Data specification

Date: 2022

Author: Andreas Lenander Ægidius

Project: Test methods to explore the evolution of 'stream'-related terms on the Danish Web archive

The following four search queries to be extracted from the Danish Web archive, Netarkivet:

- text:streamingtjeneste* OR text:streaming?tjeneste* OR url_search:streamingtjeneste* OR url_search:streaming?tjeneste*
- text:streamingservice* OR text:streaming?service* OR url_search:streamingservice* OR url_search:streaming?service*
- (text:stream* OR url_search:stream*) AND content_language:en
- (text:stream* OR url_search:stream*) AND content_language:da

All four extractions split in three files with the following data fields:

- Links.csv: id, crawl_date, hash, url, links_domains
- Text.csv: id, crawl_date, hash, url, content
- Metadata.csv: id, arc_harvest, arc_job, crawl_date, wayback_date, hash, domain, url, title, content_type_norm