeout.DownloadTimeoutMiddleware',
 'scrapy.downloadermiddlewares.defaultheaders.DefaultHeadersMiddleware',
 'scrapy.downloadermiddlewares.useragent.UserAgentMiddleware',
 'scrapy.downloadermiddlewares.retry.RetryMiddleware',
 'scrapy.downloadermiddlewares.redirect.MetaRefreshMiddleware',
 'scrapy.downloadermiddlewares.httpcompression.HttpCompressionMiddleware',
 'scrapy.downloadermiddlewares.redirect.RedirectMiddleware',
 'scrapy.downloadermiddlewares.cookies.CookiesMiddleware',
 'scrapy.downloadermiddlewares.httpproxy.HttpProxyMiddleware',
 'scrapy.downloadermiddlewares.stats.DownloaderStats']
2025-01-21 07:31:45 [scrapy.middleware] INFO: Enabled spider middlewares:
['scrapy.spidermiddlewares.httperror.HttpErrorMiddleware',
 'scrapy.spidermiddlewares.referer.RefererMiddleware',
 'scrapy.spidermiddlewares.urllength.UrlLengthMiddleware',
 'scrapy.spidermiddlewares.depth.DepthMiddleware']
```

```
2025-01-21 07:31:45 [scrapy.middleware] INFO: Enabled item pipelines:
[]
2025-01-21 07:31:45 [scrapy.core.engine] INFO: Spider opened
2025-01-21 07:31:45 [scrapy.extensions.logstats] INFO: Crawled 0 pages (at 0 pages/min), scraped 0 items (at 0 items/min)
2025-01-21 07:31:45 [scrapy.extensions.telnet] INFO: Telnet console listening on 127.0.0.1:6023
2025-01-21 07:31:45 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/> (referer: None)
2025-01-21 07:31:46 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/page-2.html> (referer: http://books.toscrape.com/)
2025-01-21 07:31:47 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/its-only-the-himalayas_981/index.html> (referer: http://books.toscrape.com/)
2025-01-21 07:31:48 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/libertarianism-for-beginners_982/index.html> (referer: http://books.toscrape.com/)
2025-01-21 07:31:49 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/mesaerion-the-best-science-fiction-stories-1800-1849_983/index.html> (referer: http://books.toscrape.com/
)
2025-01-21 07:31:51 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/olio_984/index.html> (referer: http://books.toscrape.com/)
2025-01-21 07:31:52 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/our-band-could-be-your-life-scenes-from-the-american-indie-underground-1981-1991_985/index.html> (referer
: http://books.toscrape.com/)
2025-01-21 07:31:54 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/rip-it-up-and-start-again_986/index.html> (referer: http://books.toscrape.com/)
2025-01-21 07:31:55 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/scott-pilgrims-precious-little-life-scott-pilgrim-1_987/index.html> (referer: http://books.toscrape.com/)
2025-01-21 07:31:56 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/set-me-free_988/index.html> (referer: http://books.toscrape.com/)
2025-01-21 07:31:58 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/shakespeares-sonnets_989/index.html> (referer: http://books.toscrape.com/)
2025-01-21 07:31:59 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/starving-hearts-triangular-trade-trilogy-1_990/index.html> (referer: http://books.toscrape.com/)
2025-01-21 07:32:00 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/the-black-maria_991/index.html> (referer: http://books.toscrape.com/)
2025-01-21 07:32:02 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/the-boys-in-the-boat-nine-americans-and-their-epic-quest-for-gold-at-the-1936-berlin-olympics_992/index.h
tml> (referer: http://books.toscrape.com/)
2025-01-21 07:32:03 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/the-coming-woman-a-novel-based-on-the-life-of-the-infamous-feminist-victoria-woodhull_993/index.html> (re
ferer: http://books.toscrape.com/)
2025-01-21 07:32:04 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/the-dirty-little-secrets-of-getting-your-dream-job_994/index.html> (referer: http://books.toscrape.com/)
2025-01-21 07:32:05 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/the-requiem-red_995/index.html> (referer: http://books.toscrape.com/)
2025-01-21 07:32:06 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/sapiens-a-brief-history-of-humankind_996/index.html> (referer: http://books.toscrape.com/)
2025-01-21 07:32:07 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/page-3.html> (referer: http://books.toscrape.com/catalogue/page-2.html)
2025-01-21 07:32:09 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/you-cant-bury-them-all-poems_961/index.html> (referer: http://books.toscrape.com/catalogue/page-2.html)
2025-01-21 07:32:10 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/behind-closed-doors_962/index.html> (referer: http://books.toscrape.com/catalogue/page-2.html)
2025-01-21 07:32:11 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/in-a-dark-dark-wood_963/index.html> (referer: http://books.toscrape.com/catalogue/page-2.html)
2025-01-21 07:32:12 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/maude-1883-1993she-grew-up-with-the-country_964/index.html> (referer: http://books.toscrape.com/catalogue
/page-2.html)
2025-01-21 07:32:14 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/penny-maybe_965/index.html> (referer: http://books.toscrape.com/catalogue/page-2.html)
2025-01-21 07:32:15 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/sophies-world_966/index.html> (referer: http://books.toscrape.com/catalogue/page-2.html)
2025-01-21 07:32:16 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/the-bear-and-the-piano_967/index.html> (referer: http://books.toscrape.com/catalogue/page-2.html)
2025-01-21 07:32:18 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/the-elephant-tree_968/index.html> (referer: http://books.toscrape.com/catalogue/page-2.html)
2025-01-21 07:32:18 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/the-five-love-languages-how-to-express-heartfelt-commitment-to-your-mate_969/index.html> (referer: http:/
/books.toscrape.com/catalogue/page-2.html)
2025-01-21 07:32:20 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/the-four-agreements-a-practical-guide-to-personal-freedom_970/index.html> (referer: http://books.toscrape
.com/catalogue/page-2.html)
2025-01-21 07:32:21 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/wall-and-piece_971/index.html> (referer: http://books.toscrape.com/catalogue/page-2.html)
2025-01-21 07:32:22 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/worlds-elsewhere-journeys-around-shakespeares-globe_972/index.html> (referer: http://books.toscrape.com/c
atalogue/page-2.html)
2025-01-21 07:32:24 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/aladdin-and-his-wonderful-lamp_973/index.html> (referer: http://books.toscrape.com/catalogue/page-2.html)
2025-01-21 07:32:25 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/americas-cradle-of-quarterbacks-western-pennsylvanias-football-factory-from-johnny-unitas-to-joe-montana_
974/index.html> (referer: http://books.toscrape.com/catalogue/page-2.html)
2025-01-21 07:32:26 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/birdsong-a-story-in-pictures_975/index.html> (referer: http://books.toscrape.com/catalogue/page-2.html)
2025-01-21 07:32:27 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/black-dust_976/index.html> (referer: http://books.toscrape.com/catalogue/page-2.html)
2025-01-21 07:32:29 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/page-4.html> (referer: http://books.toscrape.com/catalogue/page-3.html)
```

```
2025-01-21 07:32:30 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/the-natural-history-of-us-the-fine-art-of-pretending-2_941/index.html> (referer: http://books.toscrape.com/catalogue/page-3.html)
2025-01-21 07:32:30 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/the-past-never-ends_942/index.html> (referer: http://books.toscrape.com/catalogue/page-3.html)
2025-01-21 07:32:32 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/the-pioneer-woman-cooks-dinnertime-comfort-classics-freezer-food-16-minute-meals-and-other-delicious-ways-to-solve-supper_943/index.html> (referer: http://books.toscrape.com/catalogue/page-3.html)
2025-01-21 07:32:33 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/the-secret-of-dreadwillow-carse_944/index.html> (referer: http://books.toscrape.com/catalogue/page-3.html)
2025-01-21 07:32:34 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/the-torch-is-passed-a-harding-family-story_945/index.html> (referer: http://books.toscrape.com/catalogue/page-3.html)
2025-01-21 07:32:35 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/thirst_946/index.html> (referer: http://books.toscrape.com/catalogue/page-3.html)
2025-01-21 07:32:37 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/this-one-summer_947/index.html> (referer: http://books.toscrape.com/catalogue/page-3.html)
2025-01-21 07:32:38 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/throwing-rocks-at-the-google-bus-how-growth-became-the-enemy-of-prosperity_948/index.html> (referer: http://books.toscrape.com/catalogue/page-3.html)
2025-01-21 07:32:39 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/tsubasa-world-chronicle-2-tsubasa-world-chronicle-2_949/index.html> (referer: http://books.toscrape.com/catalogue/page-3.html)
2025-01-21 07:32:40 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/unbound-how-eight-technologies-made-us-human-transformed-society-and-brought-our-world-to-the-brink_950/index.html> (referer: http://books.toscrape.com/catalogue/page-3.html)
2025-01-21 07:32:42 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/unicorn-tracks_951/index.html> (referer: http://books.toscrape.com/catalogue/page-3.html)
2025-01-21 07:32:43 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/unseen-city-the-majesty-of-pigeons-the-discreet-charm-of-snails-other-wonders-of-the-urban-wilderness_952/index.html> (referer: http://books.toscrape.com/catalogue/page-3.html)
2025-01-21 07:32:44 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/untitled-collection-sabbath-poems-2014_953/index.html> (referer: http://books.toscrape.com/catalogue/page-3.html)
2025-01-21 07:32:45 [scrapy.extensions.logstats] INFO: Crawled 49 pages (at 49 pages/min), scraped 0 items (at 0 items/min)
2025-01-21 07:32:45 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/we-love-you-charlie-freeman_954/index.html> (referer: http://books.toscrape.com/catalogue/page-3.html)
2025-01-21 07:32:47 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/when-we-collided_955/index.html> (referer: http://books.toscrape.com/catalogue/page-3.html)
2025-01-21 07:32:48 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/without-borders-wanderlove-1_956/index.html> (referer: http://books.toscrape.com/catalogue/page-3.html)
2025-01-21 07:32:49 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/page-5.html> (referer: http://books.toscrape.com/catalogue/page-4.html)
2025-01-21 07:32:50 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/rat-queens-vol-3-demons-rat-queens-collected-editions-11-15_921/index.html> (referer: http://books.toscrape.com/catalogue/page-4.html)
2025-01-21 07:32:51 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/reskilling-america-learning-to-labor-in-the-twenty-first-century_922/index.html> (referer: http://books.toscrape.com/catalogue/page-4.html)
2025-01-21 07:32:53 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/saga-volume-5-saga-collected-editions-5_923/index.html> (referer: http://books.toscrape.com/catalogue/page-4.html)
2025-01-21 07:32:54 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/saga-volume-6-saga-collected-editions-6_924/index.html> (referer: http://books.toscrape.com/catalogue/page-4.html)
2025-01-21 07:32:55 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/security_925/index.html> (referer: http://books.toscrape.com/catalogue/page-4.html)
2025-01-21 07:32:57 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/soul-reader_926/index.html> (referer: http://books.toscrape.com/catalogue/page-4.html)
2025-01-21 07:32:58 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/spark-joy-an-illustrated-master-class-on-the-art-of-organizing-and-tidying-up_927/index.html> (referer: http://books.toscrape.com/catalogue/page-4.html)
2025-01-21 07:33:00 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/the-activists-tao-te-ching-ancient-advice-for-a-modern-revolution_928/index.html> (referer: http://books.toscrape.com/catalogue/page-4.html)
2025-01-21 07:33:01 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/the-age-of-genius-the-seventeenth-century-and-the-birth-of-the-modern-mind_929/index.html> (referer: http://books.toscrape.com/catalogue/page-4.html)
2025-01-21 07:33:02 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/the-art-forger_930/index.html> (referer: http://books.toscrape.com/catalogue/page-4.html)
2025-01-21 07:33:03 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/the-bulletproof-diet-lose-up-to-a-pound-a-day-reclaim-energy-and-focus-upgrade-your-life_931/index.html> (referer: http://books.toscrape.com/catalogue/page-4.html)
2025-01-21 07:33:05 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/the-death-of-humanity-and-the-case-for-life_932/index.html> (referer: http://books.toscrape.com/catalogue/page-4.html)
2025-01-21 07:33:06 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/the-electric-pencil-drawings-from-inside-state-hospital-no-3_933/index.html> (referer: http://books.toscrape.com/catalogue/page-4.html)
2025-01-21 07:33:07 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/the-gutsy-girl-escapades-for-your-life-of-epic-adventure_934/index.html> (referer: http://books.toscrape.com/catalogue/page-4.html)
2025-01-21 07:33:08 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/the-inefficiency-assassin-time-management-tactics-for-working-smarter-not-longer_935/index.html> (referer: http://books.toscrape.com/catalogue/page-4.html)
2025-01-21 07:33:08 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/the-life-changing-magic-of-tidying-up-the-japanese-art-of-decluttering-and-organizing_936/index.html> (referer: http://books.toscrape.com/catalogue/page-4.html)
```

```
2025-01-21 07:52:02 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/walt-disneys-alice-in-wonderland_777/index.html> (referer: http://books.toscrape.com/catalogue/page-12.html)
2025-01-21 07:52:04 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/we-are-robin-vol-1-the-vigilante-business-we-are-robin-1_778/index.html> (referer: http://books.toscrape.com/catalogue/page-12.html)
2025-01-21 07:52:05 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/whats-it-like-in-space-stories-from-astronauts-whove-been-there_779/index.html> (referer: http://books.toscrape.com/catalogue/page-12.html)
2025-01-21 07:52:05 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/whole-lotta-creativity-going-on-60-fun-and-unusual-exercises-to-awaken-and-strengthen-your-creativity_780/index.html> (referer: http://books.toscrape.com/catalogue/page-12.html)
2025-01-21 07:52:06 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/chasing-heaven-what-dying-taught-me-about-living_797/index.html> (referer: http://books.toscrape.com/catalogue/page-11.html)
2025-01-21 07:52:07 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/close-to-you_798/index.html> (referer: http://books.toscrape.com/catalogue/page-11.html)
2025-01-21 07:52:09 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/daring-greatly-how-the-courage-to-be-vulnerable-transforms-the-way-we-live-love-parent-and-lead_799/index.html> (referer: http://books.toscrape.com/catalogue/page-11.html)
2025-01-21 07:52:10 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/dark-notes_800/index.html> (referer: http://books.toscrape.com/catalogue/page-11.html)
2025-01-21 07:52:11 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/little-red_817/index.html> (referer: http://books.toscrape.com/catalogue/page-10.html)
2025-01-21 07:52:12 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/louisa-the-extraordinary-life-of-mrs-adams_818/index.html> (referer: http://books.toscrape.com/catalogue/page-10.html)
2025-01-21 07:52:14 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/miss-peregrines-home-for-peculiar-children-miss-peregrines-peculiar-children-1_819/index.html> (referer: http://books.toscrape.com/catalogue/page-10.html)
2025-01-21 07:52:15 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/modern-romance_820/index.html> (referer: http://books.toscrape.com/catalogue/page-10.html)
2025-01-21 07:52:16 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://books.toscrape.com/catalogue/the-argonauts_837/index.html> (referer: http://books.toscrape.com/catalogue/page-9.html)
2025-01-21 07:52:16 [bookspider] DEBUG: Web archive saved to web_archive.json
2025-01-21 07:52:16 [scrapy.statscollectors] INFO: Dumping Scrapy stats:
{'downloader/request_bytes': 352361,
 'downloader/request_count': 1015,
 'downloader/request_method_count/GET': 1015,
 'downloader/response_bytes': 3925713,
 'downloader/response_count': 1015,
 'downloader/response_status_count/200': 1015,
 'elapsed_time_seconds': 1230.948533,
 'finish_reason': 'closespider_pagecount',
 'finish_time': datetime.datetime(2025, 1, 21, 12, 52, 16, 527151, tzinfo=datetime.timezone.utc),
 'httpcompression/response_bytes': 21265873,
 'httpcompression/response_count': 1015,
 'items_per_minute': None,
 'log_count/DEBUG': 1021,
 'log_count/INFO': 30,
 'memusage/max': 77725696,
 'memusage/startup': 66830336,
 'request_depth_max': 50,
 'response_received_count': 1015,
 'responses_per_minute': None,
 'scheduler/dequeued': 1015,
 'scheduler/dequeued/memory': 1015,
 'scheduler/enqueued': 1050,
 'scheduler/enqueued/memory': 1050,
 'start_time': datetime.datetime(2025, 1, 21, 12, 31, 45, 578618, tzinfo=datetime.timezone.utc)}
2025-01-21 07:52:16 [scrapy.core.engine] INFO: Spider closed (closespider_pagecount)
```

Full output is shared in the zip file and could also be found here – https://github.com/himap2569/Web-Crawling/blob/main/output.txt

b) web_archive.json : JSON of data extracted (truncated)


```
[
    {
        "title": "It's Only the Himalayas",
        "price": "\u00a345.17",
        "availability": "19 available",
        "category": "Travel",
        "description": "\u201cWherever you go, whatever you do, just . . . don\u2019t do anything stupid.\u201d \u2014My MotherDuring her yearlong adventure backpacking from South Africa to Singapore, S. Bedford definitely did a few things her mother might classify as \"stupid.\" She swam with great white sharks in South Africa, ran from lions in Zimbabwe, climbed a Himalayan mountain without training in Nepal, and wa \u201cWherever you go, whatever you do, just . . . don\u2019t do anything stupid.\u201d \u2014My MotherDuring her yearlong adventure backpacking from South Africa to Singapore, S. Bedford definitely did a few things her mother might classify as \"stupid.\" She swam with great white sharks in South Africa, ran from lions in Zimbabwe, climbed a Himalayan mountain without training in Nepal, and watched as her friend was attacked by a monkey in Indonesia.But interspersed in those slightly more crazy moments, Sue Bedford and her friend \"Sara the Stoic\" experienced the sights, sounds, life, and culture of fifteen countries. Joined along the way by a few friends and their aging fathers here and there, Sue and Sara experience the trip of a lifetime. They fall in love with the world, cultivate an appreciation for home, and discover who, or what, they want to become.It's Only the Himalayas is the incredibly funny, sometimes outlandish, always entertaining confession of a young backpacker that will inspire you to take your own adventure. ...more",
        "rating": "Two",
        "url": "http://books.toscrape.com/catalogue/its-only-the-himalayas_981/index.html"
    },
    {
        "title": "Libertarianism for Beginners",
        "price": "\u00a351.33",
        "availability": "19 available",
        "category": "Politics",
        "description": "Libertarianism isn't about winning elections; it is first and foremost a political philosophy--a description of how, in the opinion of libertarians, free people ought to treat one another, at least when they use the law, which they regard as potentially dangerous. If libertarians are correct, the law should intrude into people's lives as little as possible, rarely telling Libertarianism isn't about winning elections; it is first and foremost a political philosophy--a description of how, in the opinion of libertarians, free people ought to treat one another, at least when they use the law, which they regard as potentially dangerous. If libertarians are correct, the law should intrude into people's lives as little as possible, rarely telling them what to do or how to live.A political and economic philosophy as old as John Locke and John Stuart Mill, but as alive and timely as Rand Paul, the Tea Party, and the novels of Ayn Rand, libertarianism emphasizes individual rights and calls for a radical reduction in the power and size of government. \"Libertarianism For Beginners\" lays out the history and principles of this often-misunderstood philosophy in lucid, dispassionate terms that help illuminate today's political dialogue.\" ...more",
        "rating": "Two",
        "url": "http://books.toscrape.com/catalogue/libertarianism-for-beginners_982/index.html"
    },
    {
        "title": "Mesaerion: The Best Science Fiction Stories 1800-1849",
        "price": "\u00a337.59",
        "availability": "19 available",
        "category": "Science Fiction",
        "description": "Andrew Barger, award-winning author and engineer, has extensively researched forgotten journals and magazines of the early 19th century to locate groundbreaking science fiction short stories in the English language. In doing so, he found what is possibly the first science fiction story by a female (and it is not from Mary Shelley). Andrew located the first steampunk short Andrew Barger, award-winning author and engineer, has extensively researched forgotten journals and magazines of the early 19th century to locate groundbreaking science fiction short stories in the English language. In doing so, he found what is possibly the first science fiction story by a female (and it is not from Mary Shelley). Andrew located the first steampunk short story, which has not been republished since 1844. There is the first voyage to the moon in a balloon, republished for the first time since 1820 that further tells of a darkness machine and a lunarian named Zuloc. Other sci-stories include the first robotic insect and an electricity gun. Once again, Andrew has searched old texts to find the very best science fiction stories from the period when the genre automated to life, some of the stories are published for the first time in nearly 200 years. Read these fantastic stories today!OUR OWN COUNTRY So mechanical has the age become, that men seriously talk of flying machines, to go by steam, --not your air-balloons, but real Daedalian wings, made of wood and joints, nailed to your shoulder, -- not wings of feathers and wax like the wings of Icarus, who fell into the Cretan sea, but real, solid, substantial, rock-maple wings with wrought-iron hinges, and huge concavities, to propel us through the air. Knickerbocker Magazine, May 18 ...more",
        "rating": "One",
        "url": "http://books.toscrape.com/catalogue/mesaerion-the-best-science-fiction-stories-1800-1849_983/index.html"
    },
    {
```


Full output is shared in the zip file and could also be found here – https://github.com/himap2569/Web-Crawling/blob/main/books/books/web_archive.json

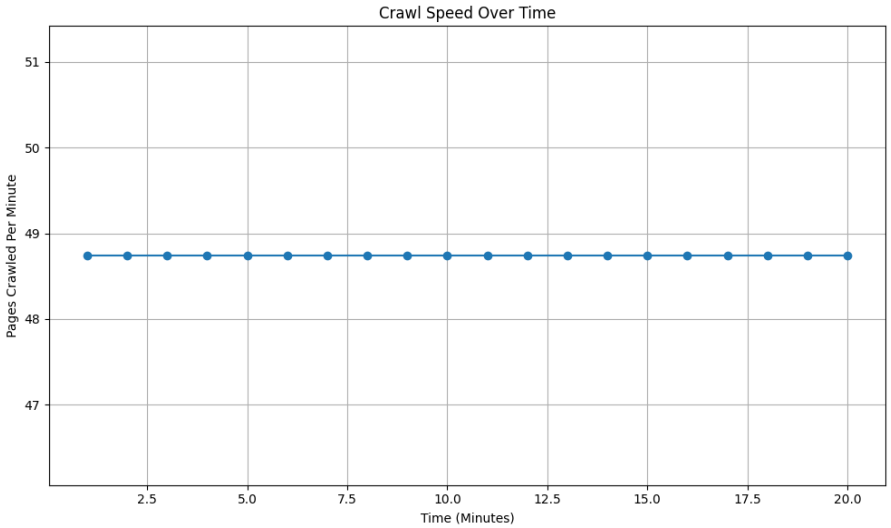
c) Keywords (truncated)

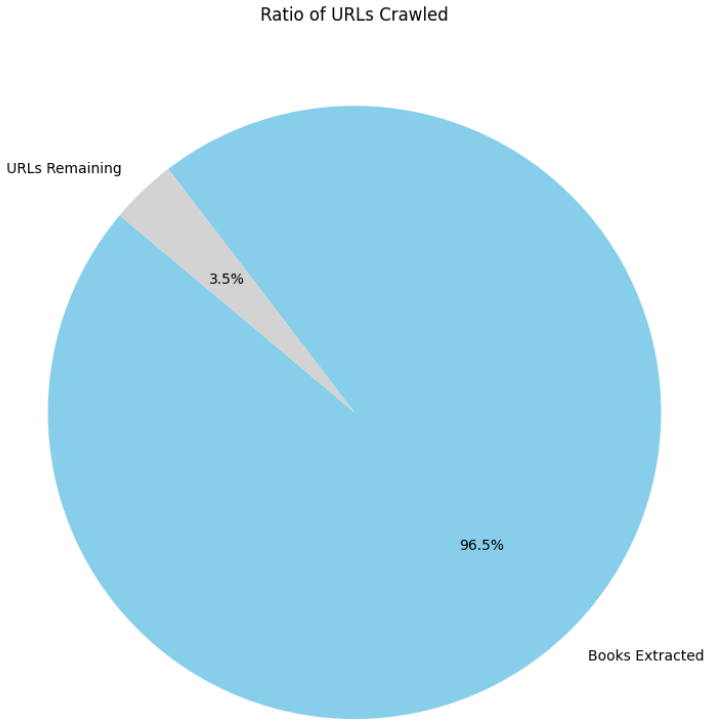

```
{
    "keywords": {
        "a": 165,
        "light": 3,
        "in": 69,
        "the": 632,
        "attic": 1,
        "tipping": 3,
        "velvet": 1,
        "soumission": 1,
        "sharp": 1,
        "objects": 1,
        "sapiens:": 1,
        "brief": 3,
        "history": 16,
        "of": 256,
        "humankind": 1,
        "requiem": 1,
        "red": 8,
        "dirty": 2,
        "little": 14,
        "secrets": 8,
        "getting": 1,
        "your": 26,
        "dream": 8,
        "job": 2,
        "coming": 1,
        "woman:": 3,
        "novel": 6,
        "based": 2,
        "on": 34,
        "life": 35,
        "infamous": 1,
        "feminist": 1,
        "victoria": 1,
        "woodhull": 1,
        "boys": 3,
        "boat:": 1,
        "nine": 1,
        "americans": 1,
        "and": 185,
        "their": 2,
        "epic": 3,
        "quest": 6,
        "for": 58,
        "gold": 2,
        "at": 11,
        "1936": 1,
        "berlin": 2,
```

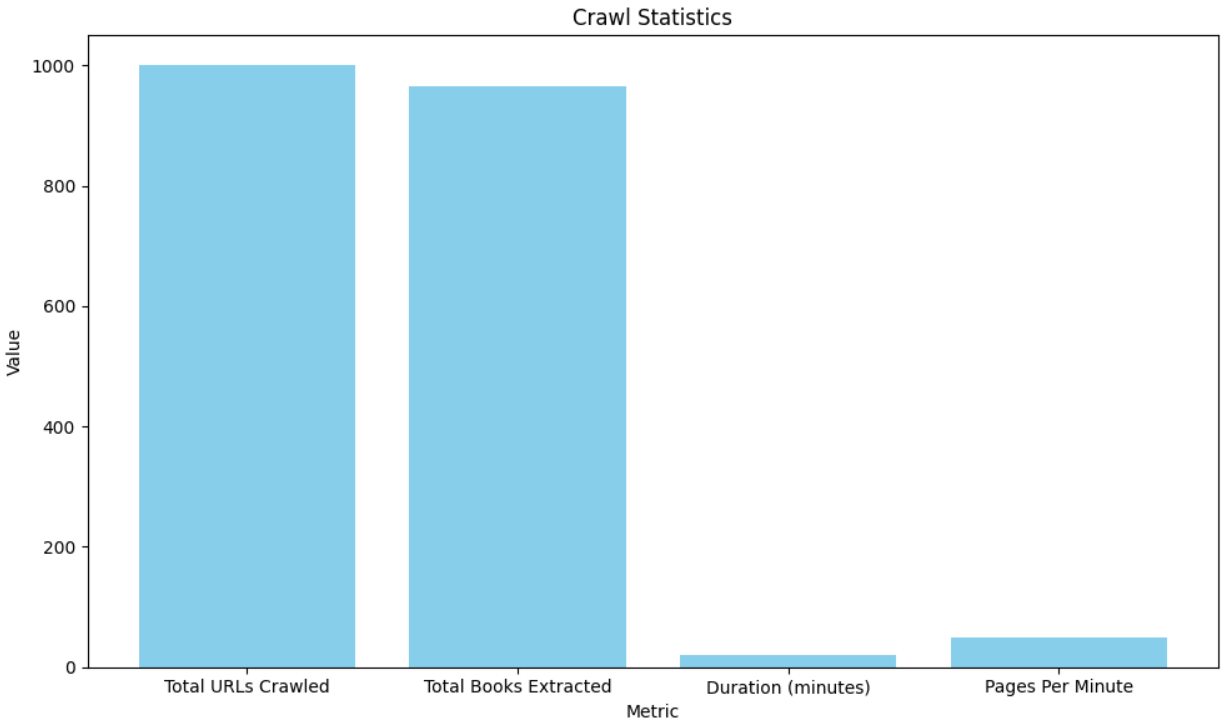

Full output is shared in the zip file and could also be found here – https://github.com/himap2569/Web-Crawling/blob/main/books/books/crawled_data.json

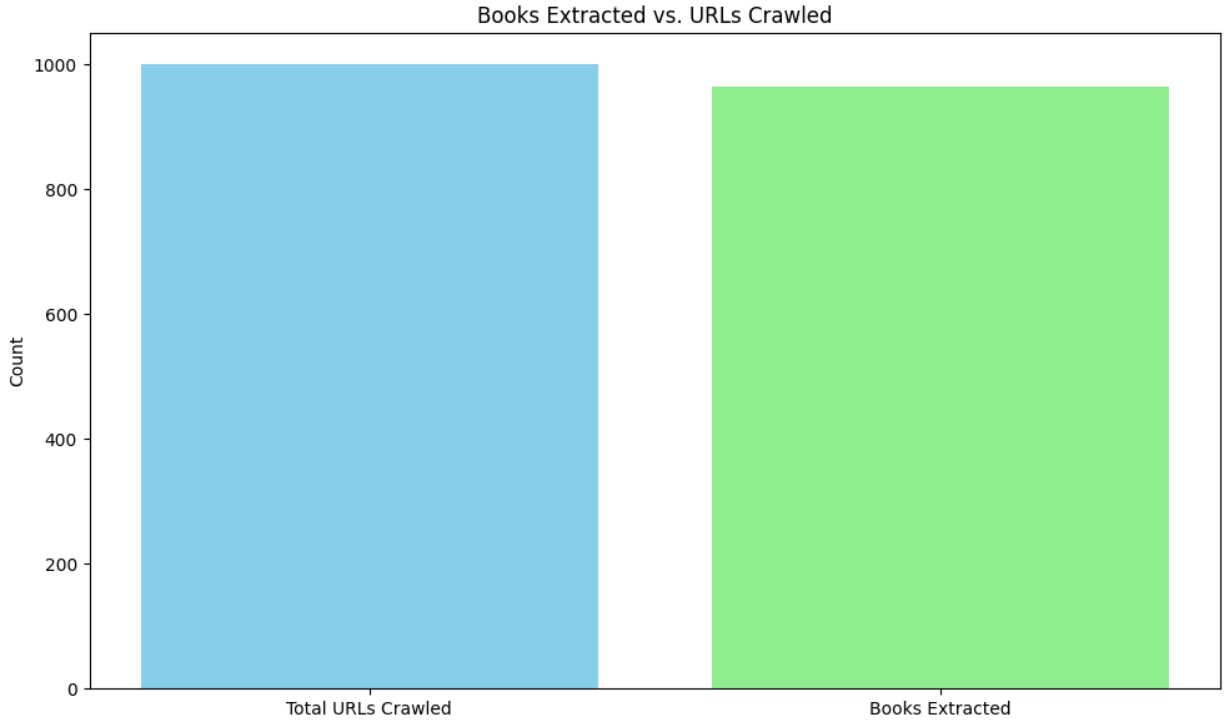
4. STATISTICS

Metric	Value
Total URLs Crawled	1100
Total Books Extracted	965
Duration (minutes)	20.52
Pages Per Minute	48.74






Crawl Statistics
Value
1000
800
600
400
200
0
Total URLs Crawled
Total Books Extracted
Duration (minutes)
Pages Per Minute
Metric


Books Extracted vs. URLs Crawled
Count
1000
800
600
400
200
0
Total URLs Crawled
Books Extracted

5. LESSONS LEARNED

a) Observations:

1. Efficiency of the Crawler:

- The crawler successfully handled 1100 URLs and extracted book details from 965 pages in 20.52 minutes, achieving a 96.5% extraction success rate.
- A steady crawl speed of approximately 48.74 pages per minute demonstrates the efficiency of the Scrapy framework for small-scale crawling tasks.

2. Challenges Encountered:

- Handling Non-Book Pages: Some URLs did not lead to book pages, resulting in incomplete data extraction.
- Server Politeness: A DOWNLOAD_DELAY of 1 second limited the crawl speed but ensured adherence to server rules.

3. Observations on Scaling:

The performance was consistent, but crawling at scale (e.g., millions of pages) would require significant optimizations to avoid excessive runtime.

b) Notes:

1. *Crawling Efficiency:*
   The crawler's steady speed and high accuracy rate validate the use of Scrapy for data extraction tasks. However, handling edge cases (e.g., non-book pages) can further improve efficiency.
2. *Server Politeness:*
   A DOWNLOAD_DELAY is necessary for ethical crawling. However, for large-scale crawling, strategies like distributed crawling or request batching can balance speed and politeness.
3. *Scalability Challenges:*
   Crawling millions or billions of pages would be infeasible with a single crawler due to time constraints. Distributed crawling across multiple machines or nodes is essential for scalability.
4. *Data Storage:*

Storing data in a JSON file worked well for a small dataset, but for millions or billions of pages, a database (e.g., MongoDB or PostgreSQL) would be more efficient for storage and querying.

c) Time Predictions for Scaling

Assumptions:

- Crawl Speed: 48.74 pages/minute.
- Efficiency: The current setup (1-second delay) is used for scaling.

For 10 Million Pages:

1. *Time Required:*

   At 48.74 pages/minute:

   Time (minutes) = 10,000,000 ÷ 48.74 ≈ 205,169.92 minutes

   Convert to days: Time (days) = 205,169.92 ÷ 1440 ≈ 142.5 days (continuous operation)

2. *Optimization Suggestions:*

   By reducing the delay (e.g., to 0.5 seconds) or using distributed crawlers, this time could be halved or further reduced.

For 1 Billion Pages:

1. *Time Required:*

   At 48.74 pages/minute: Time (minutes) = 1,000,000,000 ÷ 48.74 ≈ 20,516,992.06 minutes

   Convert to years: Time (years) = 20,516,992.06 ÷ 525,600 ≈ 39 years (continuous operation)

2. *Scaling Strategies:*

   - Distributed Crawling: Split the workload across 10 crawlers, reducing time to approximately 3.9 years.

   - Further optimizations, like reducing delays or increasing concurrency, could bring this down significantly.

6. REFERENCES

[1] https://www.youtube.com/watch?v=m_3gjHGxIJc

[2] https://www.youtube.com/watch?v=w3XcMfyUGxY

[3] https://medium.com/@joerosborne/intro-to-web-scraping-build-your-first-scraper-in-5-minutes-1c36b5c4b110

[4] https://www.freecodecamp.org/news/web-scraping-python-tutorial-how-to-scrape-data-from-a-website/