Web Crawler and Visualization Tool

Final Project Report

Name: Ibrahim Mohammed
NetID: im2707

May 9, 2025

Overview

This project implements a multi-threaded Java-based web crawler with real-time graphical visualization using Swing. It not only discovers hyperlinks starting from a given URL but also presents interactive analytics like domain filtering, live graph updates, depth-based coloring, and PageRank-based influence scores.

Technologies Used

- Java 17
- Swing (for UI and graph visualization)
- SQLite (local persistence)
- Jsoup (for HTML parsing)
- Git + GitHub for version control

Key Features

1. **Multi-threaded Crawl Engine:** Crawls pages concurrently while respecting robots.txt.
2. **Live Visualization:** Displays a graph where each node is a URL and edges represent hyperlinks.
3. **Depth and Domain Coloring:** Nodes are colored by crawl depth, and edges distinguish internal/external links.
4. **PageRank Analytics:** Computes the importance of pages based on link structure.
5. **Export Options:** Graphs can be exported as PNG images and crawl data as CSV reports.

How to Run

1. **Prerequisites:** Ensure Java (JDK 17+) and SQLite JDBC driver are available.
2. **From Eclipse:**
    - Import the project.

- Run `Main.java` as a Java Application.

3. **From CLI:**

```
javac -cp ".:sqlite-jdbc.jar:jsoup.jar" com/ibrahim/webcrawler/*.java
java -cp ".:sqlite-jdbc.jar:jsoup.jar" com.ibrahim.webcrawler.CrawlerUI
```

How to Use

- Enter a starting URL and desired depth.
- Use the complete link including https.
- Optionally enable domain filtering and specify crawl keywords.
- Click `Start Crawl`. The live log will update in real time.
- View analytics, export graphs and reports via respective buttons.

Screenshots

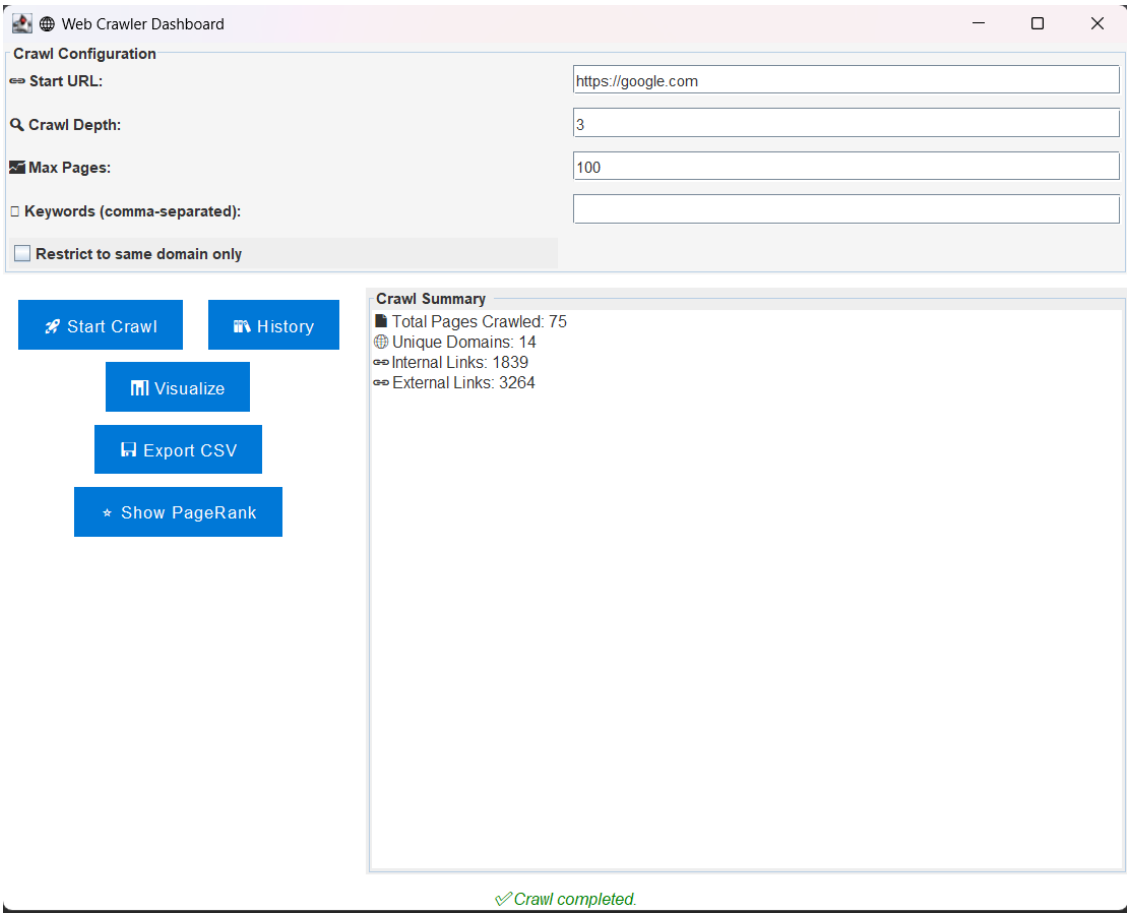
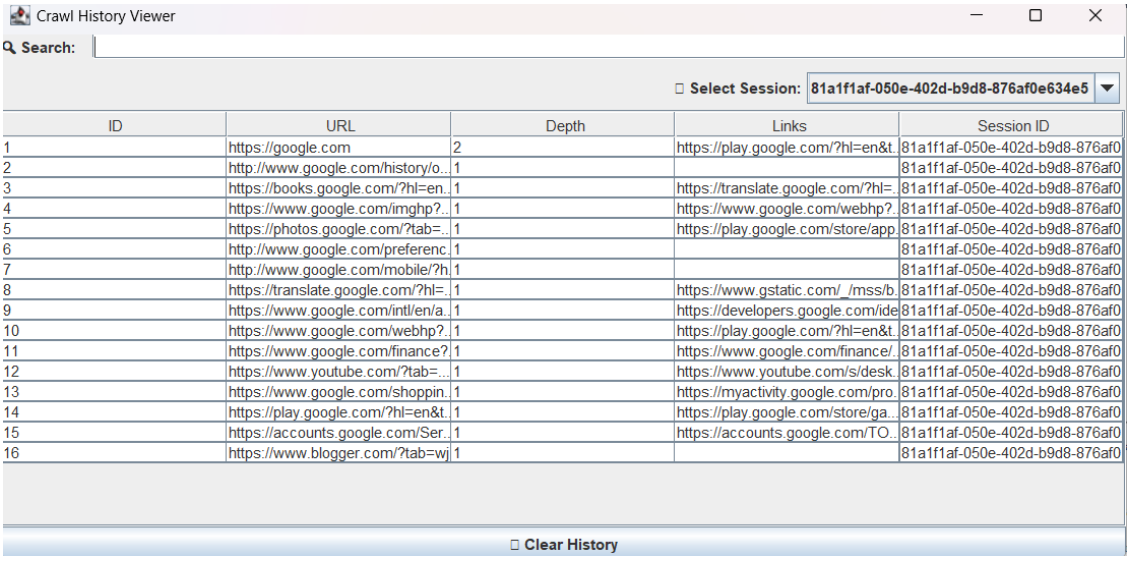


Figure 1: Main Crawler Dashboard UI



ID	URL	Depth	Links	Session ID
1	https://google.com	2	https://play.google.com/?hl=en&t..	81a1f1af-050e-402d-b9d8-876af0
2	http://www.google.com/history/o...	1		81a1f1af-050e-402d-b9d8-876af0
3	https://books.google.com/?hl=en..	1	https://translate.google.com/?hl=..	81a1f1af-050e-402d-b9d8-876af0
4	https://www.google.com/imghp?..	1	https://www.google.com/webhp?..	81a1f1af-050e-402d-b9d8-876af0
5	https://photos.google.com/?tab=..	1	https://play.google.com/store/app..	81a1f1af-050e-402d-b9d8-876af0
6	http://www.google.com/preferenc..	1		81a1f1af-050e-402d-b9d8-876af0
7	http://www.google.com/mobile/?h..	1		81a1f1af-050e-402d-b9d8-876af0
8	https://translate.google.com/?hl=..	1	https://www.gstatic.com/_/mss/b..	81a1f1af-050e-402d-b9d8-876af0
9	https://www.google.com/intl/en/a..	1	https://developers.google.com/ide..	81a1f1af-050e-402d-b9d8-876af0
10	https://www.google.com/webhp?..	1	https://play.google.com/?hl=en&t..	81a1f1af-050e-402d-b9d8-876af0
11	https://www.google.com/finance?..	1	https://www.google.com/finance/..	81a1f1af-050e-402d-b9d8-876af0
12	https://www.youtube.com/?tab=...	1	https://www.youtube.com/s/desk..	81a1f1af-050e-402d-b9d8-876af0
13	https://www.google.com/shoppin..	1	https://myactivity.google.com/pro..	81a1f1af-050e-402d-b9d8-876af0
14	https://play.google.com/?hl=en&t..	1	https://play.google.com/store/ga..	81a1f1af-050e-402d-b9d8-876af0
15	https://accounts.google.com/Ser..	1	https://accounts.google.com/TO..	81a1f1af-050e-402d-b9d8-876af0
16	https://www.blogger.com/?tab=wj	1		81a1f1af-050e-402d-b9d8-876af0

Figure 2: Database History of All the Crawls

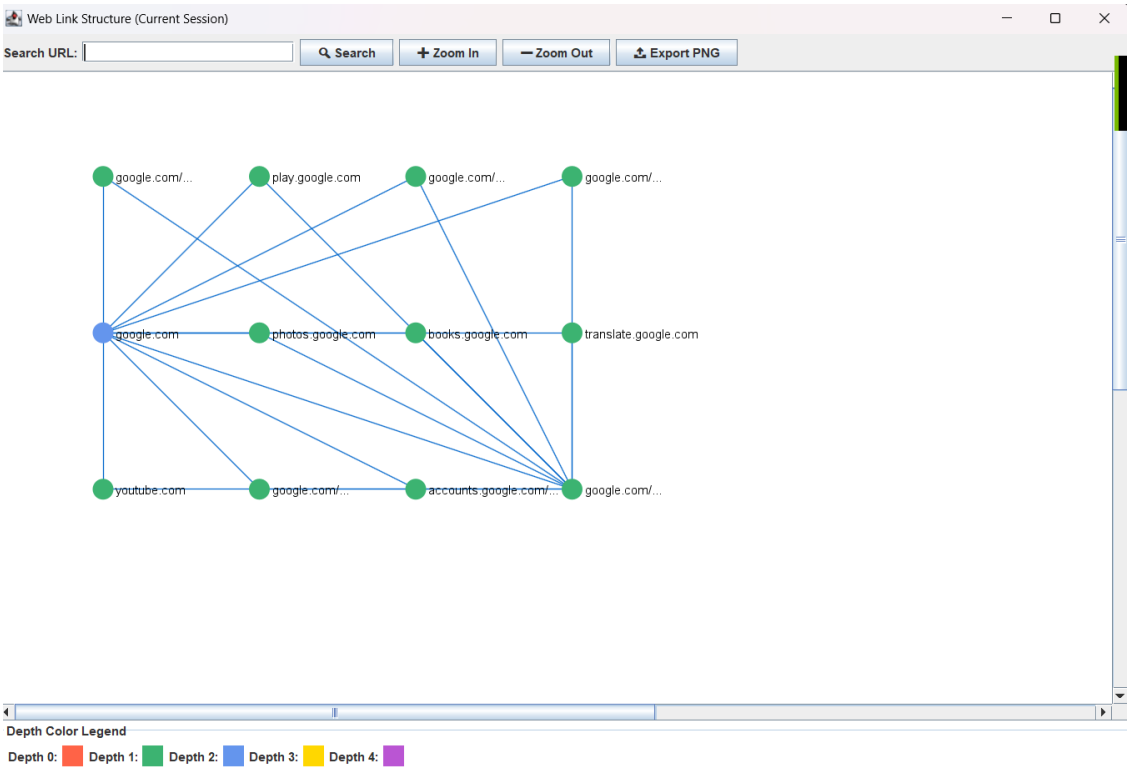


Figure 3: Live Graph Visualization of Web Crawl

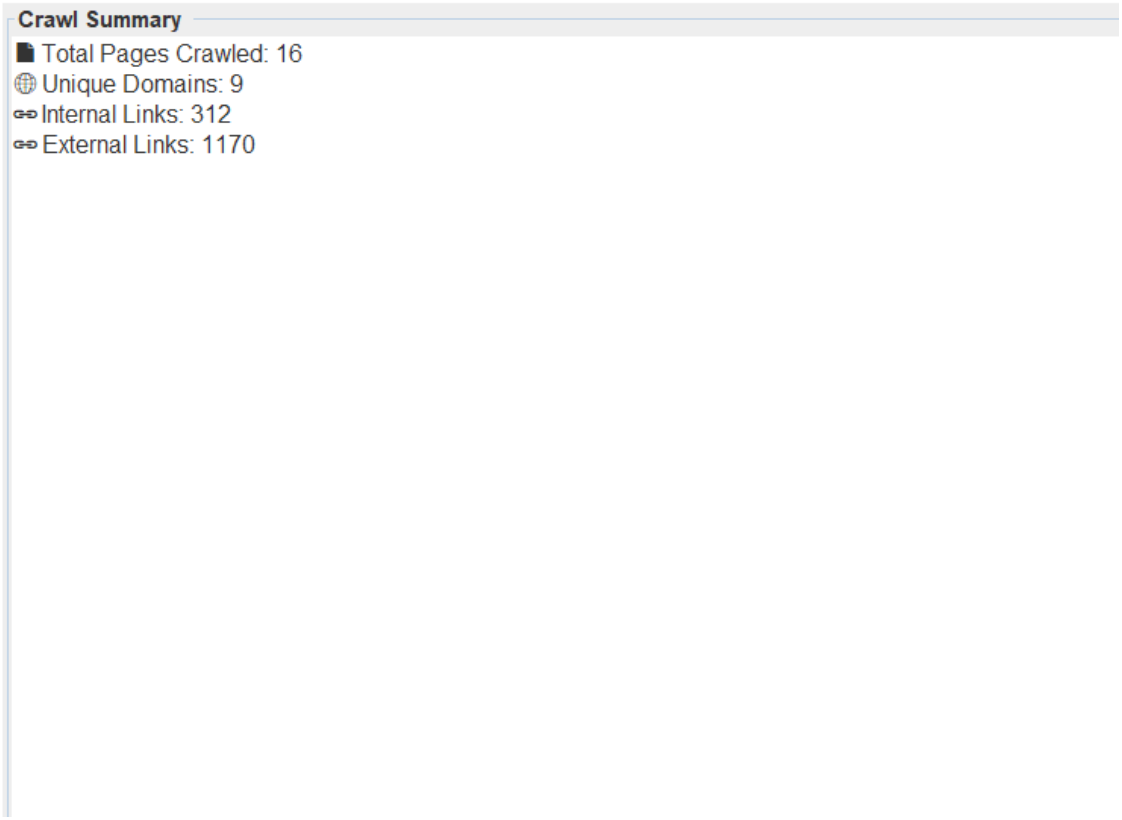


Figure 4: Analytics Panel Showing Page Count and Link Stats

Figure 5: Live Crawl Log Output

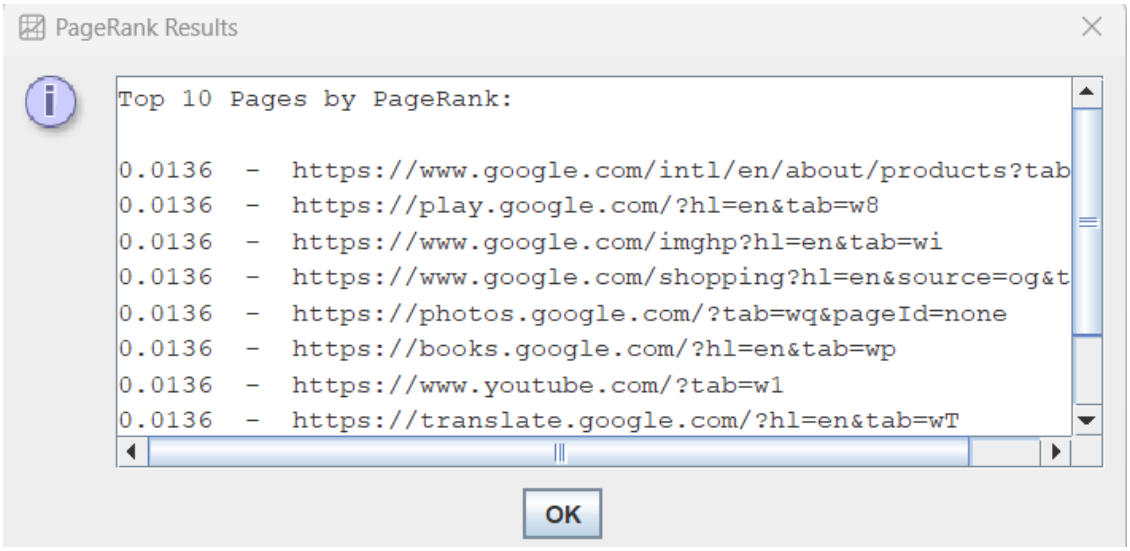


Figure 6: Top 10 Pages of the Crawl by Page Rank

Results and Observations

The crawler was tested in multiple domains including `https://facebook.com`, `https://google.com` and `https://nyu.edu`.

Key observations:

- PageRank scores revealed hubs and authorities in the link structure.
- Internal links outnumbered external links in most test cases.
- Robots.txt compliance effectively restricted certain domains.

Conclusion

This project demonstrates a fully functional and extensible crawling and analytics engine built with native Java tools. Future extensions may include semantic analysis, clustering, or integration with Apache Kafka for distributed crawling.