The Journal of Open Source Software

Tag-Pag: A Dedicated Tool for Systematic Web Page Annotations

Anton Pogrebnjak[1], **Julian Schelb**[1], **Andreas Spitz**[1], **Celina Kacperski**[2], **and Roberto Ulloa**[2]

1 Department of Computer Science, University of Konstanz, Germany **2** Cluster of Excellence "The Politics of Inequalities", University of Konstanz, Germany

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor:

Submitted: 24 February 2025
Published: unpublished

Summary

Tag-Pag is an application designed to simplify the categorization of web pages, a task increasingly common for researchers who scrape web pages to analyze individuals' browsing patterns or train machine learning classifiers. Unlike existing tools that focus on annotating sections of text, Tag-Pag systematizes page-level annotations, allowing users to determine whether an entire document relates to one or multiple predefined topics.

Tag-Pag offers an intuitive interface to configure the input web pages and annotation labels. It integrates libraries to extract content from the HTML and URL indicators to aid the annotation process. It provides direct access to both scraped and live versions of the web page. Our tool is designed to expedite the annotation process with features like quick navigation, label assignment, and export functionality, making it a versatile and efficient tool for various research applications. Tag-Pag is available at https://github.com/Pantonius/TagPag.

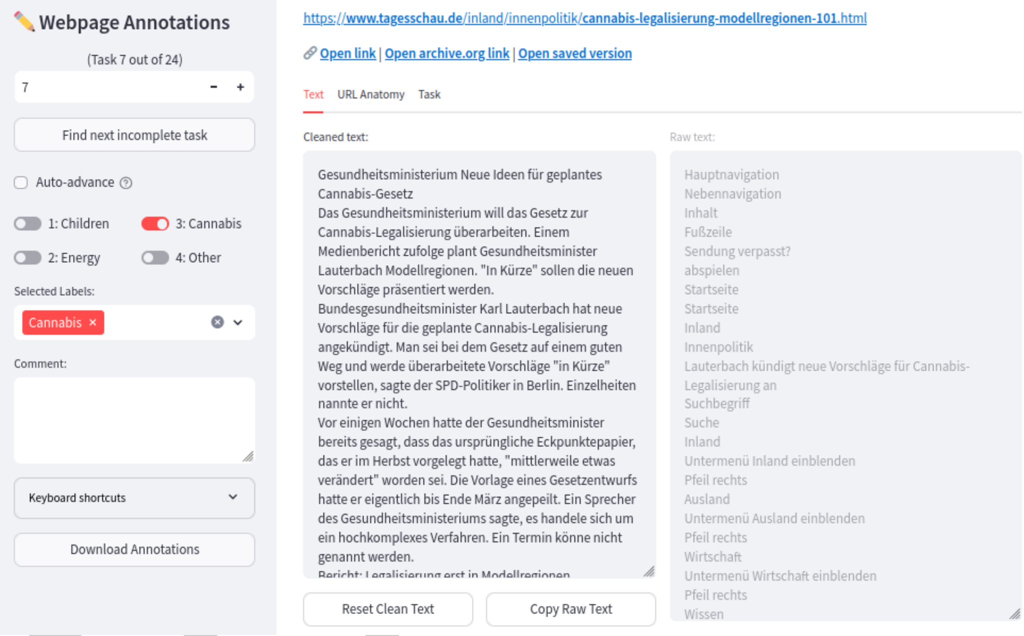
Statement of need

The annotation of web data is increasingly common across multiple disciplines, serving purposes such as analyzing online behavioral patterns (Guess, 2021; Stier et al., 2020; Ulloa & Kacperski, 2023; Wojcieszak et al., 2024), auditing the performance of online platforms (Kacperski et al., 2024; Makhortykh et al., 2020), or training and evaluating machine learning classifiers (Schelb et al., 2024). As the need for processing web data grows, researchers have turned to systematic and efficient methodologies that span the entire process—from data collection to categorization—to ensure robust results.

Online behavioral researchers have recently investigated limitations of web scraping, such as reliance on external environments that differ from individuals' computers (Ulloa, Mangold, et al., 2024) and changes due to time delays in scraping (Dahlke et al., 2023; Ulloa, Mangold, et al., 2024). To improve reliability and validity, researchers attempt to scrape data from web pages as close to the visit time of participants as possible and uniformly distribute the delay between the visit and the web page collection. Such limitations have, more recently, been addressed by developing new web tools that collect content directly from an individual's browser (e.g., Adam et al., 2024; GESIS Panel Team, 2025). Other researchers have gone to the effort of standardizing web data collections for algorithm auditings to avoid, for example, noise stemming from search engine personalization (Ulloa, Makhortykh, et al., 2024).

These academic efforts illustrate the importance placed on collecting high-quality data for annotation purposes; so far, however, there has been a lack of tools to facilitate the manual annotation process. This has led researchers to instead use inefficient methods such as relying on the URL, rarely accessing the systematically scraped content, or manually visiting (and revisiting) the related web page at different times. As a result, promising lines of inquiry — especially those requiring large-scale and consistent annotations — might be left unexplored.

Existing tools often focus on annotating specific sections within a text (Huang, 2016; Meister, 2023; Rampin & Rampin, 2021), where the user selects a portion of text and assigns a label or establishes connections between parts of speech (Strippel et al., 2022). These tools fall short when the goal is to annotate entire pages to determine, for example, if the content corresponds to very specific topics (Schelb et al., 2024), misinformation (Urman et al., 2022), or, more broadly, political content (Guess, 2021; Stier et al., 2020), news articles (Ulloa & Kacperski, 2023) or pages that restrict access such as logins (Dahlke et al., 2023). Tag-Pag [1] addresses this gap by allowing broad-level annotations of entire web pages.



Cleaned text:

Gesundheitsministerium Neue Ideen für geplantes Cannabis-Gesetz
Das Gesundheitsministerium will das Gesetz zur Cannabis-Legalisierung überarbeiten. Einem Medienbericht zufolge plant Gesundheitsminister Lauterbach Modellregionen. "In Kürze" sollen die neuen Vorschläge präsentiert werden.
Bundesgesundheitsminister Karl Lauterbach hat neue Vorschläge für die geplante Cannabis-Legalisierung angekündigt. Man sei bei dem Gesetz auf einem guten Weg und werde überarbeitete Vorschläge "in Kürze" vorstellen, sagte der SPD-Politiker in Berlin. Einzelheiten nannte er nicht.
Vor einigen Wochen hatte der Gesundheitsminister bereits gesagt, dass das ursprüngliche Eckpunktepapier, das er im Herbst vorgelegt hatte, "mittlerweile etwas verändert" worden sei. Die Vorlage eines Gesetzentwurfs hatte er eigentlich bis Ende März angepeilt. Ein Sprecher des Gesundheitsministeriums sagte, es handele sich um ein hochkomplexes Verfahren. Ein Termin könne nicht genannt werden.
Bericht: Legalisierung erst in Modellregionen

Raw text:

Hauptnavigation
Nebennavigation
Inhalt
Fußzeile
Sendung verpasst?
abspielen
Startseite
Startseite
Inland
Innenpolitik
Lauterbach kündigt neue Vorschläge für Cannabis-Legalisierung an
Suchbegriff
Suche
Inland
Untermenü Inland einblenden
Pfeil rechts
Ausland
Untermenü Ausland einblenden
Pfeil rechts
Wirtschaft
Untermenü Wirtschaft einblenden
Pfeil rechts
Wissen

Figure 1: Annotation interface for web pages classification tasks. The interface of Tag-Pag is divided into a left sidebar and a main panel. The sidebar provides task navigation, annotation selection labels, and additional tools such as key shortcut references and annotation downloads. The main panel displays web page content in multiple views, including a cleaned text version, raw text, and URL decomposition. Annotators can label data using predefined categories and add comments. Tag-Pag automatically highlights relevant sections of the URL (at the top).

Tag-Pag uses libraries to extract two versions of the content from the HTML (see Figure 1): (1) cleaned text (Barbaresi, 2021), with removed boilerplate such as menus and advertisements, and (2) raw text (Artem Golubin, 2023), with only removed HTML elements. The tool also parses the URLs themselves, which often contain relevant information about the page's content, adding another layer of contextual data for annotations. For a comprehensive overview, users can open the scraped HTML, the live web page, or the latest version stored in the Wayback Machine. For researchers creating or refining training datasets to improve machine learning models, Tag-Pag allows easy text editing to retain only the relevant parts for classifiers, e.g., manually removing boilerplate to further filter the cleaned text.

Additionally, Tag-Pag includes functionality designed to speed up the annotation process: key bindings for interface actions for rapid label assignment, automatic transition between pages for single-label annotations are supported, and a feature to locate unannotated pages is included. Comments and annotations can be exported to CSV, ensuring compatibility with further steps of the analysis pipeline.

[1] Tag-Pag. https://github.com/Pantonius/TagPag

The tool also supports multiple annotators, with the functionality to hide one another's annotations and randomize the tasks' order to avoid priming effects (Mathur et al., 2017; Shen et al., 2019).

By integrating these features, Tag-Pag offers a systematic, efficient, and user-friendly approach to web page annotation, addressing the needs of researchers across various disciplines.

Acknowledgments

This research was funded by the Deutsche Forschungsgemeinschaft (DFG – German Research Foundation) under Germany's Excellence Strategy – EXC- 2035/1 – 390681379. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Adam, S., Makhortykh, M., Maier, M., Aigenseer, V., Urman, A., Lopez, T. G., Christner, C., León, E. de, & Ulloa, R. (2024). *Improving the quality of individual-level online information tracking: Challenges of existing approaches and introduction of a new content- and long-tail sensitive academic solution*. arXiv. https://doi.org/10.48550/arXiv.2403.02931

Artem Golubin. (2023). *Selectolax: Fast HTML5 parser with CSS selectors*. https://github.com/rushter/selectolax

Barbaresi, A. (2021). Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. In H. Ji, J. C. Park, & R. Xia (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations* (pp. 122–131). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-demo.15

Dahlke, R., Kumar, D., Durumeric, Z., & Hancock, J. T. (2023). Quantifying the Systematic Bias in the Accessibility and Inaccessibility of Web Scraping Content From URL-Logged Web-Browsing Digital Trace Data. *Social Science Computer Review*, 08944393231218214. https://doi.org/10.1177/08944393231218214

GESIS Panel Team. (2025). *GESIS Panel.dbd - Pre-ReleaseGESIS Panel.dbd - Pre-Release*. GESIS. https://doi.org/10.4232/1.14467

Guess, A. M. (2021). (Almost) Everything in Moderation: New Evidence on Americans' Online Media Diets. *American Journal of Political Science*, *65*(4), 1007–1022. https://doi.org/10.1111/ajps.12589

Huang, R. (2016). *RQDA: R-based Qualitative Data Analysis*. http://rqda.r-forge.r-project.org/

Kacperski, C., Bielig, M., Makhortykh, M., Sydorova, M., & Ulloa, R. (2024). Examining bias perpetuation in academic search engines: An algorithm audit of Google and Semantic Scholar. *First Monday*, *29*(11). https://doi.org/10.5210/fm.v29i11.13730

Makhortykh, M., Urman, A., & Ulloa, R. (2020). How search engines disseminate information about COVID-19 and why they should do better. *Harvard Kennedy School Misinformation Review*, *1*(COVID-19 and Misinformation). https://doi.org/10.37016/mr-2020-017

Mathur, N., Baldwin, T., & Cohn, T. (2017). Sequence Effects in Crowdsourced Annotations. In M. Palmer, R. Hwa, & S. Riedel (Eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2860–2865). Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1306

Meister, J. C. (2023). From TACT to CATMA, or, a mindful approach to text annotation and analysis. In J. Nyhan, G. Rockwell, S. Sinclair, & A. Ortolja-Baird (Eds.), *On Making in the Digital Humanities: The scholarship of digital humanities development in honour of John Bradley* (pp. 213–250). UCL Press. ISBN: 978-1-80008-420-9

Rampin, R., & Rampin, V. (2021). Taguette: Open-source qualitative data analysis. *Journal of Open Source Software*, *6*(68), 3522. https://doi.org/10.21105/joss.03522

Schelb, J., Ulloa, R., & Spitz, A. (2024). Assessing In-context Learning and Fine-tuning for Topic Classification of German Web Data. In X. Fu & E. Fleisig (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)* (pp. 144–158). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-srw.22

Shen, J. H., Lapedriza, A., & Picard, R. W. (2019). Unintentional affective priming during labeling may bias labels. *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 587–593. https://doi.org/10.1109/ACII.2019.8925466

Stier, S., Kirkizh, N., Froio, C., & Schroeder, R. (2020). Populist Attitudes and Selective Exposure to Online News: A Cross-Country Analysis Combining Web Tracking and Surveys. *The International Journal of Press/Politics*, *25*(3), 426–446. https://doi.org/10.1177/1940161220907018

Strippel, C., Laugwitz, L., Paasch-Colberg, S., Esau, K., & Heft, A. (2022). BRAT Rapid Annotation Tool. *M&K Medien & Kommunikationswissenschaft*, *70*(4), 446–461. https://doi.org/10.5771/1615-634X-2022-4-446

Ulloa, R., & Kacperski, C. S. (2023). Search engine effects on news consumption: Ranking and representativeness outweigh familiarity in news selection. *New Media & Society*, 14614448231154926. https://doi.org/10.1177/14614448231154926

Ulloa, R., Makhortykh, M., & Urman, A. (2024). Scaling up search engine audits: Practical insights for algorithm auditing. *Journal of Information Science*, *50*(2), 404–419. https://doi.org/10.1177/01655515221093029

Ulloa, R., Mangold, F., Schmidt, F., Gilsbach, J., & Stier, S. (2024). *Beyond time delays: How web scraping distorts measures of online news consumption*. arXiv. https://doi.org/10.48550/arXiv.2412.00479

Urman, A., Makhortykh, M., Ulloa, R., & Kulshrestha, J. (2022). Where the earth is flat and 9/11 is an inside job: A comparative algorithm audit of conspiratorial information in web search results. *Telematics and Informatics*, *72*, 101860. https://doi.org/10.1016/j.tele.2022.101860

Wojcieszak, M., Menchen-Trevino, E., Clemm von Hohenberg, B., Leeuw, S. de, Gonçalves, J., Davidson, S., & Gonçalves, A. (2024). Non-News Websites Expose People to More Political Content Than News Websites: Evidence from Browsing Data in Three Countries. *Political Communication*, *41*(1), 129–151. https://doi.org/10.1080/10584609.2023.2238641