6
Conclusion

We presented an approach that makes use of some visual features of webpages while being approximately an order of magnitude faster than a full rendering approach. This required some engineering effort, which is worthwhile given the obtained results.

We applied it to the task of segmenting a news webpage, obtaining satisfactory results with a small number of features and regardless of the language of the story and website where it has been published. In our experiments, we achieve at least 91% of F1 for title detection, 77% of F1 for date detection and 88% of F1 for body detection on the RCD4 corpus. On the Cardoso3000, our numbers revolve around 94% of F1 for title detection, 76% of F1 for date detection and 89% of F1 for body detection.

We also implemented two works in the literature that used a full rendering. The first, by Luo, et al. (2009) [22], didn't perform as well as expected, but we showed that our simplified CSS can be used as an alternative to their full rendering, presenting similar end results. The second, by Wang, et al. (2009) [40, 41], was more successful. Our results were close enough to the ones reported that we feel confident to say we reproduced their results. However, we found some potential issues with their method, as depicted by our content-based metrics for their approach.

We also learned more about rendering engines, and feel the end result was not as robust as we would like. We ran into memory leaks and had difficulties processing a large batch of documents at once. Sticking with simple HTML parsers and our simplified CSS parser proved to be much more robust, and we feel that it requires less resources to run. Given that the chosen method allows it and the results are satisfactory to the task at hand, we prefer to avoid rendering webpages.

As future work, we would like to extend our approach to other relevant metadata in the news domain, such as author and news agency, when available, and other tasks and domains in the web where visual features might be useful, such as e-commerce and product price comparison.