The Journal of Open Source Software

Inscriptis - A Python-based HTML to text conversion library optimized for knowledge extraction from the Web

Albert Weichselbraun[1]

1 University of Applied Sciences of the Grisons, Chur, Switzerland

DOI: DOIunavailable

Software

- Review
- Repository
- Archive

Editor: Pending Editor

Reviewers:

- @Pending Reviewers

Submitted: N/A
Published: N/A

Summary

`Inscriptis` provides a library, command line client and Web service for converting HTML content to plain text. In contrast to existing software packages such as HTML2text, jusText and Lynx, it has been tailored towards knowledge extraction pipelines by

1. providing a layout-aware conversion of textual output that more closely resembles the rendering obtained from standard Web browsers. `Inscriptis` excels in terms of conversion quality, since it correctly converts complex HTML constructs such as nested tables and also interprets a subset of HTML (e.g., `align`, `valign`) and CSS (e.g., `display`, `white-space`, `margin-top`, `vertical-algin`, etc.) attributes that determine the text alignment.
2. supporting annotation rules, i.e., user-provided mappings that allow for annotating the extracted text based on structural and semantic information encoded in HTML tags and attributes used for controlling structure and layout in the original HTML document.

These unique features ensure that downstream Knowledge Extraction components can operate on accurate text representations without drawing upon a heavyweight solution such as Selenium which requires interaction with a full-fledged Web browser. In addition, its optional annotation support enables downstream components to use information on the structure of the original HTML document.

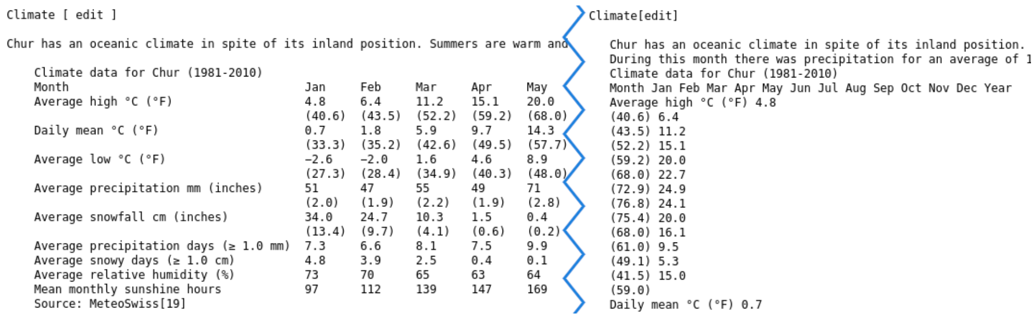
Statement of need

Research in a growing number of scientific disciplines relies upon Web content. Li et al. (2014), for instance, studied the impact of company-specific News coverage on stock prices, in medicine and pharmacovigilance social media listening plays an important role in gathering insights into patient needs and the monitoring of adverse drug effects (Convertino et al., 2018), and communication sciences draw upon media coverage to obtain information on the perception and framing of issues as well as on the rise and fall of topics within News and social media (Scharl et al., 2017; Weichselbraun et al., 2021).

Computer science focuses on analyzing content by applying knowledge extraction techniques such as entity recognition (Fu et al., 2021) to automatically identify entities (e.g., persons, organizations, locations, products, etc.) within text documents, entity linking (Ding et al., 2021) to link these entities to knowledge bases (e.g., Wikidata and DBPedia), and sentiment analysis to automatically assess sentiment polarity (i.e., positive versus negative coverage) and emotions expressed towards these entities (Wang et al., 2020).

Most knowledge extraction methods operate on text and, therefore, require an accurate conversion of HTML content which also preserves the spatial alignment between text elements. This is particularly true for methods drawing upon algorithms which directly or indirectly leverage information on the proximity between terms, such as word embeddings (Mikolov et al., 2013; Pennington et al., 2014) and language models (Reis et al., 2021), sentiment analysis which often also considers the distance between target and sentiment terms, and automatic keyword and phrase extraction techniques.

Despite this need from within the research community, many standard HTML to text conversion techniques are not layout aware, yielding text representations that fail to preserve the spatial properties of text snippets, as illustrated below.



```
Climate [ edit ]

Chur has an oceanic climate in spite of its inland position. Summers are warm and

    Climate data for Chur (1981-2010)
    Month                                  Jan     Feb     Mar     Apr     May
    Average high °C (°F)                   4.8     6.4     11.2    15.1    20.0
                                           (40.6)  (43.5)  (52.2)  (59.2)  (68.0)
    Daily mean °C (°F)                     0.7     1.8     5.9     9.7     14.3
                                           (33.3)  (35.2)  (42.6)  (49.5)  (57.7)
    Average low °C (°F)                    -2.6    -2.0    1.6     4.6     8.9
                                           (27.3)  (28.4)  (34.9)  (40.3)  (48.0)
    Average precipitation mm (inches)      51      47      55      49      71
                                           (2.0)   (1.9)   (2.2)   (1.9)   (2.8)
    Average snowfall cm (inches)           34.0    24.7    10.3    1.5     0.4
                                           (13.4)  (9.7)   (4.1)   (0.6)   (0.2)
    Average precipitation days (≥ 1.0 mm)  7.3     6.6     8.1     7.5     9.9
    Average snowy days (≥ 1.0 cm)          4.8     3.9     2.5     0.4     0.1
    Average relative humidity (%)          73      70      65      63      64
    Mean monthly sunshine hours            97      112     139     147     169
    Source: MeteoSwiss[19]
```

```
Climate[edit]

   Chur has an oceanic climate in spite of its inland position.
   During this month there was precipitation for an average of 1
   Climate data for Chur (1981-2010)
   Month Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Year
   Average high °C (°F) 4.8
   (40.6) 6.4
   (43.5) 11.2
   (52.2) 15.1
   (59.2) 20.0
   (68.0) 22.7
   (72.9) 24.9
   (76.8) 24.1
   (75.4) 20.0
   (68.0) 16.1
   (61.0) 9.5
   (49.1) 5.3
   (41.5) 15.0
   (59.0)
   Daily mean °C (°F) 0.7
```

Figure 1: Text representation of a table from DBpedia computed by `Inscriptis` (left) and lynx (right). Lynx fails to correctly interpret the cascaded table and, therefore, does not properly align the temperature values.

`Inscriptis` is not only able to correctly render such pages but also offers the option to preserve parts of the original HTML document's semantics (e.g., information on headings, emphasised text, tables, etc.) by complementing the extracted text with annotations obtained from the document. Figure 2 provides an example of annotations extracted from a Wikipedia page. These annotations might be useful for

- aiding downstream knowledge extraction components with additional information that may be leveraged to improve their respective performance. Text summarization techniques, for instance, can put a stronger emphasis on paragraphs that contain bold and italic text, and sentiment analysis may consider this information in addition to textual clues such as uppercase text.
- assisting manual document annotation processes (e.g., for qualitative analysis or gold standard creation). `Inscriptis` supports multiple export formats such as XML, annotated HTML and the JSONL format that is used by the open source annotation tool doccano[1]. Support for further annotation formats can be easily added by implementing custom annotation processors.
- enabling the use of `Inscriptis` for tasks such as content extraction (i.e., extract task-specific relevant content from a Web page) which rely on information on the HTML document's structure.

[1] Please note that doccano currently does not support overlapping annotations and, therefore, cannot import files containing overlapping annotations.

The Journal of Open Source Software

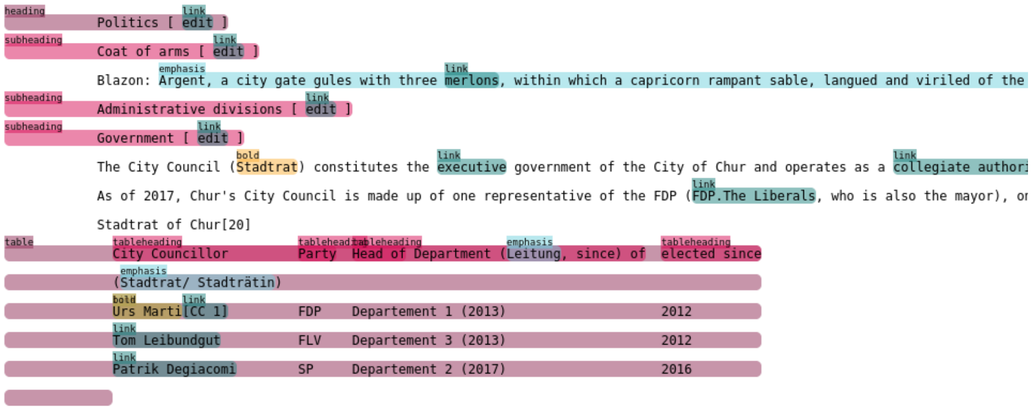


Figure 2: Annotations extracted from the DBpedia entry for Chur using the `--postprocessor html` command line option.

In conclusion, `Inscriptis` provides knowledge extraction components with high quality conversions of HTML documents. Since its first public release in March 2016, `Inscriptis` has been downloaded over 121,000 times from the Python Package Index (PyPI)[2], has proven its capabilities in national and European research projects and has been integrated into commercial products such as the webLyzard Web Intelligence and Visual Analytics Platform.

Mentions

The following research projects use `Inscriptis` within their knowledge extraction pipelines:

- CareerCoach: Automatic Knowledge Extraction and Recommender Systems for Personalized Re- and Upskilling suggestions funded by Innosuisse.
- EPOCH project funded by the Austrian Federal Ministry for Climate Action, Environment, Energy, Mobility and Technology (BMK) via the ICT of the Future Program.
- MedMon: Monitoring of Internet Resources for Pharamceutical Research and Development funded by Innosuisse.
- ReTV project funded by the European Union's Horizon 2020 Research and Innovation Programme.

Acknowledgements

Work on `Inscriptis` has been conducted within the MedMon and CareerCoach projects funded by Innosuisse.

References

Convertino, I., Ferraro, S., Blandizzi, C., & Tuccori, M. (2018). The usefulness of listening social media for pharmacovigilance purposes: A systematic review. *Expert Opinion on Drug Safety*, *17*(11), 1081–1093. https://doi.org/10.1080/14740338.2018.1531847

Ding, W., Chaudhri, V. K., Chittar, N., & Konakanchi, K. (2021). JEL: Applying End-to-End Neural Entity Linking in JPMorgan Chase. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(17), 15301–15308. https://ojs.aaai.org/index.php/AAAI/article/view/17796

[2] Source: https://pepy.tech/project/inscriptis

Fu, J., Huang, X., & Liu, P. (2021). SpanNER: Named Entity Re-/Recognition as Span Prediction. *arXiv:2106.00641 [Cs]*. http://arxiv.org/abs/2106.00641

Li, Q., Wang, T., Li, P., Liu, L., Gong, Q., & Chen, Y. (2014). The effect of news and public mood on stock movements. *Information Sciences*, *278*, 826–840. https://doi.org/10.1016/j.ins.2014.03.096

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a Meeting Held December 5-8, 2013, Lake Tahoe, Nevada, United States*, 3111–3119. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Reis, E. S. D., Costa, C. A. D., Silveira, D. E. D., Bavaresco, R. S., Righi, R. D. R., Barbosa, J. L. V., Antunes, R. S., Gomes, M. M., & Federizzi, G. (2021). Transformers aftermath: Current research and rising trends. *Communications of the ACM*, *64*(4), 154–163. https://doi.org/10.1145/3430937

Scharl, A., Herring, D., Rafelsberger, W., Hubmann-Haidvogel, A., Kamolov, R., Fischl, D., Föls, M., & Weichselbraun, A. (2017). Semantic Systems and Visual Tools to Support Environmental Communication. *IEEE Systems Journal*, *11*(2), 762–771. https://doi.org/10.1109/JSYST.2015.2466439

Wang, Z., Ho, S.-B., & Cambria, E. (2020). A review of emotion sensing: Categorization models and algorithms. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-019-08328-z

Weichselbraun, A., Kuntschik, P., Fancolino, V., Saner, M., & Wyss, V. (2021). Adapting Data-Driven Research to the Fields of Social Sciences and the Humanities. *Future Internet*, *13*(3). https://doi.org/10.3390/fi13030059