Structured data extraction from the Web

Web Information Extraction and Retrieval 2018/19, Faculty of Computer and Information Science, University of Ljubljana

Matej Klemen, Andraž Povše, Jaka Stavanja

***Abstract*—In this work we present our implementation of 2 different approaches for structured data extraction from the Web: using regular expressions and using XPath. We test the implemented methods on 6 webpages from 3 different sources (Overstock, Rtvslo and Avto.net) and provide the outputs for the pages, generated by the methods.**

I. Introduction

The Web is an ever-increasing collection of information. Probably the most used format for representing this information is HTML. After crawling a webpage (which was done in the previous assignment) we would like to be able to automatically extract useful parts of the page. Unfortunately, the popularity of HTML makes this task a little harder, since its primary goal is to make pages readable by humans, not necessarily computers.

For this assignment, we implement 2 different approaches to structured data extraction for 6 webpages from 3 different sources (Overstock, Rtvslo and Avto.net). These are data extraction using regular expressions and using XPath query language.

The rest of this report is structured as follows. [**TODO: update this**] In chapter ... we present ....

II. Used data

In addition to the provided webpages from *Overstock* and *Rtvslo*, we select another source from which we obtain 2 similar webpages. The third source on which we test our implemented methods are 2 webpages from *Avto.net*. The data items and data records we are interested in are shown on Figure [**TODO: ref image once its done**].

TODO: describe the 2 selected webpages (from Avto.net) and show a picture with identification of data items and data records.

III. Methodology

In this section we describe our implementations of the approaches using regular expressions and XPath.

A. Regular expressions approach

[**TODO: describe implementation and reference the Appendix, where you put the output of regex method for the webpages**]

B. XPath

[**TODO: describe implementation and reference the Appendix, where you put the output of xpath method for the webpages**]

IV. Conclusion

We presented 2 approaches to structured data extraction from the Web: using regular expressions and using XPath query language. We applied these methods to 6 webpages and provided the methods' output. The third, more general, approach to structured data extraction, using RoadRunner algorithm was not implemented.

Figure 1. A nice plot showing something really cool and awesome.

Appendix
Outputs of the methods

[**TODO: provide outputs for all the webpages for each method (probably in the form of some dank lstlisting?**]