README

Author

Developed by Zijie Yu on 3/7/2023.

Website Crawler

This program is a web crawler that crawls a news website and collects statistics on the visited pages.

Prerequisites

- JDK 11

Libraries

The crawler uses version 4.4 of the crawler4j library and requires importing the 23 jar files specified in the homework assignment.

Features

- The crawler makes use of various JDK 11 features such as `var`, `Optional`, and stream API.
- The crawler visits pages on the domain foxnews.com and ignores certain file types specified in the EXCLUSIONS pattern.
- The crawler collects various statistics on each visited page such as download size, number of outgoing links, and content type.
- The statistics are collected in the `pageStats` class and saved to three CSV files and one text file at the end of crawling.

Code Overview

- The `Crawler` class extends the `WebCrawler` class and overrides several methods to implement the desired behavior.
- The `shouldVisit` method filters URLs based on the `EXCLUSIONS` pattern and the `newsSiteDomain` variable.

- The `handlePageStatusCode` and `visit` methods collect statistics on visited pages and store them in the `pageStats` object.
- The `pageStats` class defines three `ArrayLists` to store the statistics collected by the `Crawler` class.
- The `UrlStatus`, `UrlInfo`, and `UrlDetail` classes define the structure of the statistics stored in the `ArrayLists`.
- The `Main` class starts the crawling process and saves the collected statistics to files.

Usage

1. Clone the repository and navigate to the project directory.
2. Run the program using `java Main`.
3. The program will crawl the news site and save the collected statistics to files.