A web scrapper case

Rafael Campos Nunes

April 16, 2019

Contents

1 Requirements **1**

2 Inspection **1**

1 Requirements

The objective of the web scrapper is to list every product with the following characteristics encoded as a JSON:

1. Name of the product
2. Link to one or more of its images
3. Price of the product

The page from which all of this information will be scrapped is: `https://nerdstore.com.br/categoria/especiais/game-of-throne/s`

2 Inspection

When inspecting the page I found that the elements are enclosed in the following structure:

```
<ul class="products ...">
  <li class="... product ...">
    <a class="woocommerce-LoopProduct-link woocommerce-loop-product__link">
      <img class="" src="">
```

```
      <img class="" src="">
      <h2 class="woocommerce-loop-product__title">
        <span class="ellip">text ellipsed?</span>
        <span class="ellip-line">text ellipsed 2?</span>
      </h2>
      <span class="price">
        <span class="woocommerce-Price-amount amount">
          <span class="woocommerce-Price-currencySymbol">R$</span>
          "49,90"
        </span>
      </span>
    </a>
  </li>
  ...
</ul>
```

By inspection I see that a product may or may not have more than one image and that are specific classes to grab the specified element, though this shouldn't be enough I think because they (the developers) might change on the long run. I should return a JSON that looks somewhat like this:

```
o = {
  'name'  : '',
  'img'   : ['', ..., ''],
  'price' : '44,90'
}
```

The price comes with a currency tag and that should be important to have but it's not asked by the head hunters.

What I should do now is to take the big block of products that are contained inside the *ul* tag with the *products* class and work from there.