Crawl_Search

December 1, 2025

0.1 Abstract

This project focuses on designing and implementing a web-based search pipeline, including document acquisition, indexing, and query retrieval. A Scrapy-powered crawler is used to download and store web documents in HTML format. An inverted index is then generated using TF-IDF weighting to support efficient similarity-based retrieval. Finally, a Flask-driven query processor enables users to submit free-text searches and receive ranked document results based on cosine similarity.

Future enhancements can focus on improving scalability and retrieval quality. Optional features such as concurrent and distributed crawling using AutoThrottle and Scrapyd would allow faster and broader document collection. Search accuracy could be increased by integrating semantic vector embeddings like Word2Vec or FAISS-based k-nearest neighbor similarity. Additionally, a front-end search interface and production deployment would make the system more user-accessible and robust.

0.2 Overview

The project focus on web document retrieval and query processing. It consists of three compononent: a web crawler, an indexer, and a query procesor. These components works together to enable efficiant crawling, indexing, and querying of web docuemnts. ### Soultion Outline 1. **Web Crawler**: Use Scrapy to roam the web and download web documents. 2. **Indexer**: Build an inverted index using the TF-IDF and Cosine similarities to search and retrieve 3. **Query Processor**: Accepts user queries and returns ranked document results using TF-IDF and cosine similarity. ### Relevant Literature: - Scrapy Documentation:
This official tutorial provides a comprehensive introduction to Scrapy, a powerful Python framework for web crawling and scraping. It covers spider creation, data extraction, and best practices for scalable web data collection, which directly informs the design of the project's web crawler component.

- Flask Documentation:
  The Flask documentation offers detailed guidance on building lightweight web applications and APIs in Python. It is the primary reference for implementing the query processor and REST API endpoints, enabling user interaction and search result delivery in this project.

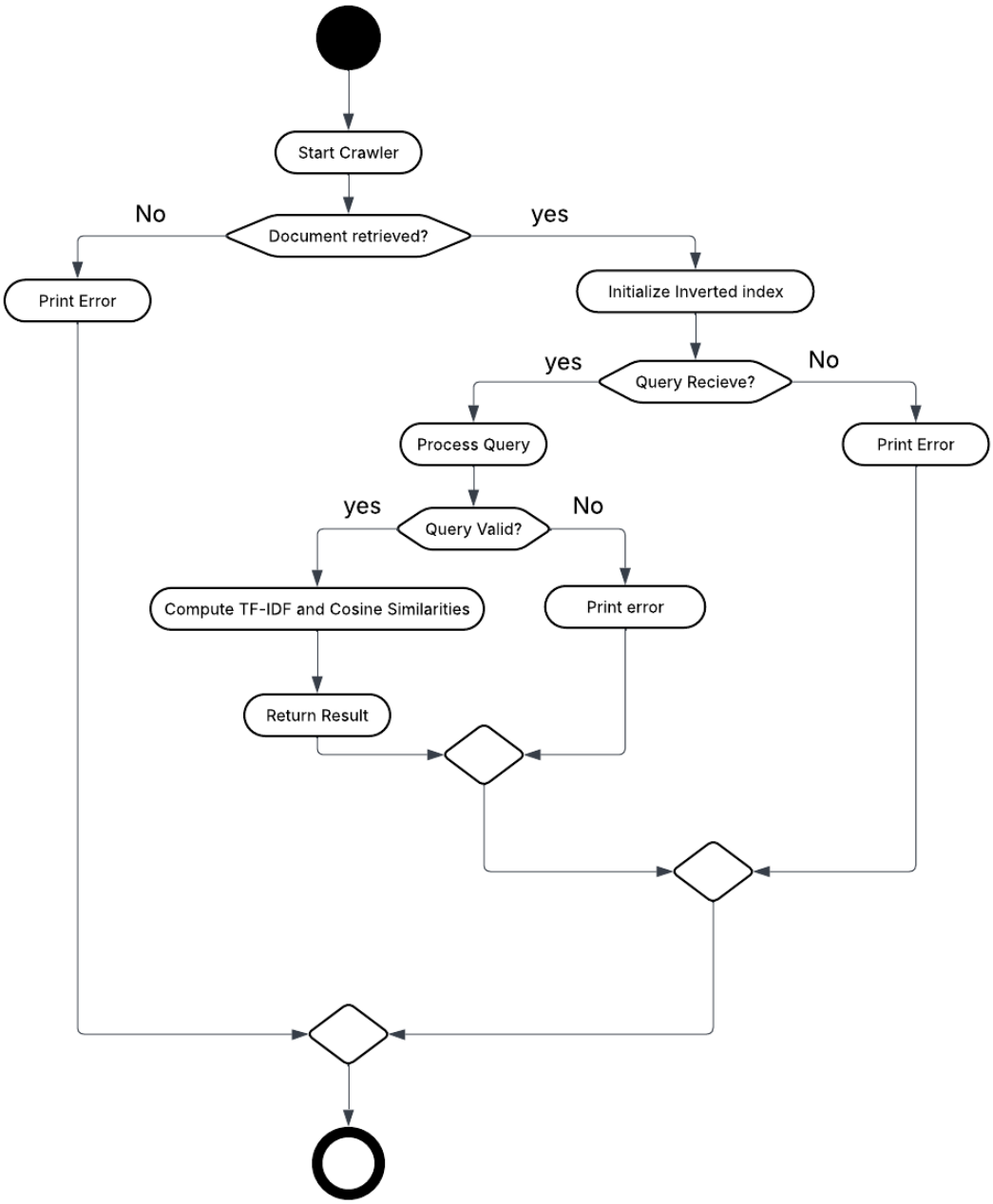
For the search indexing section, I have already studied and applied established techniques, as demonstrated in this project.

0.2.1 Proposed System

The system use the power of scrapy for web crawling, Skit-Learn for TF-IDF indexing and Cosine similarities, and Flask for query processing. The combination of these technologies, it delivers a solution for web docuement retrieval and query processing.

0.3 Design

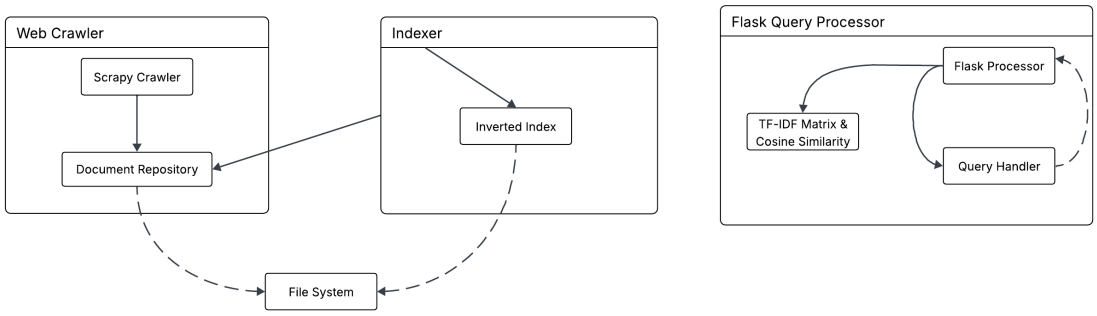
The system's capabilities includes downloaading web documents, constucting an inverted index, and processing user queries. it involves the Scrapy crawler to collect web documents, th indexer building the TF-IDF matrix, and the Flask processor handeling user queries. Integration is achieved through data exchange between components and adherence to defined interfaces.



This activity diagram outlines the flow of activities within the system, including crawling, indexing, processing queries, and returning results. Each component interacts with the others to achieve the overall functionality of the system.

0.4 Architecture

The architecture of the system involves three main components: the Scrapy crawler, the inverted indexer, and the Flask query processor. These components interact with each other to enable web document crawling, indexing, and query processing.



This diagram illustrates the interaction between the components and how they utilize interfaces such as file I/O and HTTP endpoints for communication and data exchange. The implementation relies on external libraries such as Scikit-Learn, BeutifulSoup, Spider, and Flask to enable various functionality.

0.5 5. Operation

Installation Instructions:
- Ensure Python 3.8 or higher is installed.
- Create and activate a virtual environment:
- `python -m venv venv`
- `venv\Scripts\activate` (Windows) or `source venv/bin/activate` (Linux/Mac)
- Install all required dependencies using:
- `pip install -r requirements.txt`

Software Commands:
- **Running the crawler:** Execute the relevant notebook cells or script to start the Scrapy crawler.
- **Building the index:** Run the indexing cells or script to generate the inverted index (`index.json`).
- **Running the Flask server and querying:** Start the Flask server by running the appropriate notebook cells or script. Submit queries via the web interface or batch CSV upload.

Inputs:
- **Seed URL:** The initial URL for the crawler.
- **CSV queries:** A CSV file containing `query_id` and `query_text` columns for batch search.
- **Configuration options:** Parameters such as maximum pages to crawl, crawl depth, and allowed domains.

Outputs:
- **Crawled HTML files:** Stored in the `pages/` directory.
- **Cleaned text files:** Stored in the `cleaned_text/` directory.
- **Inverted index:** Saved as `index.json`.
- **Search results:** Ranked results provided as a downloadable CSV file via the Flask API. ## 6. Conclusion - **Success/Failure:**
The project successfully demonstrates the implementation of a modular web search pipeline, encompassing web crawling, document cleaning, inverted indexing, and query processing. The system is capable of acquiring web documents, extracting and cleaning their content, constructing a positional inverted index, and retrieving relevant documents in response to user queries using TF-IDF

and cosine similarity. Error handling is incorporated to manage missing files or invalid queries, ensuring robust operation.

- **Outputs:**
  The outputs of the system are as follows:
  - Crawled HTML files are stored in the `pages/` directory.
  - Cleaned text files are saved in the `cleaned_text/` directory.
  - The positional inverted index is generated and saved as `index.json`.
  - Search results for user queries are provided as ranked CSV files, which can be downloaded via the Flask API.
  - Example outputs, including ranked document results and similarity scores for sample queries, are displayed below as evidence of successful retrieval and ranking.

0.6 7. Data Sources

- Web documents are sourced from quotes.toscrape.com
- Additional data sources can be integrated as needed for testing and evaluation.

0.7 8. Test Cases

Test cases involve validating the functionality of the crawler, indexer, and query processor. Frameworks such as Scrapy's testing tools and unit testing libraries for Python can be utilized. Test coverage includes scenarios for crawling, indexing, and querying. ## 9. Source Code

The complete source code for this project, including the web crawler, indexer, and query processor, is provided in the accompanying Jupyter notebook (`Crawl_Search.ipynb`) and related Python scripts within the project directory. All implementation details, code cells, and configuration files are included to enable full reproducibility of the results.

Key files and directories: - `Crawl_Search.ipynb`: Main notebook containing the code for crawling, indexing, and query processing. - `pages/`: Directory where crawled HTML files are stored. - `cleaned_text/`: Directory containing cleaned text files extracted from HTML pages. - `index.json`: Generated inverted index used for search and retrieval. - `requirements.txt`: List of required Python packages for the project.

To access or modify the source code, open the notebook or scripts in your preferred Python environment (e.g., Jupyter Notebook or Visual Studio Code).

0.8 10. Bibliography

- https://docs.scrapy.org/en/latest/intro/tutorial.html
  Official Scrapy tutorial providing step-by-step guidance on building web crawlers, extracting data, and managing large-scale scraping projects.
- https://beautiful-soup-4.readthedocs.io/en/latest
  Comprehensive documentation for Beautiful Soup, a Python library used for parsing HTML and XML documents and extracting useful information.

- https://www.w3schools.com/python/python_json.asp
  Introductory guide to working with JSON data in Python, including reading, writing, and parsing JSON files.
- https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html
  Reference documentation for the cosine similarity function in scikit-learn, used to measure similarity between vectors, such as document and query vectors.
- https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html
  Documentation for the TfidfVectorizer class in scikit-learn, which converts text documents to a matrix of TF-IDF features for information retrieval and text mining.

[8]:
```
import nest_asyncio
nest_asyncio.apply()

import scrapy
from scrapy.crawler import CrawlerProcess
from scrapy.linkextractors import LinkExtractor
import os
import uuid
```

0.8.1 Step 1: Web Crawling

[ ]:
```
os.makedirs("pages", exist_ok=True)

class NotebookCrawler(scrapy.Spider):
    name = "notebook_crawler"

    def __init__(self, seed_url, allowed_domain, max_pages, max_depth, *args,␣
 ↪**kwargs):
        super(NotebookCrawler, self).__init__(*args, **kwargs)
        self.start_urls = [seed_url]
        self.allowed_domains = [allowed_domain]
        self.max_pages = max_pages
        self.visited = set()

        # Set max depth in custom settings
        self.custom_settings = {
            'DEPTH_LIMIT': max_depth,
            'AUTOTHROTTLE_ENABLED': True,
            'LOG_ENABLED': True,
            'CLOSESPIDER_PAGECOUNT': max_pages
        }

    def parse(self, response):
        # Stop if max pages reached
        if len(self.visited) >= self.max_pages:
            self.logger.info(f"Reached max pages limit: {self.max_pages}")
```

```
            return

        # Generate UUID as the complete filename
        page_uuid = str(uuid.uuid4())
        filename = f"{page_uuid}.html"

        # Save HTML content
        with open(f"pages/{filename}", "w", encoding="utf-8") as f:
            f.write(response.text)

        self.visited.add(response.url)
        self.logger.info(f"Saved page {len(self.visited)}/{self.max_pages}:␣
 ↪{filename}")

        # Extract and follow links
        if len(self.visited) < self.max_pages:
            links = LinkExtractor(allow_domains=self.allowed_domains).
 ↪extract_links(response)
            for link in links:
                if link.url not in self.visited:
                    yield response.follow(link.url, self.parse)

# Usage with configurable parameters
process = CrawlerProcess(settings={
    'LOG_LEVEL': 'INFO',
})

process.crawl(
    NotebookCrawler,
    seed_url='https://quotes.toscrape.com',
    allowed_domain='quotes.toscrape.com',
    max_pages=100,
    max_depth=5
)
try:
    process.start()
except:
    pass
```

```
2025-11-23 17:43:55 [scrapy.utils.log] INFO: Scrapy 2.13.3 started (bot:
scrapybot)
2025-11-23 17:43:55 [scrapy.utils.log] INFO: Versions:
{'lxml': '6.0.2',
 'libxml2': '2.11.9',
 'cssselect': '1.3.0',
 'parsel': '1.10.0',
 'w3lib': '2.3.1',
```

```
 'Twisted': '25.5.0',
 'Python': '3.14.0 (tags/v3.14.0:ebf955d, Oct  7 2025, 10:15:03) [MSC v.1944 '
           '64 bit (AMD64)]',
 'pyOpenSSL': '25.3.0 (OpenSSL 3.5.4 30 Sep 2025)',
 'cryptography': '46.0.3',
 'Platform': 'Windows-11-10.0.26200-SP0'}
2025-11-23 17:43:55 [scrapy.addons] INFO: Enabled addons:
[]
2025-11-23 17:43:55 [scrapy.extensions.telnet] INFO: Telnet Password:
e50700739db06670
2025-11-23 17:43:55 [scrapy.middleware] INFO: Enabled extensions:
['scrapy.extensions.corestats.CoreStats',
 'scrapy.extensions.telnet.TelnetConsole',
 'scrapy.extensions.logstats.LogStats']
2025-11-23 17:43:55 [scrapy.crawler] INFO: Overridden settings:
{'LOG_LEVEL': 'INFO'}
2025-11-23 17:43:55 [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']
2025-11-23 17:43:55 [scrapy.utils.log] INFO: Versions:
{'lxml': '6.0.2',
 'libxml2': '2.11.9',
 'cssselect': '1.3.0',
 'parsel': '1.10.0',
 'w3lib': '2.3.1',
 'Twisted': '25.5.0',
 'Python': '3.14.0 (tags/v3.14.0:ebf955d, Oct  7 2025, 10:15:03) [MSC v.1944 '
           '64 bit (AMD64)]',
 'pyOpenSSL': '25.3.0 (OpenSSL 3.5.4 30 Sep 2025)',
 'cryptography': '46.0.3',
 'Platform': 'Windows-11-10.0.26200-SP0'}
2025-11-23 17:43:55 [scrapy.addons] INFO: Enabled addons:
[]
2025-11-23 17:43:55 [scrapy.extensions.telnet] INFO: Telnet Password:
e50700739db06670
2025-11-23 17:43:55 [scrapy.middleware] INFO: Enabled extensions:
['scrapy.extensions.corestats.CoreStats',
 'scrapy.extensions.telnet.TelnetConsole',
```

```
 'scrapy.extensions.logstats.LogStats']
2025-11-23 17:43:55 [scrapy.crawler] INFO: Overridden settings:
{'LOG_LEVEL': 'INFO'}
2025-11-23 17:43:55 [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']
2025-11-23 17:43:55 [scrapy.middleware] INFO: Enabled spider middlewares:
['scrapy.spidermiddlewares.start.StartSpiderMiddleware',
 'scrapy.spidermiddlewares.httperror.HttpErrorMiddleware',
 'scrapy.spidermiddlewares.referer.RefererMiddleware',
 'scrapy.spidermiddlewares.urllength.UrlLengthMiddleware',
 'scrapy.spidermiddlewares.depth.DepthMiddleware']
2025-11-23 17:43:55 [scrapy.middleware] INFO: Enabled item pipelines:
[]
2025-11-23 17:43:55 [scrapy.middleware] INFO: Enabled spider middlewares:
['scrapy.spidermiddlewares.start.StartSpiderMiddleware',
 'scrapy.spidermiddlewares.httperror.HttpErrorMiddleware',
 'scrapy.spidermiddlewares.referer.RefererMiddleware',
 'scrapy.spidermiddlewares.urllength.UrlLengthMiddleware',
 'scrapy.spidermiddlewares.depth.DepthMiddleware']
2025-11-23 17:43:55 [scrapy.middleware] INFO: Enabled item pipelines:
[]

2025-11-23 17:43:55 [scrapy.core.engine] INFO: Spider opened
2025-11-23 17:43:55 [scrapy.extensions.logstats] INFO: Crawled 0 pages (at 0
pages/min), scraped 0 items (at 0 items/min)
2025-11-23 17:43:55 [scrapy.extensions.logstats] INFO: Crawled 0 pages (at 0
pages/min), scraped 0 items (at 0 items/min)
2025-11-23 17:43:55 [scrapy.extensions.telnet] INFO: Telnet console listening on
127.0.0.1:6023
2025-11-23 17:43:55 [scrapy.extensions.telnet] INFO: Telnet console listening on
127.0.0.1:6023
2025-11-23 17:43:55 [notebook_crawler] INFO: Saved page 1/100:
356811ed-d500-4aa1-9fcc-6f844fbd1b04.html
2025-11-23 17:43:55 [notebook_crawler] INFO: Saved page 1/100:
356811ed-d500-4aa1-9fcc-6f844fbd1b04.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 2/100:
03533d63-e885-4bdf-8e74-b63882da4dd8.html
```

```
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 2/100:
03533d63-e885-4bdf-8e74-b63882da4dd8.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 3/100:
86a78166-6e2c-4cdd-a773-245dd4220698.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 3/100:
86a78166-6e2c-4cdd-a773-245dd4220698.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 4/100:
19a234b6-7df4-49e1-b6bd-364fd0d6ea23.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 4/100:
19a234b6-7df4-49e1-b6bd-364fd0d6ea23.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 5/100:
ed6c0083-6310-4bf5-90c1-ee9f13dcc377.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 5/100:
ed6c0083-6310-4bf5-90c1-ee9f13dcc377.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 6/100:
f354fcf9-3722-45ab-8b46-150a6ee60db0.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 6/100:
f354fcf9-3722-45ab-8b46-150a6ee60db0.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 7/100:
fa0536d9-9f38-4846-a8e0-4efeb72994c7.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 7/100:
fa0536d9-9f38-4846-a8e0-4efeb72994c7.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 8/100:
3ff65671-55f0-4d2b-8d87-2dd67d1f24c9.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 8/100:
3ff65671-55f0-4d2b-8d87-2dd67d1f24c9.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 9/100:
0dae95d3-3135-4caa-b44f-ade7cc014e0b.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 9/100:
0dae95d3-3135-4caa-b44f-ade7cc014e0b.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 10/100:
33075000-3e3f-4b11-ae90-84f9f6721d2c.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 10/100:
33075000-3e3f-4b11-ae90-84f9f6721d2c.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 11/100:
9e1c6951-eea1-4c51-b3b2-09b404d54702.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 11/100:
9e1c6951-eea1-4c51-b3b2-09b404d54702.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 12/100:
bce5f152-2c66-4ee6-bcc7-a9624b5b287e.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 12/100:
bce5f152-2c66-4ee6-bcc7-a9624b5b287e.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 13/100:
44508a25-ed7f-4b22-bf1b-9b8550effb03.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 13/100:
44508a25-ed7f-4b22-bf1b-9b8550effb03.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 14/100:
afa844e2-31db-4006-bd08-eb5ab376e2c5.html
```

```
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 14/100:
afa844e2-31db-4006-bd08-eb5ab376e2c5.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 15/100:
0da8c278-15cf-49ba-893f-4f01d48c736e.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 15/100:
0da8c278-15cf-49ba-893f-4f01d48c736e.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 16/100:
bbb34dab-8dcd-42f0-b128-8b50315f274b.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 16/100:
bbb34dab-8dcd-42f0-b128-8b50315f274b.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 17/100:
28cb4559-e777-4a78-8f94-1b6ee3e14898.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 17/100:
28cb4559-e777-4a78-8f94-1b6ee3e14898.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 18/100:
c49d8fb5-fe9e-4eb9-8b80-d74b2028575c.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 18/100:
c49d8fb5-fe9e-4eb9-8b80-d74b2028575c.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 19/100:
e2d43c61-1903-4d7b-9768-55cd9d3348ea.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 19/100:
e2d43c61-1903-4d7b-9768-55cd9d3348ea.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 20/100:
cf0af01b-2e5f-46ec-91df-3b630a692c05.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 20/100:
cf0af01b-2e5f-46ec-91df-3b630a692c05.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 21/100:
3cf8318d-2d93-4a68-bcc0-4b04e7183cc1.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 21/100:
3cf8318d-2d93-4a68-bcc0-4b04e7183cc1.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 22/100:
8ac0576f-ebac-4302-b974-aeec096704ae.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 22/100:
8ac0576f-ebac-4302-b974-aeec096704ae.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 23/100:
9294c9d3-67cb-4793-bcf5-f50cc699eeeb.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 23/100:
9294c9d3-67cb-4793-bcf5-f50cc699eeeb.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 24/100:
a01a8fd3-176d-415e-ac51-64c517d9e902.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 24/100:
a01a8fd3-176d-415e-ac51-64c517d9e902.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 25/100:
9509b7e6-5e75-482b-a07c-029755a27c34.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 25/100:
9509b7e6-5e75-482b-a07c-029755a27c34.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 26/100:
6f1c372e-9227-43bb-9a31-58f558b8386a.html
```

```
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 26/100:
6f1c372e-9227-43bb-9a31-58f558b8386a.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 27/100:
94455b9f-aad4-4864-a83f-2a4718210b6f.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 27/100:
94455b9f-aad4-4864-a83f-2a4718210b6f.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 28/100:
832078ac-d3e3-4288-924a-9f7881321439.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 28/100:
832078ac-d3e3-4288-924a-9f7881321439.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 29/100:
f64b257f-c7fe-4bc5-81e4-db485722b1b1.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 29/100:
f64b257f-c7fe-4bc5-81e4-db485722b1b1.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 30/100:
e33c6fda-e024-40e0-90dd-11353d0c5487.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 30/100:
e33c6fda-e024-40e0-90dd-11353d0c5487.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 31/100:
21109878-e79f-4b27-9f51-b7e70480a36b.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 31/100:
21109878-e79f-4b27-9f51-b7e70480a36b.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 32/100:
4c96d6fd-1918-41d5-90f0-9a3fdb9ea8b3.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 32/100:
4c96d6fd-1918-41d5-90f0-9a3fdb9ea8b3.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 33/100:
fa7f9b73-ea06-4599-957f-33961bd04b5c.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 33/100:
fa7f9b73-ea06-4599-957f-33961bd04b5c.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 34/100:
a377702e-3849-4c58-827d-ef182f147f52.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 34/100:
a377702e-3849-4c58-827d-ef182f147f52.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 35/100:
577c0249-565d-4754-91c9-8555c1dcfbf6.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 35/100:
577c0249-565d-4754-91c9-8555c1dcfbf6.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 36/100:
fce510ef-2bdb-460c-9ca3-9d99aba229cc.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 36/100:
fce510ef-2bdb-460c-9ca3-9d99aba229cc.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 37/100:
ac091e85-7356-4bc5-869e-fbcc1c73cc9c.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 37/100:
ac091e85-7356-4bc5-869e-fbcc1c73cc9c.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 38/100:
e8c15b23-f845-4b91-b1e7-818e98197d1c.html
```

```
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 38/100:
e8c15b23-f845-4b91-b1e7-818e98197d1c.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 39/100:
6354b280-385f-4fa9-9d5f-fe164e2af3be.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 39/100:
6354b280-385f-4fa9-9d5f-fe164e2af3be.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 40/100:
706c4909-8213-4fc6-b8c2-be88d789c613.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 40/100:
706c4909-8213-4fc6-b8c2-be88d789c613.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 41/100:
8f24df9b-b331-4882-abc6-4a8830cfaed2.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 41/100:
8f24df9b-b331-4882-abc6-4a8830cfaed2.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 42/100:
775c6d24-f25c-49f9-8120-be4120844239.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 42/100:
775c6d24-f25c-49f9-8120-be4120844239.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 43/100:
f62bb571-8465-4745-b309-7d7aa42398cd.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 43/100:
f62bb571-8465-4745-b309-7d7aa42398cd.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 44/100:
7ad7d822-3ba6-4b49-94a7-c3ff7043864a.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 44/100:
7ad7d822-3ba6-4b49-94a7-c3ff7043864a.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 45/100:
34456388-96a6-4451-99f8-13757e0144a1.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 45/100:
34456388-96a6-4451-99f8-13757e0144a1.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 46/100:
0bb881a2-c200-43e5-abcf-e202b6f564fd.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 46/100:
0bb881a2-c200-43e5-abcf-e202b6f564fd.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 47/100:
dbab8f31-87ad-4e58-ab77-3927ca50145e.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 47/100:
dbab8f31-87ad-4e58-ab77-3927ca50145e.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 48/100:
6f06b128-5e1c-4208-983f-baf965292b97.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 48/100:
6f06b128-5e1c-4208-983f-baf965292b97.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 49/100:
e275c575-d82f-4323-a965-ebbabce2a70e.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 49/100:
e275c575-d82f-4323-a965-ebbabce2a70e.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 50/100:
d26943d7-dd22-4c77-8a3b-61b8ae2058e3.html
```

```
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 50/100:
d26943d7-dd22-4c77-8a3b-61b8ae2058e3.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 51/100:
13252d39-d3c6-4918-bb23-718b1f3578bc.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 51/100:
13252d39-d3c6-4918-bb23-718b1f3578bc.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 52/100:
a857a3fe-d52b-4e3c-81d9-adcc2e9e8d00.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 52/100:
a857a3fe-d52b-4e3c-81d9-adcc2e9e8d00.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 53/100:
e5b8dc4c-bd39-4d89-9538-6611019908a4.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 53/100:
e5b8dc4c-bd39-4d89-9538-6611019908a4.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 54/100: 90be2eee-
ec38-4b5f-a38d-4082ee5368a8.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 54/100: 90be2eee-
ec38-4b5f-a38d-4082ee5368a8.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 55/100:
b8c0b8eb-350c-4d87-9258-8884b165c531.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 55/100:
b8c0b8eb-350c-4d87-9258-8884b165c531.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 56/100:
a0587338-7399-4e30-9059-17dc04c906a2.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 56/100:
a0587338-7399-4e30-9059-17dc04c906a2.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 57/100:
d7b7beee-1eda-43b5-9c92-0b5e9b0b3171.html
2025-11-23 17:43:56 [notebook_crawler] INFO: Saved page 57/100:
d7b7beee-1eda-43b5-9c92-0b5e9b0b3171.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 58/100:
b3a79842-10cc-4c1e-8e39-c38e10dc62e5.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 58/100:
b3a79842-10cc-4c1e-8e39-c38e10dc62e5.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 59/100:
ce797a1f-21d5-444e-b638-1eef757125b8.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 59/100:
ce797a1f-21d5-444e-b638-1eef757125b8.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 60/100:
a07ffa78-4320-4c1f-8bcd-4398bcddade5.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 60/100:
a07ffa78-4320-4c1f-8bcd-4398bcddade5.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 61/100:
c5d3fc30-0fff-41a3-a371-16c771242c82.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 61/100:
c5d3fc30-0fff-41a3-a371-16c771242c82.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 62/100:
aab5b301-220e-491e-a613-74d0a9751631.html
```

```
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 62/100:
aab5b301-220e-491e-a613-74d0a9751631.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 63/100:
be2eeaa7-4a29-42a9-9f70-1b06a61db909.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 63/100:
be2eeaa7-4a29-42a9-9f70-1b06a61db909.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 64/100:
11c964eb-994d-4584-9666-e8cc7088125f.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 64/100:
11c964eb-994d-4584-9666-e8cc7088125f.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 65/100:
c3e12949-39b0-4999-a70a-2203d94dd4d2.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 65/100:
c3e12949-39b0-4999-a70a-2203d94dd4d2.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 66/100:
d6e393ec-676f-4a36-b535-5f113d250ff8.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 66/100:
d6e393ec-676f-4a36-b535-5f113d250ff8.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 67/100:
87038c58-f208-418b-b43a-059448b3ca17.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 67/100:
87038c58-f208-418b-b43a-059448b3ca17.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 68/100:
ce9d3307-c4c7-4c61-bbc8-f5e2ea2151b8.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 68/100:
ce9d3307-c4c7-4c61-bbc8-f5e2ea2151b8.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 69/100:
ead23f54-d066-4e5d-b03e-5ab9e88f1329.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 69/100:
ead23f54-d066-4e5d-b03e-5ab9e88f1329.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 70/100:
b11d8b06-22a2-4e50-bacf-0eed3e782dc8.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 70/100:
b11d8b06-22a2-4e50-bacf-0eed3e782dc8.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 71/100:
869fbe05-5186-446c-a1be-a6d9994a53a6.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 71/100:
869fbe05-5186-446c-a1be-a6d9994a53a6.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 72/100:
bf1d535b-20de-4cb4-aa5c-90c4fda16d6a.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 72/100:
bf1d535b-20de-4cb4-aa5c-90c4fda16d6a.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 73/100:
d4f5e2ef-137d-486e-96f8-d24c1ebb7f43.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 73/100:
d4f5e2ef-137d-486e-96f8-d24c1ebb7f43.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 74/100:
2d5c1a88-0838-4189-9c35-b9293a1979b0.html
```

```
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 74/100:
2d5c1a88-0838-4189-9c35-b9293a1979b0.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 75/100:
865eee11-c261-4491-a8b6-211827713e85.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 75/100:
865eee11-c261-4491-a8b6-211827713e85.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 76/100:
e980d629-1e20-451f-8cef-f9db9abffe84.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 76/100:
e980d629-1e20-451f-8cef-f9db9abffe84.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 77/100:
697749bf-0180-4bf5-b544-4a6b2a63f77d.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 77/100:
697749bf-0180-4bf5-b544-4a6b2a63f77d.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 78/100:
b6c71894-4e10-4ef2-8c71-693093de6004.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 78/100:
b6c71894-4e10-4ef2-8c71-693093de6004.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 79/100:
2d246143-0b39-4795-b73f-9e8e50f65345.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 79/100:
2d246143-0b39-4795-b73f-9e8e50f65345.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 80/100:
0151dabc-1519-4c16-89c2-7e01dbef2723.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 80/100:
0151dabc-1519-4c16-89c2-7e01dbef2723.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 81/100: 99c5e5ec-
cba0-4ea7-82eb-007c2fe6977b.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 81/100: 99c5e5ec-
cba0-4ea7-82eb-007c2fe6977b.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 82/100:
b2738930-3269-4e68-9bc2-847138865862.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 82/100:
b2738930-3269-4e68-9bc2-847138865862.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 83/100:
7ed635be-13ec-4320-a5bd-bfa2af345171.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 83/100:
7ed635be-13ec-4320-a5bd-bfa2af345171.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 84/100:
0200b28a-d172-4eaf-bcb8-f6cbefa3d371.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 84/100:
0200b28a-d172-4eaf-bcb8-f6cbefa3d371.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 85/100:
8b7c0c82-80f9-488f-b8ec-4241df9cea46.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 85/100:
8b7c0c82-80f9-488f-b8ec-4241df9cea46.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 86/100:
398bbde8-32ef-4b43-90cd-68231196e67a.html
```

```
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 86/100:
398bbde8-32ef-4b43-90cd-68231196e67a.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 87/100:
9f7b0f49-f860-4ef3-930f-c83c1612c01f.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 87/100:
9f7b0f49-f860-4ef3-930f-c83c1612c01f.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 88/100:
05d98930-8c67-4077-a16b-230742c1a196.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 88/100:
05d98930-8c67-4077-a16b-230742c1a196.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 89/100:
370de3a7-7817-4ad1-9a22-370d948cdce0.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 89/100:
370de3a7-7817-4ad1-9a22-370d948cdce0.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 90/100:
c1c301e6-e1ee-41e5-b8c7-5c22a010e352.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 90/100:
c1c301e6-e1ee-41e5-b8c7-5c22a010e352.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 91/100: 2343ecfb-
cea0-403b-883d-2809b1ce179e.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 91/100: 2343ecfb-
cea0-403b-883d-2809b1ce179e.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 92/100:
199fd625-5091-4809-bdef-684692e18622.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 92/100:
199fd625-5091-4809-bdef-684692e18622.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 93/100:
ba5fd02d-bca3-4b22-8336-ffe6ef036ea3.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 93/100:
ba5fd02d-bca3-4b22-8336-ffe6ef036ea3.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 94/100:
1c550459-5f94-403a-a234-2cea3d84b090.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 94/100:
1c550459-5f94-403a-a234-2cea3d84b090.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 95/100:
bc5d9424-7277-4cae-a102-82c3a53798bd.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 95/100:
bc5d9424-7277-4cae-a102-82c3a53798bd.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 96/100:
b5d68d86-a95e-46a3-923c-4b8744f64ac0.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 96/100:
b5d68d86-a95e-46a3-923c-4b8744f64ac0.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 97/100:
2d6c762b-03ec-4204-b718-83e7b3ff3eb5.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 97/100:
2d6c762b-03ec-4204-b718-83e7b3ff3eb5.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 98/100:
cbebda24-a8b8-4e37-bc60-f2371a5b8d5b.html
```

```
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 98/100:
cbebda24-a8b8-4e37-bc60-f2371a5b8d5b.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 99/100:
802da25f-289f-42b2-b27f-e39de840a04f.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 99/100:
802da25f-289f-42b2-b27f-e39de840a04f.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 100/100:
2f269255-0763-433d-9744-69babf18e0a1.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Saved page 100/100:
2f269255-0763-433d-9744-69babf18e0a1.html
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
```

```
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:57 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
```

```
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
```

```
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
```

```
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:58 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
```

```
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [notebook_crawler] INFO: Reached max pages limit: 100
2025-11-23 17:43:59 [scrapy.core.engine] INFO: Closing spider (finished)
2025-11-23 17:43:59 [scrapy.core.engine] INFO: Closing spider (finished)
2025-11-23 17:43:59 [scrapy.statscollectors] INFO: Dumping Scrapy stats:
{'downloader/request_bytes': 114573,
 'downloader/request_count': 295,
 'downloader/request_method_count/GET': 295,
 'downloader/response_bytes': 1239219,
 'downloader/response_count': 295,
 'downloader/response_status_count/200': 230,
 'downloader/response_status_count/308': 65,
 'dupefilter/filtered': 912,
 'elapsed_time_seconds': 3.683602,
 'finish_reason': 'finished',
 'finish_time': datetime.datetime(2025, 11, 23, 23, 43, 59, 186569,
tzinfo=datetime.timezone.utc),
 'items_per_minute': 0.0,
 'log_count/INFO': 240,
 'request_depth_max': 4,
```

```
 'response_received_count': 230,
 'responses_per_minute': 4600.0,
 'scheduler/dequeued': 295,
 'scheduler/dequeued/memory': 295,
 'scheduler/enqueued': 295,
 'scheduler/enqueued/memory': 295,
 'start_time': datetime.datetime(2025, 11, 23, 23, 43, 55, 502967,
tzinfo=datetime.timezone.utc)}
2025-11-23 17:43:59 [scrapy.statscollectors] INFO: Dumping Scrapy stats:
{'downloader/request_bytes': 114573,
 'downloader/request_count': 295,
 'downloader/request_method_count/GET': 295,
 'downloader/response_bytes': 1239219,
 'downloader/response_count': 295,
 'downloader/response_status_count/200': 230,
 'downloader/response_status_count/308': 65,
 'dupefilter/filtered': 912,
 'elapsed_time_seconds': 3.683602,
 'finish_reason': 'finished',
 'finish_time': datetime.datetime(2025, 11, 23, 23, 43, 59, 186569,
tzinfo=datetime.timezone.utc),
 'items_per_minute': 0.0,
 'log_count/INFO': 240,
 'request_depth_max': 4,
 'response_received_count': 230,
 'responses_per_minute': 4600.0,
 'scheduler/dequeued': 295,
 'scheduler/dequeued/memory': 295,
 'scheduler/enqueued': 295,
 'scheduler/enqueued/memory': 295,
 'start_time': datetime.datetime(2025, 11, 23, 23, 43, 55, 502967,
tzinfo=datetime.timezone.utc)}
2025-11-23 17:43:59 [scrapy.core.engine] INFO: Spider closed (finished)
2025-11-23 17:43:59 [scrapy.core.engine] INFO: Spider closed (finished)
```

0.8.2 Step 2: Indexing

[25]:
```python
from bs4 import BeautifulSoup
import os

# Create folder for cleaned text files
os.makedirs("cleaned_text", exist_ok=True)

def html_to_text(html_file):
    with open(html_file, 'r', encoding='utf-8') as f:
        soup = BeautifulSoup(f, 'html.parser')
```

```
    for script in soup(["script", "style"]):
        script.decompose()

    text = soup.get_text()

    lines = (line.strip() for line in text.splitlines())
    chunks = (phrase.strip() for line in lines for phrase in line.split("  "))
    text = ' '.join(chunk for chunk in chunks if chunk)

    return text

pages_dir = "pages"
documents = {}

for filename in os.listdir(pages_dir):
    if filename.endswith('.html'):
        filepath = os.path.join(pages_dir, filename)
        clean_text = html_to_text(filepath)
        documents[filename] = clean_text

        text_filename = filename.replace('.html', '.txt')
        with open(f"cleaned_text/{text_filename}", 'w', encoding='utf-8') as f:
            f.write(clean_text)

        print(f"Processed: {filename} ({len(clean_text)} characters)")
```

```
Processed: 0151dabc-1519-4c16-89c2-7e01dbef2723.html (352 characters)
Processed: 0200b28a-d172-4eaf-bcb8-f6cbefa3d371.html (295 characters)
Processed: 03533d63-e885-4bdf-8e74-b63882da4dd8.html (465 characters)
Processed: 05d98930-8c67-4077-a16b-230742c1a196.html (465 characters)
Processed: 0bb881a2-c200-43e5-abcf-e202b6f564fd.html (2525 characters)
Processed: 0da8c278-15cf-49ba-893f-4f01d48c736e.html (564 characters)
Processed: 0dae95d3-3135-4caa-b44f-ade7cc014e0b.html (1936 characters)
Processed: 11c964eb-994d-4584-9666-e8cc7088125f.html (373 characters)
Processed: 13252d39-d3c6-4918-bb23-718b1f3578bc.html (413 characters)
Processed: 199fd625-5091-4809-bdef-684692e18622.html (1673 characters)
Processed: 19a234b6-7df4-49e1-b6bd-364fd0d6ea23.html (401 characters)
Processed: 1c550459-5f94-403a-a234-2cea3d84b090.html (385 characters)
Processed: 21109878-e79f-4b27-9f51-b7e70480a36b.html (3967 characters)
Processed: 2343ecfb-cea0-403b-883d-2809b1ce179e.html (1936 characters)
Processed: 28cb4559-e777-4a78-8f94-1b6ee3e14898.html (751 characters)
Processed: 2d246143-0b39-4795-b73f-9e8e50f65345.html (356 characters)
Processed: 2d5c1a88-0838-4189-9c35-b9293a1979b0.html (356 characters)
Processed: 2d6c762b-03ec-4204-b718-83e7b3ff3eb5.html (751 characters)
Processed: 2f269255-0763-433d-9744-69babf18e0a1.html (315 characters)
Processed: 199fd625-5091-4809-bdef-684692e18622.html (1673 characters)
Processed: 19a234b6-7df4-49e1-b6bd-364fd0d6ea23.html (401 characters)
Processed: 1c550459-5f94-403a-a234-2cea3d84b090.html (385 characters)
```

```
Processed: 21109878-e79f-4b27-9f51-b7e70480a36b.html (3967 characters)
Processed: 2343ecfb-cea0-403b-883d-2809b1ce179e.html (1936 characters)
Processed: 28cb4559-e777-4a78-8f94-1b6ee3e14898.html (751 characters)
Processed: 2d246143-0b39-4795-b73f-9e8e50f65345.html (356 characters)
Processed: 2d5c1a88-0838-4189-9c35-b9293a1979b0.html (356 characters)
Processed: 2d6c762b-03ec-4204-b718-83e7b3ff3eb5.html (751 characters)
Processed: 2f269255-0763-433d-9744-69babf18e0a1.html (315 characters)
Processed: 33075000-3e3f-4b11-ae90-84f9f6721d2c.html (2843 characters)
Processed: 34456388-96a6-4451-99f8-13757e0144a1.html (2674 characters)
Processed: 356811ed-d500-4aa1-9fcc-6f844fbd1b04.html (1681 characters)
Processed: 370de3a7-7817-4ad1-9a22-370d948cdce0.html (392 characters)
Processed: 398bbde8-32ef-4b43-90cd-68231196e67a.html (355 characters)
Processed: 3cf8318d-2d93-4a68-bcc0-4b04e7183cc1.html (1447 characters)
Processed: 3ff65671-55f0-4d2b-8d87-2dd67d1f24c9.html (360 characters)
Processed: 44508a25-ed7f-4b22-bf1b-9b8550effb03.html (378 characters)
Processed: 4c96d6fd-1918-41d5-90f0-9a3fdb9ea8b3.html (676 characters)
Processed: 33075000-3e3f-4b11-ae90-84f9f6721d2c.html (2843 characters)
Processed: 34456388-96a6-4451-99f8-13757e0144a1.html (2674 characters)
Processed: 356811ed-d500-4aa1-9fcc-6f844fbd1b04.html (1681 characters)
Processed: 370de3a7-7817-4ad1-9a22-370d948cdce0.html (392 characters)
Processed: 398bbde8-32ef-4b43-90cd-68231196e67a.html (355 characters)
Processed: 3cf8318d-2d93-4a68-bcc0-4b04e7183cc1.html (1447 characters)
Processed: 3ff65671-55f0-4d2b-8d87-2dd67d1f24c9.html (360 characters)
Processed: 44508a25-ed7f-4b22-bf1b-9b8550effb03.html (378 characters)
Processed: 4c96d6fd-1918-41d5-90f0-9a3fdb9ea8b3.html (676 characters)
Processed: 577c0249-565d-4754-91c9-8555c1dcfbf6.html (2822 characters)
Processed: 6354b280-385f-4fa9-9d5f-fe164e2af3be.html (380 characters)
Processed: 697749bf-0180-4bf5-b544-4a6b2a63f77d.html (476 characters)
Processed: 6f06b128-5e1c-4208-983f-baf965292b97.html (411 characters)
Processed: 6f1c372e-9227-43bb-9a31-58f558b8386a.html (330 characters)
Processed: 706c4909-8213-4fc6-b8c2-be88d789c613.html (1325 characters)
Processed: 775c6d24-f25c-49f9-8120-be4120844239.html (389 characters)
Processed: 7ad7d822-3ba6-4b49-94a7-c3ff7043864a.html (399 characters)
Processed: 7ed635be-13ec-4320-a5bd-bfa2af345171.html (2148 characters)
Processed: 802da25f-289f-42b2-b27f-e39de840a04f.html (2359 characters)
Processed: 577c0249-565d-4754-91c9-8555c1dcfbf6.html (2822 characters)
Processed: 6354b280-385f-4fa9-9d5f-fe164e2af3be.html (380 characters)
Processed: 697749bf-0180-4bf5-b544-4a6b2a63f77d.html (476 characters)
Processed: 6f06b128-5e1c-4208-983f-baf965292b97.html (411 characters)
Processed: 6f1c372e-9227-43bb-9a31-58f558b8386a.html (330 characters)
Processed: 706c4909-8213-4fc6-b8c2-be88d789c613.html (1325 characters)
Processed: 775c6d24-f25c-49f9-8120-be4120844239.html (389 characters)
Processed: 7ad7d822-3ba6-4b49-94a7-c3ff7043864a.html (399 characters)
Processed: 7ed635be-13ec-4320-a5bd-bfa2af345171.html (2148 characters)
Processed: 802da25f-289f-42b2-b27f-e39de840a04f.html (2359 characters)
Processed: 832078ac-d3e3-4288-924a-9f7881321439.html (1976 characters)
Processed: 865eee11-c261-4491-a8b6-211827713e85.html (351 characters)
Processed: 869fbe05-5186-446c-a1be-a6d9994a53a6.html (1934 characters)
```

```
Processed: 86a78166-6e2c-4cdd-a773-245dd4220698.html (2880 characters)
Processed: 87038c58-f208-418b-b43a-059448b3ca17.html (561 characters)
Processed: 8ac0576f-ebac-4302-b974-aeec096704ae.html (1936 characters)
Processed: 8b7c0c82-80f9-488f-b8ec-4241df9cea46.html (2375 characters)
Processed: 8f24df9b-b331-4882-abc6-4a8830cfaed2.html (397 characters)
Processed: 90be2eee-ec38-4b5f-a38d-4082ee5368a8.html (3936 characters)
Processed: 832078ac-d3e3-4288-924a-9f7881321439.html (1976 characters)
Processed: 865eee11-c261-4491-a8b6-211827713e85.html (351 characters)
Processed: 869fbe05-5186-446c-a1be-a6d9994a53a6.html (1934 characters)
Processed: 86a78166-6e2c-4cdd-a773-245dd4220698.html (2880 characters)
Processed: 87038c58-f208-418b-b43a-059448b3ca17.html (561 characters)
Processed: 8ac0576f-ebac-4302-b974-aeec096704ae.html (1936 characters)
Processed: 8b7c0c82-80f9-488f-b8ec-4241df9cea46.html (2375 characters)
Processed: 8f24df9b-b331-4882-abc6-4a8830cfaed2.html (397 characters)
Processed: 90be2eee-ec38-4b5f-a38d-4082ee5368a8.html (3936 characters)
Processed: 9294c9d3-67cb-4793-bcf5-f50cc699eeeb.html (1673 characters)
Processed: 94455b9f-aad4-4864-a83f-2a4718210b6f.html (647 characters)
Processed: 9509b7e6-5e75-482b-a07c-029755a27c34.html (512 characters)
Processed: 99c5e5ec-cba0-4ea7-82eb-007c2fe6977b.html (363 characters)
Processed: 9e1c6951-eea1-4c51-b3b2-09b404d54702.html (385 characters)
Processed: 9f7b0f49-f860-4ef3-930f-c83c1612c01f.html (751 characters)
Processed: a01a8fd3-176d-415e-ac51-64c517d9e902.html (2129 characters)
Processed: a0587338-7399-4e30-9059-17dc04c906a2.html (1646 characters)
Processed: a07ffa78-4320-4c1f-8bcd-4398bcddade5.html (466 characters)
Processed: a377702e-3849-4c58-827d-ef182f147f52.html (600 characters)
Processed: a857a3fe-d52b-4e3c-81d9-adcc2e9e8d00.html (409 characters)
Processed: aab5b301-220e-491e-a613-74d0a9751631.html (1794 characters)
Processed: 9294c9d3-67cb-4793-bcf5-f50cc699eeeb.html (1673 characters)
Processed: 94455b9f-aad4-4864-a83f-2a4718210b6f.html (647 characters)
Processed: 9509b7e6-5e75-482b-a07c-029755a27c34.html (512 characters)
Processed: 99c5e5ec-cba0-4ea7-82eb-007c2fe6977b.html (363 characters)
Processed: 9e1c6951-eea1-4c51-b3b2-09b404d54702.html (385 characters)
Processed: 9f7b0f49-f860-4ef3-930f-c83c1612c01f.html (751 characters)
Processed: a01a8fd3-176d-415e-ac51-64c517d9e902.html (2129 characters)
Processed: a0587338-7399-4e30-9059-17dc04c906a2.html (1646 characters)
Processed: a07ffa78-4320-4c1f-8bcd-4398bcddade5.html (466 characters)
Processed: a377702e-3849-4c58-827d-ef182f147f52.html (600 characters)
Processed: a857a3fe-d52b-4e3c-81d9-adcc2e9e8d00.html (409 characters)
Processed: aab5b301-220e-491e-a613-74d0a9751631.html (1794 characters)
Processed: ac091e85-7356-4bc5-869e-fbcc1c73cc9c.html (1003 characters)
Processed: afa844e2-31db-4006-bd08-eb5ab376e2c5.html (377 characters)
Processed: b11d8b06-22a2-4e50-bacf-0eed3e782dc8.html (1740 characters)
Processed: b2738930-3269-4e68-9bc2-847138865862.html (358 characters)
Processed: b3a79842-10cc-4c1e-8e39-c38e10dc62e5.html (321 characters)
Processed: b5d68d86-a95e-46a3-923c-4b8744f64ac0.html (1226 characters)
Processed: b6c71894-4e10-4ef2-8c71-693093de6004.html (1130 characters)
Processed: b8c0b8eb-350c-4d87-9258-8884b165c531.html (102 characters)
Processed: ba5fd02d-bca3-4b22-8336-ffe6ef036ea3.html (383 characters)
```

```
Processed: bbb34dab-8dcd-42f0-b128-8b50315f274b.html (505 characters)
Processed: ac091e85-7356-4bc5-869e-fbcc1c73cc9c.html (1003 characters)
Processed: afa844e2-31db-4006-bd08-eb5ab376e2c5.html (377 characters)
Processed: b11d8b06-22a2-4e50-bacf-0eed3e782dc8.html (1740 characters)
Processed: b2738930-3269-4e68-9bc2-847138865862.html (358 characters)
Processed: b3a79842-10cc-4c1e-8e39-c38e10dc62e5.html (321 characters)
Processed: b5d68d86-a95e-46a3-923c-4b8744f64ac0.html (1226 characters)
Processed: b6c71894-4e10-4ef2-8c71-693093de6004.html (1130 characters)
Processed: b8c0b8eb-350c-4d87-9258-8884b165c531.html (102 characters)
Processed: ba5fd02d-bca3-4b22-8336-ffe6ef036ea3.html (383 characters)
Processed: bbb34dab-8dcd-42f0-b128-8b50315f274b.html (505 characters)
Processed: bc5d9424-7277-4cae-a102-82c3a53798bd.html (505 characters)
Processed: bce5f152-2c66-4ee6-bcc7-a9624b5b287e.html (330 characters)
Processed: be2eeaa7-4a29-42a9-9f70-1b06a61db909.html (448 characters)
Processed: bf1d535b-20de-4cb4-aa5c-90c4fda16d6a.html (1299 characters)
Processed: c1c301e6-e1ee-41e5-b8c7-5c22a010e352.html (368 characters)
Processed: c3e12949-39b0-4999-a70a-2203d94dd4d2.html (1681 characters)
Processed: c49d8fb5-fe9e-4eb9-8b80-d74b2028575c.html (1447 characters)
Processed: c5d3fc30-0fff-41a3-a371-16c771242c82.html (2141 characters)
Processed: cbebda24-a8b8-4e37-bc60-f2371a5b8d5b.html (1768 characters)
Processed: ce797a1f-21d5-444e-b638-1eef757125b8.html (326 characters)
Processed: bc5d9424-7277-4cae-a102-82c3a53798bd.html (505 characters)
Processed: bce5f152-2c66-4ee6-bcc7-a9624b5b287e.html (330 characters)
Processed: be2eeaa7-4a29-42a9-9f70-1b06a61db909.html (448 characters)
Processed: bf1d535b-20de-4cb4-aa5c-90c4fda16d6a.html (1299 characters)
Processed: c1c301e6-e1ee-41e5-b8c7-5c22a010e352.html (368 characters)
Processed: c3e12949-39b0-4999-a70a-2203d94dd4d2.html (1681 characters)
Processed: c49d8fb5-fe9e-4eb9-8b80-d74b2028575c.html (1447 characters)
Processed: c5d3fc30-0fff-41a3-a371-16c771242c82.html (2141 characters)
Processed: cbebda24-a8b8-4e37-bc60-f2371a5b8d5b.html (1768 characters)
Processed: ce797a1f-21d5-444e-b638-1eef757125b8.html (326 characters)
Processed: ce9d3307-c4c7-4c61-bbc8-f5e2ea2151b8.html (318 characters)
Processed: cf0af01b-2e5f-46ec-91df-3b630a692c05.html (1000 characters)
Processed: d26943d7-dd22-4c77-8a3b-61b8ae2058e3.html (411 characters)
Processed: d4f5e2ef-137d-486e-96f8-d24c1ebb7f43.html (3743 characters)
Processed: d6e393ec-676f-4a36-b535-5f113d250ff8.html (367 characters)
Processed: d7b7beee-1eda-43b5-9c92-0b5e9b0b3171.html (994 characters)
Processed: dbab8f31-87ad-4e58-ab77-3927ca50145e.html (2114 characters)
Processed: e275c575-d82f-4323-a965-ebbabce2a70e.html (408 characters)
Processed: e2d43c61-1903-4d7b-9768-55cd9d3348ea.html (1768 characters)
Processed: e33c6fda-e024-40e0-90dd-11353d0c5487.html (522 characters)
Processed: ce9d3307-c4c7-4c61-bbc8-f5e2ea2151b8.html (318 characters)
Processed: cf0af01b-2e5f-46ec-91df-3b630a692c05.html (1000 characters)
Processed: d26943d7-dd22-4c77-8a3b-61b8ae2058e3.html (411 characters)
Processed: d4f5e2ef-137d-486e-96f8-d24c1ebb7f43.html (3743 characters)
Processed: d6e393ec-676f-4a36-b535-5f113d250ff8.html (367 characters)
Processed: d7b7beee-1eda-43b5-9c92-0b5e9b0b3171.html (994 characters)
Processed: dbab8f31-87ad-4e58-ab77-3927ca50145e.html (2114 characters)
```

```
Processed: e275c575-d82f-4323-a965-ebbabce2a70e.html (408 characters)
Processed: e2d43c61-1903-4d7b-9768-55cd9d3348ea.html (1768 characters)
Processed: e33c6fda-e024-40e0-90dd-11353d0c5487.html (522 characters)
Processed: e5b8dc4c-bd39-4d89-9538-6611019908a4.html (410 characters)
Processed: e8c15b23-f845-4b91-b1e7-818e98197d1c.html (337 characters)
Processed: e980d629-1e20-451f-8cef-f9db9abffe84.html (353 characters)
Processed: ead23f54-d066-4e5d-b03e-5ab9e88f1329.html (364 characters)
Processed: ed6c0083-6310-4bf5-90c1-ee9f13dcc377.html (328 characters)
Processed: f354fcf9-3722-45ab-8b46-150a6ee60db0.html (400 characters)
Processed: f62bb571-8465-4745-b309-7d7aa42398cd.html (393 characters)
Processed: f64b257f-c7fe-4bc5-81e4-db485722b1b1.html (526 characters)
Processed: fa0536d9-9f38-4846-a8e0-4efeb72994c7.html (397 characters)
Processed: fa7f9b73-ea06-4599-957f-33961bd04b5c.html (518 characters)
Processed: fce510ef-2bdb-460c-9ca3-9d99aba229cc.html (520 characters)
Processed: e5b8dc4c-bd39-4d89-9538-6611019908a4.html (410 characters)
Processed: e8c15b23-f845-4b91-b1e7-818e98197d1c.html (337 characters)
Processed: e980d629-1e20-451f-8cef-f9db9abffe84.html (353 characters)
Processed: ead23f54-d066-4e5d-b03e-5ab9e88f1329.html (364 characters)
Processed: ed6c0083-6310-4bf5-90c1-ee9f13dcc377.html (328 characters)
Processed: f354fcf9-3722-45ab-8b46-150a6ee60db0.html (400 characters)
Processed: f62bb571-8465-4745-b309-7d7aa42398cd.html (393 characters)
Processed: f64b257f-c7fe-4bc5-81e4-db485722b1b1.html (526 characters)
Processed: fa0536d9-9f38-4846-a8e0-4efeb72994c7.html (397 characters)
Processed: fa7f9b73-ea06-4599-957f-33961bd04b5c.html (518 characters)
Processed: fce510ef-2bdb-460c-9ca3-9d99aba229cc.html (520 characters)
```

[ ]:
```python
import json
import re
from collections import defaultdict

def tokenize_with_positions(text):
    text = text.lower()
    tokens = re.findall(r'\b[a-z]+\b', text)

    token_positions = defaultdict(list)
    for pos, token in enumerate(tokens):
        token_positions[token].append(pos)

    return dict(token_positions)

def build_positional_inverted_index(documents):
    inverted_index = defaultdict(list)

    for doc_id, text in documents.items():
        clean_doc_id = doc_id.replace('.html', '')

        token_positions = tokenize_with_positions(text)
```

```
        for token, positions in token_positions.items():
            inverted_index[token].append([clean_doc_id, positions])

    return dict(inverted_index)

inverted_index = build_positional_inverted_index(documents)

# Save to JSON file with custom formatting
with open('index.json', 'w', encoding='utf-8') as f:
    f.write('{\n')
    items = list(inverted_index.items())
    for i, (token, entries) in enumerate(items):
        f.write(f'  "{token}": [\n')
        for j, entry in enumerate(entries):
            entry_json = json.dumps(entry)
            if j < len(entries) - 1:
                f.write(f'    {entry_json},\n')
            else:
                f.write(f'    {entry_json}\n')
        if i < len(items) - 1:
            f.write('  ],\n')
        else:
            f.write('  ]\n')
    f.write('}\n')

print(f"Inverted index saved to index.json")
print(f"Total unique tokens: {len(inverted_index)}")
```

```
Inverted index saved to index.json
Total unique tokens: 2586
```

[1]:
```
import json
import os
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
import numpy as np

# Load the positional inverted index from index.json
with open('index.json', 'r', encoding='utf-8') as f:
    index = json.load(f)

# Get all doc_ids from the index
all_doc_ids = set()
for token_entries in index.values():
    for doc_id, _ in token_entries:
        all_doc_ids.add(doc_id)
```

```
all_doc_ids = list(all_doc_ids)

def get_doc_text(doc_id):
    filepath = f"cleaned_text/{doc_id}.txt"
    if os.path.exists(filepath):
        with open(filepath, 'r', encoding='utf-8') as f:
            return f.read()
    return ""

def search_tfidf(query_text, top_k):
    # Use only doc_ids from index.json
    doc_id_list = sorted(all_doc_ids)
    corpus = [get_doc_text(doc_id) for doc_id in doc_id_list]
    vectorizer = TfidfVectorizer(lowercase=True, token_pattern=r'\b[a-z]+\b')
    tfidf_matrix = vectorizer.fit_transform(corpus)
    feature_names = vectorizer.get_feature_names_out()
    query_terms = [t for t in query_text.lower().split() if t.isalpha() or t.
 ↪isalnum()]
    query_string = ' '.join(query_terms)
    query_vector = vectorizer.transform([query_string])
    similarities = cosine_similarity(query_vector, tfidf_matrix).flatten()
    results = []
    for idx, doc_id in enumerate(doc_id_list):
        score = similarities[idx]
        tfidf_weights = {}
        doc_vector = tfidf_matrix[idx]
        for query_term in query_terms:
            if query_term in feature_names:
                term_idx = np.where(feature_names == query_term)[0]
                tfidf_weights[query_term] = doc_vector[0, term_idx[0]] if␣
 ↪len(term_idx) > 0 else 0.0
            else:
                tfidf_weights[query_term] = 0.0
        results.append((doc_id, score, tfidf_weights))
    results.sort(key=lambda x: x[1], reverse=True)
    return results[:top_k]

# Example query
query = "better to be"
results = search_tfidf(query, top_k=5)
print(f"\nTop {5} results for query: {query}\n")
for rank, (doc_id, score, tfidf_weights) in enumerate(results, 1):
    print(f"{rank}. Document: {doc_id}")
    print(f"   Cosine Similarity Score: {score:.4f}")
    print(f"   TF-IDF Weights: {', '.join([f'{term}: {weight:.4f}' for term,␣
 ↪weight in tfidf_weights.items()])}")
    print()
```

```
Top 5 results for query: better to be

1. Document: f75ec424-708b-4640-ac8b-3a640a13a4b6
   Cosine Similarity Score: 0.3420
   TF-IDF Weights: better: 0.1790, to: 0.1715, be: 0.3064

2. Document: 3f86168d-959e-4134-83ac-3adf63083b23
   Cosine Similarity Score: 0.2103
   TF-IDF Weights: better: 0.1006, to: 0.1125, be: 0.2010

3. Document: 6f4b0467-50a1-42c9-bf00-ead1e8651c44
   Cosine Similarity Score: 0.2103
   TF-IDF Weights: better: 0.1006, to: 0.1125, be: 0.2010

4. Document: 9ecbcd87-0af8-471e-a0e2-1168d0a8495c
   Cosine Similarity Score: 0.2103
   TF-IDF Weights: better: 0.1006, to: 0.1125, be: 0.2010

5. Document: ad531183-964e-46d7-bafa-ff9869d87f75
   Cosine Similarity Score: 0.2103
   TF-IDF Weights: better: 0.1006, to: 0.1125, be: 0.2010
```

0.8.3 Step 3: Query Processing

[ ]:
```python
import csv
import uuid
import random

sample_queries = [
    "to be or not to be",
    "We read to know we're not alone.",
    "cup of tea large enough"
]

with open('queries.csv', 'w', newline='', encoding='utf-8') as f:
    writer = csv.writer(f)
    writer.writerow(['query_id', 'query_text'])
    for q in sample_queries:
        writer.writerow([str(uuid.uuid4()).upper(), q])
print("Sample queries.csv created.")

with open('queries.csv', 'r', encoding='utf-8') as f:
    reader = csv.DictReader(f)
    queries = list(reader)
```

```
Sample queries.csv created.
```

[ ]:
```
from flask import Flask, request, jsonify, send_file
import os
import csv
import io
import json
import numpy as np
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
import re

app = Flask(__name__)

# Global variables
all_doc_ids = None
inverted_index = None


def load_index_metadata(index_path='index.json'):
    """
    Load index.json to get all document IDs
    """
    global all_doc_ids, inverted_index

    print(f"Loading index from {index_path}...")

    with open(index_path, 'r', encoding='utf-8') as f:
        inverted_index = json.load(f)

    # Extract all unique document IDs from the index
    all_doc_ids = set()
    for term, postings in inverted_index.items():
        for doc_id, positions in postings:
            all_doc_ids.add(doc_id)

    all_doc_ids = sorted(list(all_doc_ids))

    print(f"  Index loaded: {len(inverted_index)} terms, {len(all_doc_ids)}␣
 ↪documents")

    return all_doc_ids


def get_doc_text(doc_id):
    """
    Load document text from cleaned_text directory
    """
    file_path = os.path.join('cleaned_text', f'{doc_id}.txt')
```

```
    if not os.path.exists(file_path):
        return ""
    with open(file_path, 'r', encoding='utf-8') as f:
        return f.read()


def search_tfidf(query_text, top_k):
    doc_id_list = sorted(all_doc_ids)
    corpus = [get_doc_text(doc_id) for doc_id in doc_id_list]

    vectorizer = TfidfVectorizer(lowercase=True, token_pattern=r'\b[a-z]+\b')
    tfidf_matrix = vectorizer.fit_transform(corpus)

    query_vector = vectorizer.transform([query_text.lower()])
    similarities = cosine_similarity(query_vector, tfidf_matrix).flatten()

    ranked_results = list(zip(doc_id_list, similarities))
    ranked_results.sort(key=lambda x: x[1], reverse=True)

    return ranked_results[:top_k]


def validate_query(query_text):
    if not query_text or not isinstance(query_text, str):
        return False, None, "Query text is required"
    query_text = query_text.strip()
    if len(query_text) == 0:
        return False, None, "Query cannot be empty"
    if len(query_text) > 500:
        return False, None, "Query too long"
    return True, query_text, None


def parse_query_csv(file_content):
    queries = []
    csv_data = io.StringIO(file_content.decode('utf-8'))
    reader = csv.DictReader(csv_data)

    if 'query_id' not in reader.fieldnames or 'query_text' not in reader.
 ↪fieldnames:
        raise ValueError("CSV must contain query_id and query_text")

    for row in reader:
        if not row['query_id'] or not row['query_text']:
            continue
        queries.append((row['query_id'].strip(), row['query_text'].strip()))
```

```
    return queries


@app.route('/')
def home():
    return jsonify({
        "message": "Search Engine Query Processor",
        "documents_loaded": len(all_doc_ids) if all_doc_ids else 0,
        "endpoints": {
            "/search/batch": "Upload CSV → download ranked CSV",
        }
    })


@app.route('/search/batch', methods=['POST'])
def search_batch_endpoint():
    if all_doc_ids is None:
        return jsonify({"error": "Index not loaded"}), 500

    if 'file' not in request.files:
        return jsonify({"error": "Missing CSV file"}), 400

    file = request.files['file']
    top_k = int(request.form.get('top_k', 3))
    csv_content = file.read()

    queries = parse_query_csv(csv_content)
    results_out = []

    for query_id, text in queries:
        valid, cleaned, err = validate_query(text)
        if not valid:
            continue

        results = search_tfidf(cleaned, top_k)

        for rank, (doc_id, score) in enumerate(results, 1):
            results_out.append({
                "query_id": query_id,
                "rank": rank,
                "document_id": doc_id
            })

    # Create output CSV
    output = io.StringIO()
    writer = csv.DictWriter(output, fieldnames=["query_id", "rank",␣
 ↪"document_id"])
```

```
    writer.writeheader()
    writer.writerows(results_out)
    output.seek(0)

    return send_file(
        io.BytesIO(output.getvalue().encode('utf-8')),
        mimetype='text/csv',
        as_attachment=True,
        download_name="results.csv"
    )


# --------------------------
# Start server
# --------------------------
if __name__ == '__main__':
    if not os.path.exists('index.json'):
        print("ERROR: index.json not found!")
    else:
        load_index_metadata()

    if not os.path.exists('cleaned_text'):
        print("WARNING: cleaned_text directory missing!")

    app.run(debug=True, host='0.0.0.0', port=5000, use_reloader=False)
```

```
Loading index from index.json…
  Index loaded: 2422 terms, 100 documents
 * Serving Flask app '__main__'
 * Debug mode: on
WARNING: This is a development server. Do not use it in a production deployment.
Use a production WSGI server instead.
 * Running on all addresses (0.0.0.0)
 * Running on http://127.0.0.1:5000
 * Running on http://10.0.0.180:5000
Press CTRL+C to quit
127.0.0.1 - - [30/Nov/2025 19:09:41] "POST /search/batch HTTP/1.1" 200 -
```