Data Extractor

Timotej Kovač

I. Introduction

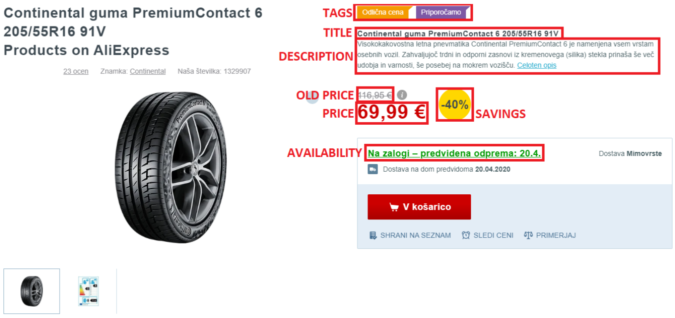
TODO

II. Selection of optional web pages

For our first website we have chosen a detailed description page of a product available on web store Mimovrste ??. Here we have chosen some interesting fields that we might be useful as shown in figure 1. These were:

- tags*, which further describe the item as having a discount, being a recommended product, etc.";
- title;
- description;
- old price*, which states the price before the now discounted price;
- price, which states the current price;
- savings*, which represents the percentage saved;
- availability, which states when the product will be available for shipment.

Fields above marked with an asterisk don't appear always and are therefore optional.



Continental guma PremiumContact 6 205/55R16 91V
Products on AliExpress

23 ocen Znamka: Continental Naša številka: 1329907

Continental guma PremiumContact 6 205/55R16 91V

Visokokakovostna letna pnevmatika Continental PremiumContact 6 je namenjena vsem vrstam osebnih vozil. Zahvaljujoč trdni in odporni zasnovi iz kremenovega (silika) stekla prinaša še več udobja in varnosti, še posebej na mokrem vozišču. Celoten opis

Na zalogi – predvidena odprema: 20.4.

Dostava Mimovrste

Dostava na dom predvidoma 20.04.2020

Figure 1. Web store mimovrste.si with tagged fields that we used for extraction of web content.

For our second website we have chosen a page containing multiple items in a grid pattern on a web site Ceneje ??. Here we have chosen the fields listed bellow:

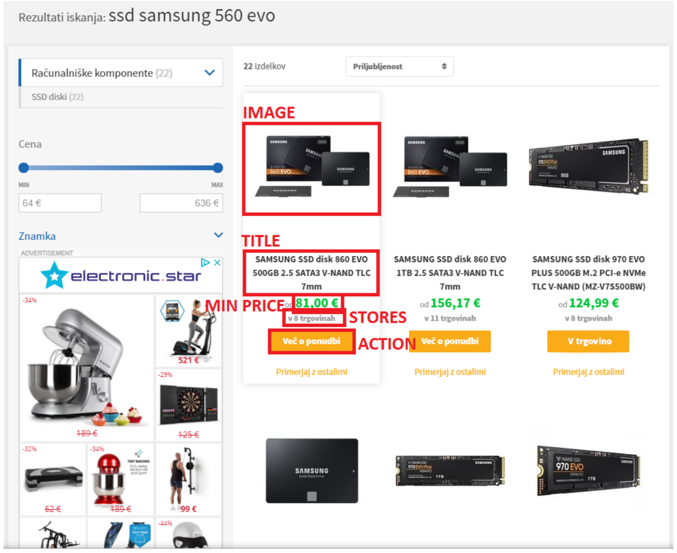
- image;
- title;
- min price, which states the minimal price in all of the stores that provide the product;
- number of stores, which states the number of stores that provide the product;
- action, which states what the button does either takes the user to a particular web store or to a list of webstores still on the same ceneje.si domain.

Here none of the fields are optional but some do vary as they may contain some other words in front of the fields or are ads which have a slightly different structure. This too can be seen in figure 2.

III. Regular Expressions Implementation

rtvslo.si

- title: <h1>(.*?)</h1>



Rezultati iskanja: ssd samsung 560 evo

Figure 2. Web site ceneje.si with tagged fields that we used for extraction of web content.

-
-
-
-

IV. XPath Implementation

TODO

V. Automatic Web Extraction Implementation

TODO

VI. Conclusion

[1], [?], [?] TODO

References

[1] "Jauntium Java Browser Automation," Accessable: https://jauntium.com/, 2020, [Accessed: 27. 03. 2020].