

URL-Archiver

Description

The goal of this project is to deliver a FLOSS-licensed, platform-independent Java-program (called "URL-Archiver") that

- (1) takes as input (the path of) a directory or any Unicode-text- (e.g.: .BIB, .TEX; .HTML; etc.) or .PDF-file (<https://www.baeldung.com/java-curl>);
- (2) scans it for any URLs (<https://stackoverflow.com/questions/4026614/extract-text-from-pdf-files> , <https://librepdf.github.io/OpenPDF> , <https://pdfbox.apache.org> ; see also https://en.wikipedia.org/wiki/List_of_PDF_software);
- (3) extracts all URLs (regular expression ;-) from the text;
- (4) optionally spring-loads all URLs in a Web-browser;
- (5) posts all URLs to <https://archive.ph> ;
- (6) gets the resulting archived URLs;
- (7) outputs a CSV-file of the resulting key-value (URL, archived URL) pairs; and
- (8) optionally inserts the archived URLs into a .BIB-file.

The program code should be minimal, modular, and self-explaining.

The project report should be concise (maximally informative, minimally long).

It must contain this project description as a quotation.

Technologies

Java, LaTeX

Advisor

Dr. Simon Kramer