Cochrane Library Crawler

Summary

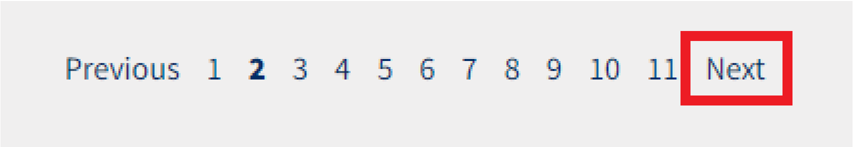
This application visits the Cochrane Library and crawls through the reviews (). It then goes through and grabs all the metadata for each article under that topic. The application will output the compiled data to a text file.

Requirements/Dependencies

Java: v1.8+
Maven: v3.8+
Apache HttpClient: v4.5+
Apache Commons: v3.12+
JSoup: v1.14+

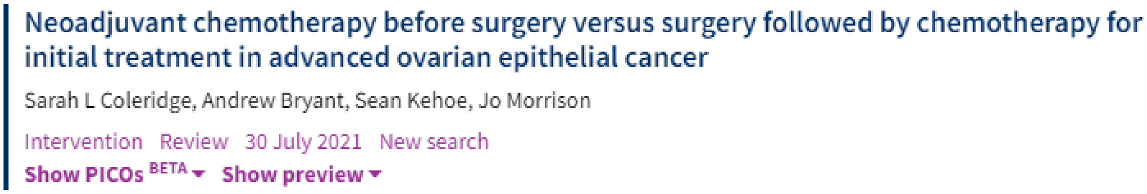
Implementation

Before starting to crawl each article, the program must first compile a list of each topic and their corresponding url. The Apache HttpClient is used to set up a connection pool allowing threads to crawl each page, vastly improving the runtime of the application. In addition, JSoup is used to parse the acquired HTML and extract the metadata required. If a particular thread visits a page that has more than 100 entries then the thread will continue on to the next page of articles assuming there are more pages to be found.



Grab the url from the "Next" link to get the next set of articles.

The "Next" link does not navigate the application to the desired webpage. Thus, the original topic url needs to be utilized alongside one of the query parameters ("cur=#") from the next link which tells the application which page to navigate to. Once all pages have been visited for each topic the threads join allowing the compiled data to then be written to a text file. Below shows an example of one of the articles output in the text file.



Neoadjuvant chemotherapy before surgery versus surgery followed by chemotherapy for initial treatment in advanced ovarian epithelial cancer

Sarah L Coleridge, Andrew Bryant, Sean Kehoe, Jo Morrison

Intervention Review 30 July 2021 New search

```
https://www.cochranelibrary.com//cdsr/doi/10.1002/14651858.CD005343.pub6/full|Cancer|Neoadjuvant chemotherapy before surgery versus surgery followed by chemotherapy for initial treatment in advanced ovarian epithelial cancer|Sarah L Coleridge, Andrew Bryant, Sean Kehoe, Jo Morrison|2021-07-30
```

Article Url | Topic Name | Article Name | Article Author | Article Date

Without multithreading, visiting each page has a drastic performance hit. Therefore, to increase the performance, threads were utilized in addition to loading 100 articles per page by adding the query parameter "&resultPerPage=100".