Measuring the Readability of Web Documents

How important is the readability of online content? And how can it be measured effectively?

Background

• Readability is what makes some texts easier to read than others.[1] it is by definition how easily written materials can be read and understood and itdepends on several factors including the average length of sentences, thenumber of new words contained, and the grammatical complexity of thelanguage used in a passage.[2]

• Measuring the readability of web content is important for enhancing user experience, improving accessibility, optimizing search engine results, aligning with the target audience.[3]

Objectives

• Identify appropriate readability formulas for English and German and optianly additional languages.

• Develop software that extracts the main component of a html document and analyse its readability.

• Compute readability measures for individual documents, and first ten search engine results.

• Use test data to evaluate main component extraction and differences in readability scores using the different formulas.

Methods

• Research pervious work to Identify appropriate readability formulas for each supported langauge.

• Design a comprehensive software architecture for content extraction and readability analysis.

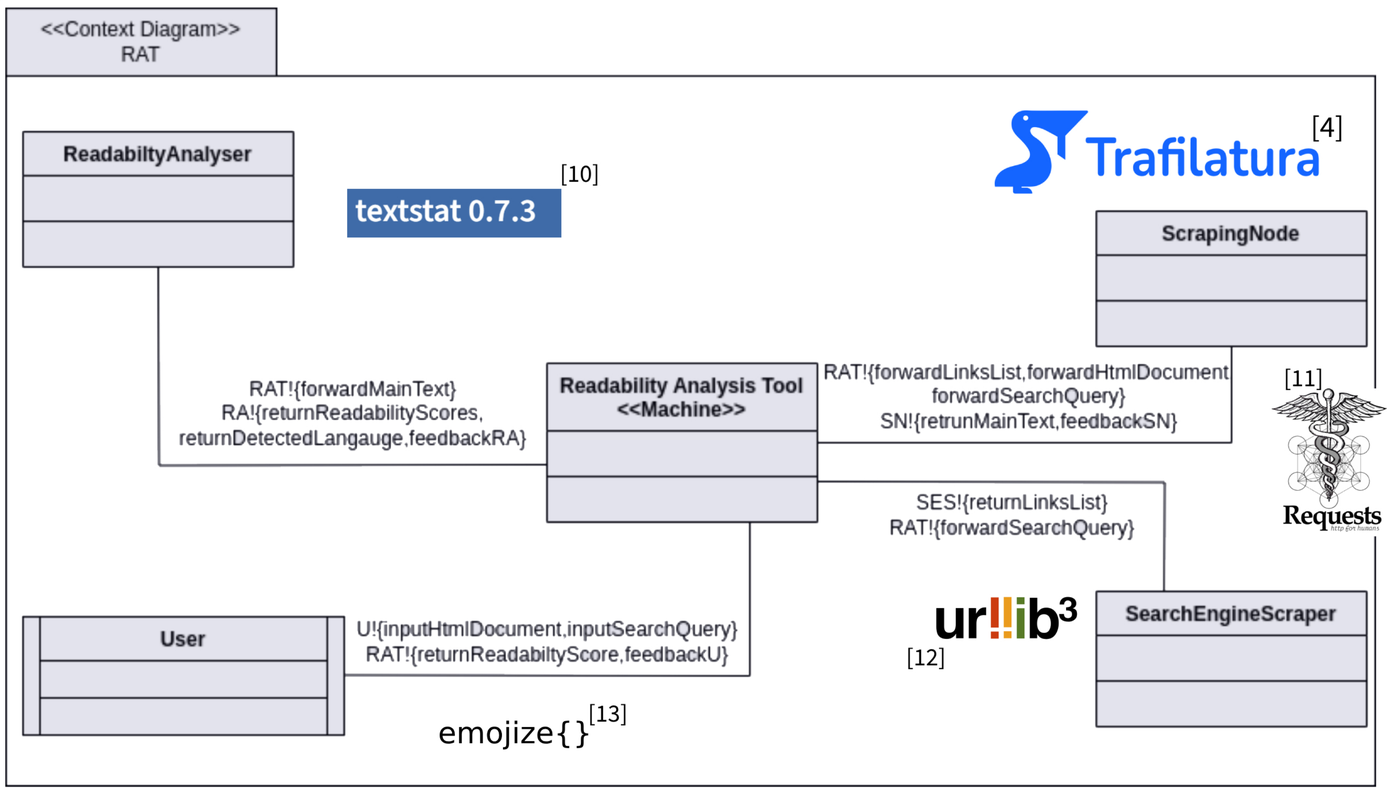
• Utilize the available Python tools, libraries and packages to implement the proposed design.

• Empirical evaluation of implemented software by utilizing available data.

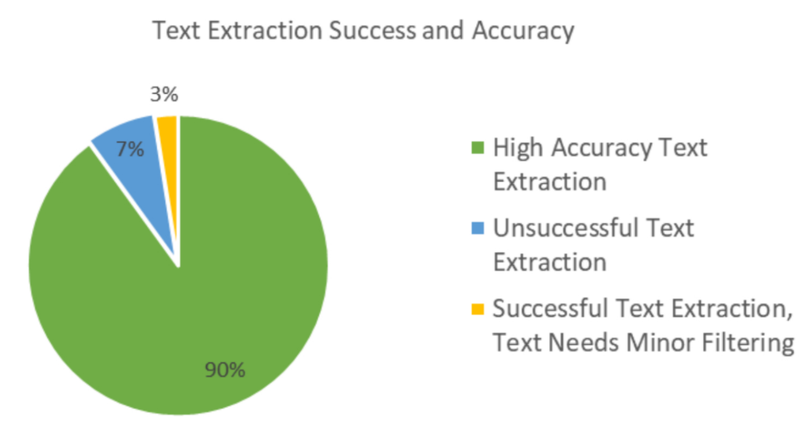
✓ Developing a simple framework for ranking the quality of extracted text.

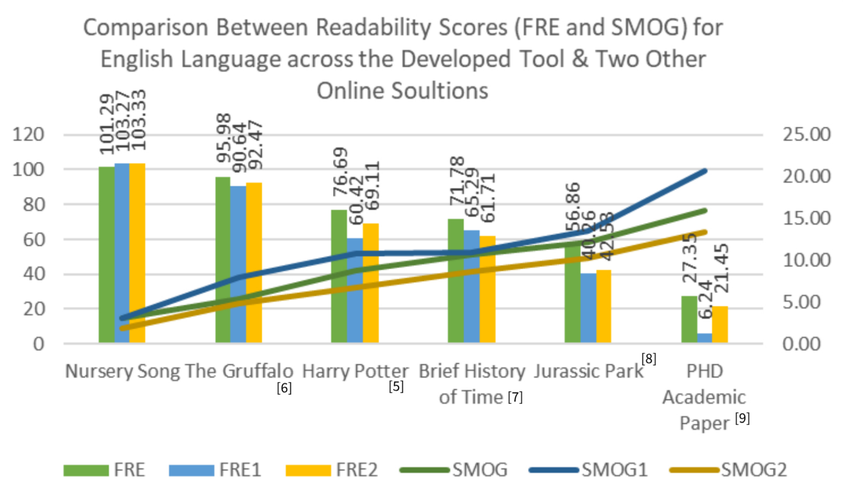
✓ Comparison of the different readability scores with the pre-calculated data.

Software Architecture



Results





Discssion/Conclusion

- A software was designed and developed to measure the readability of web documents by extracting the main text and then analyzing it with selected readability formulas.
- The developed tool performs very well in tests for English, German, and Arabic, and there is the possibility to support other languages with the same Python packages.
- The software also has some limitations, as some websites use mechanisms to prevent web scraping, but there are also some ways around this, such as using DNS.
- The tool provides a quick solution to measure the readability of any web document at a glance and has the potential to be fully or partially integrated into the RAT project.

References

[1] W. H. DuBay, “The principles of readability.,” Online Submission, 2004. [2] M. Zamanian and P. Heydari, “Readability of texts: State of the art.,”Theory & Practice in LanguageStudies, vol. 2, no. 1, 2012. [3] J. T. Richards and V. L. Hanson, “Web accessibility: a broader view,” in Proceedings of the 13th international conference on World Wide Web,pp. 72–79, 2004. [4] A. Barbaresi, “Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction,” in Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conferenceon Natural Language Processing : System Demonstrations, pp.122–131,Association for Computational Linguistics, 2021.[5] J. K. Rowling, Harry Potter and the Philosopher’s Stone, vol. 1. London:Bloomsbury Publishing, 1 ed., June 1997.[6] J. Donaldson, The Gruffalo. London, United Kingdom: Macmillan, 1999.[7]S. A. Hawking, Brief History of Time. New York, Bantam Books, 2017.[8]Crichton, Michael. Jurassic Park. Random House, 2012.[9]B. Bilecen, “How social support works among the best and the brightest: Evidence from international phd students in germany,” TransnationalSocial Review, vol. 2, no. 2, pp. 139–155, 2012.31,[10] A. Ward, “Textstat: A simple python library to calculate readability, com- plexity, and grade level of a text.” https://github.com/shivam5992/ textstat, Year of access, 2023. Accessed on: Date, August 1, 2023. 31.[11]K. Reitz, “Requests-html: Html parsing for humans (writing python3).” https://github.com/psf/requests-html, 2019.[12] “urllib python module.” https://docs.python.org/3/library/urllib.html, Year of access, 2023. Accessed on: 10.07.2023.[13] T. K. (carpedm20), “emoji - emoji for python.” https://github.com/carpedm20/emoji/, 2023. Accessed on: 15.07.2023.31

Contact

Mohamed Elnaggar

Duisburg-Essen University
Faculty of Engineering, Department of Computer Science and Applied Cognitive Science
mohamed.elnaggar@stud.uni-due.de

Supervised by

Prof. Dr. Dirk Lewandowski

Duisburg-Essen University
Faculty of Engineering, Department of Computer Science and Applied Cognitive Science
dirk.lewandowski@uni-due.org

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

RAT
RESULT ASSESSMENT TOOL
SEARCH STUDIES
HAW HAMBURG