WIER: Report on second assignment

1 Introduction

In this assignment, we explore two methods for structured data extraction from web pages: regular expressions and XPath. We apply these approaches to various examples of web pages and present the obtained results in the following sections.

2 Methodology

In this section we will go over the methods for structured data extraction as well as describe the web pages from which we extracted the data.

2.1 Regular Expressions

We start with a basic method: treating HTML content as text and using regular expressions to capture the desired information. We match elements by their class and extract data using capture groups. Each data item has its own regular expression, resulting in lists of the same data type for different records in list views.

2.2 XPath

The other method we used was XPath, where we extract data by querying the web page's DOM tree. In our examples, XPath expressions typically mirror the regular expressions, but with a much more readable syntax.

2.3 Web Pages

We analyze three types of web pages: RTV.si news articles, Overstock.com product listings, and Pokémon index pages. Our goal is to extract relevant data from each:

- **RTV.si News Articles:** Each page represents a news article from RTV Slovenia. We focus on extracting a title, subtitle, author, published time, lead, and content from these pages.

- **Overstock.com Product Listings:** Each page showcases a list of products from the Overstock web store. We aim to extract the title, list price, current price, savings, savings percentage, and product description for each item listed on the page.
- **Pokémon Index Pages:** These pages contain descriptions of Pokémon, including details such as pokemon name, number, type, weaknesses, height, weight, and its description. You can see an example of such a page on figure 1

3 Conclusion

In the previous assignment, we explored three methods for extracting structured data from websites: regular expressions, XPath, and applied them to various website types. Even the straightforward regular expression approach proved effective for our task, while XPath, being more powerful, offered a much more intuitive interface.

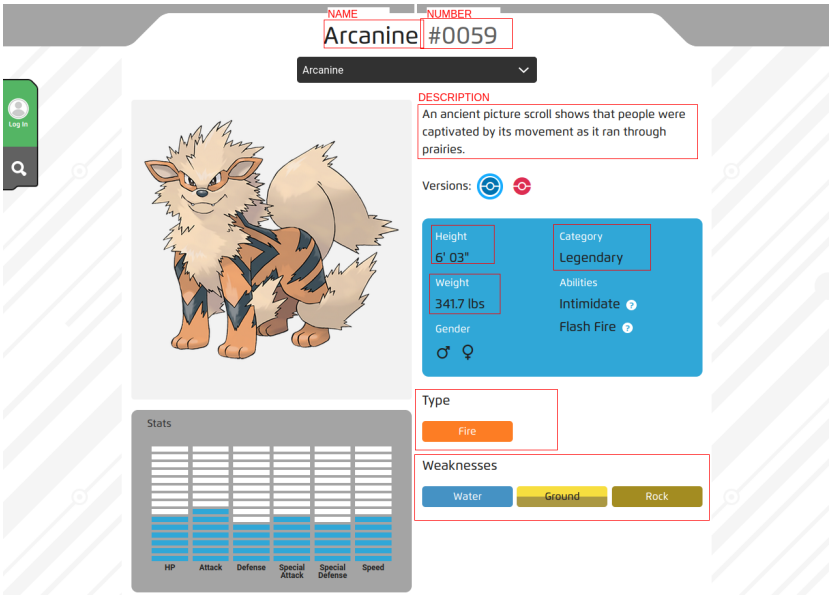

Figure 1: Example of a pokedex entry web page and what we extracted from it