Finnish Media Scrapers

Mäkelä, Eetu

2021-12-22

Mäkelä , E & Toivanen , P 2021 , ' Finnish Media Scrapers ' , Journal of open source software , vol. 6 , no. 68 . https://doi.org/10.21105/joss.03504

http://hdl.handle.net/10138/354601
https://doi.org/10.21105/joss.03504

The Journal of Open Source Software

Finnish Media Scrapers

Eetu Mäkelä[*1] **and Pihla Toivanen**[†1]

1 University of Helsinki

DOI: 10.21105/joss.03504

Software

- Review
- Repository
- Archive

Editor: Andrew Stewart

Reviewers:

- @sara-shiho
- @GaurangTandon

Submitted: 21 June 2021
Published: 22 December 2021

Summary

Finnish Media Scrapers is a package for extracting articles from Finnish journalistic media websites. Included are scrapers for the four biggest Finnish journalistic media: YLE, Helsingin Sanomat, Iltalehti and Iltasanomat. The scrapers have been designed for researchers needing a local corpus of news article texts matching a specified set of query keywords as well as temporal limitations. As a design principle, these scrapers have been designed to extract the articles in as trustworthy a manner as possible, as required for content-focused research targetting the text of those articles. Further, the process is split into distinct parts that 1) query, 2) fetch, 3) convert to text and 4) post-filter the articles separately. Each of these steps also records its output as separate files. This way, the tools can be used in a versatile manner. Further, a good record is maintained of the querying and filtering process for reproducibility as well as error analysis.

Statement of need

There is an increasing need for user-friendly computational tools in the humanities and social sciences. For example, a common workflow in media research is to collect a large amount of data and combine quantitative and qualitative methods in the analysis phase (Koivunen et al. (2021), Weber & Monge (2011)). This package responds to the research needs by providing easy-to-use tools for scraping Finnish media articles and extracting the article texts from the scraped HTML files. At the same time, the functionality has also been packaged as a Python module for the benefit of more computationally-savvy users.

The scripts were originally developed for a data journalism article (Suomen Kuvalehti et al. (2021)) analyzing how Finnish members of parliament were represented in the media in 2020. Further developing and packaging the scripts into a reusable package was based on an expressed interest from the Finnish computational science community. Since initial beta release a couple of months ago, the package is now known to be already used in at least two research projects targeting Finnish media analysis.

Related work

For a more general library for crawling media articles, have a look at newspaper3k as well as news-please, which has been built on top of it. Do note however that at the time of writing this, it is unclear whether newspaper3k is being maintained any more. More importantly for content research purposes, note that 1) newspaper3k does not handle the Finnish news sources targeted by this crawler very well and 2) it is based more on a best-effort principle (suitable for extracting masses of data for e.g. NLP training) as opposed to completeness and

*corresponding author
†co-first author

verisimilitude (required for trustworthy content-focused research targetting a particular set of news). Thus, given an article URL, newspaper3k will happily try to return something from it, but not guarantee completeness. This crawler on the other hand has been designed to be conservative, and to complain loudly through logging whenever it encounters problems that may hinder extracting the actual text of the article, such as article layouts that haven't been yet handled and verified to extract correctly.

General workflow

The general workflow for using the scrapers is as follows: 1. The scrapers support specifying a keyword as well as a timespan for extraction, and output a CSV of all matching articles with links. 2. A second set of scripts then allows downloading the matched articles in HTML format. 3. Third, there are further scripts for extracting plain text versions of the article texts out of the HTML. 4. Finally, a script exists to post-filter the resulting plain texts again with keywords.

Important to know when applying the workflow is that due to the fact that all the sources use some kind of stemming for their search, they can often return also spurious hits. Further, if searching for multiple words, the engines often perform a search for either word instead of the complete phrase. The post-filtering script above exists to counteract this by allowing the refiltering of the results more rigorously and uniformly locally.

At the same time and equally importantly, the stemming for a particular media may not cover e.g. all inflectional forms of words. Thus, it often makes sense to query for at least all common inflected variants and merge the results. For a complete worked up example of this kind of use, see the members_of_parliament folder, which demonstrates how one can collect and count how many articles in each media mention the members of the Finnish Parliament.

Acknowledgements

We acknowledge contributions from the Suomen Kuvalehti team (Samuel Nyroos, Salla Vuorikoski and Leena Sharma) during the testing phase of the scrapers.

References

Koivunen, A., Kanner, A., Janicki, M., Harju, A., Hokkanen, J., & Mäkelä, E. (2021). Emotive, evaluative, epistemic: A linguistic analysis of affectivity in news journalism. *Journalism*, *22*(5), 1190–1206. https://doi.org/10.1177/1464884920985724

Suomen Kuvalehti, Mäkelä, E., & Toivanen, P. (2021). Vuosi valokeilassa: Kuka sai medialta huomiota? Kuka jäi varjoon? Suomen kuvalehti selvitti tutkijoiden kanssa, miten kansanedustajat näkyivät neljässä suuressa uutismediassa vuonna 2020. *Suomen Kuvalehti*, 24–33.

Weber, M. S., & Monge, P. (2011). The flow of digital news in a network of sources, authorities, and hubs. *Journal of Communication*, *61*(6), 1062–1081. https://doi.org/10.1111/j.1460-2466.2011.01596.x