External libraries/Algorithms used in crawler and indexer:

1- URL Normalization :https://github.com/sentric/url-normalization which does the following to get a normalized URL:

- Converts the host (and scheme) to lower case
- Decodes percent-encoded octets of unreserved characters
- Removes the default port:
    - Removes “www” as the first domain label
    - Removes the "?" when the query is empty
    - Inverts the domain level labels to make urls unique, but we didn’t use them in this form

2- Robots.txt parsing: https://github.com/TrigonicSolutions/jrobotx

Reads robots.txt file and checks if given url is excluded or not by checking the user agent, allowed and dissallowed urls.

3- Porter Stemmer: https://tartarus.org/martin/PorterStemmer/java.txt

Transforms a word into its root form.

4- Jsoup library: https://jsoup.org/

For extracting and manipulating data of HTML pages.