

Extractor

Parte 3 -José Bezerra (jbmn2)



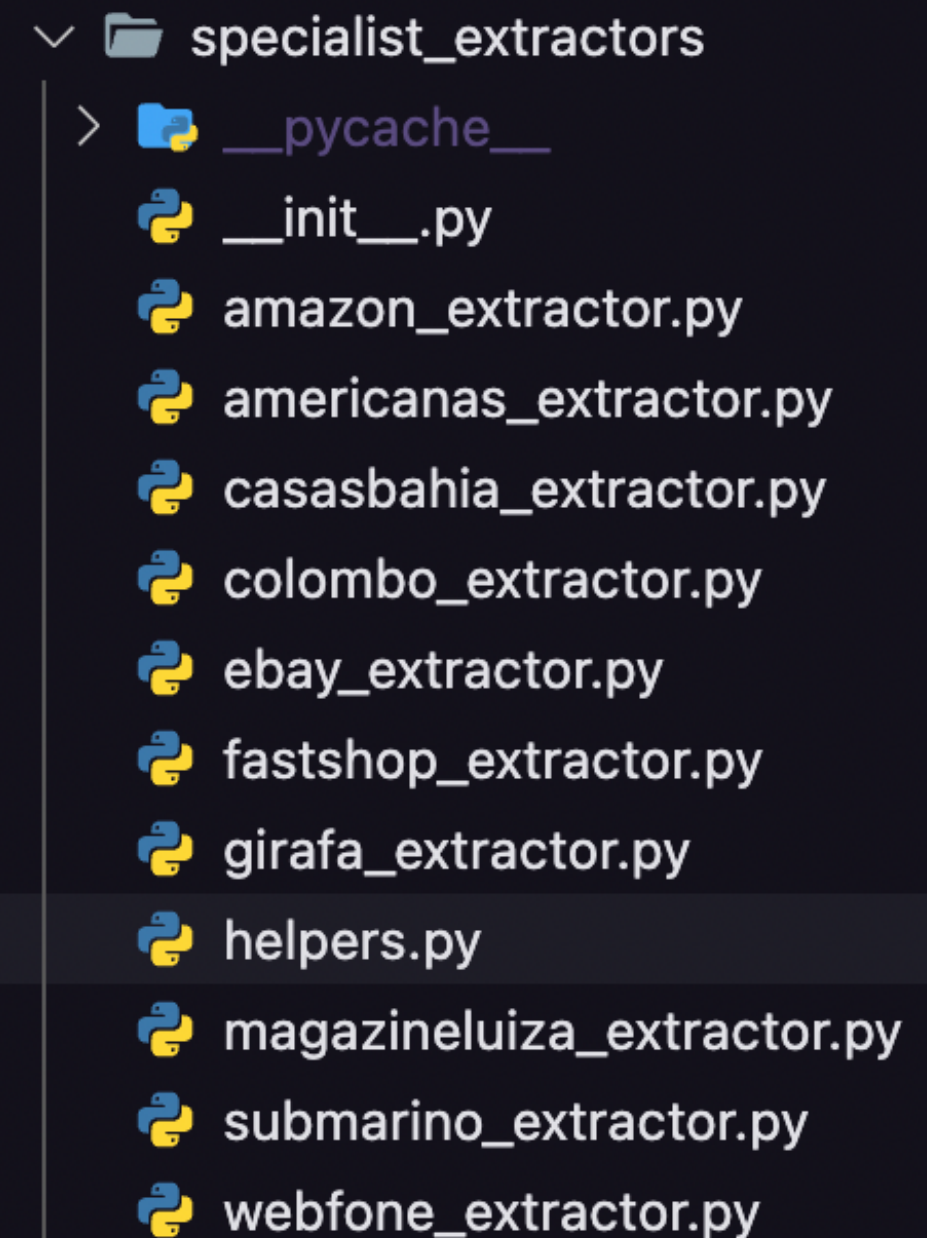


FERRAMENTAS

- BeautifulSoup
 - Navegação e parseamento de arquivo html
- Codecs
 - Lib utilizada para abrir arquivos txt com acentuação sem conflitos.
- RE (Regex)
 - Filtros de classes html
 - Processamento de valores encontrados nas páginas

COLETORES ESPECIALISTAS

Através da correta navegação de HTML e a posterior verificação se o dado encontrado é de interesse, podemos criar coletores específicos para cada tipo de domínio, sem complicações.



```

specialist_extractors
├── __pycache__
├── __init__.py
├── amazon_extractor.py
├── americanas_extractor.py
├── casasbahia_extractor.py
├── colombo_extractor.py
├── ebay_extractor.py
├── fastshop_extractor.py
├── girafa_extractor.py
├── helpers.py
├── magazineluiza_extractor.py
├── submarino_extractor.py
└── webfone_extractor.py

```

COLETOR GENERALISTA

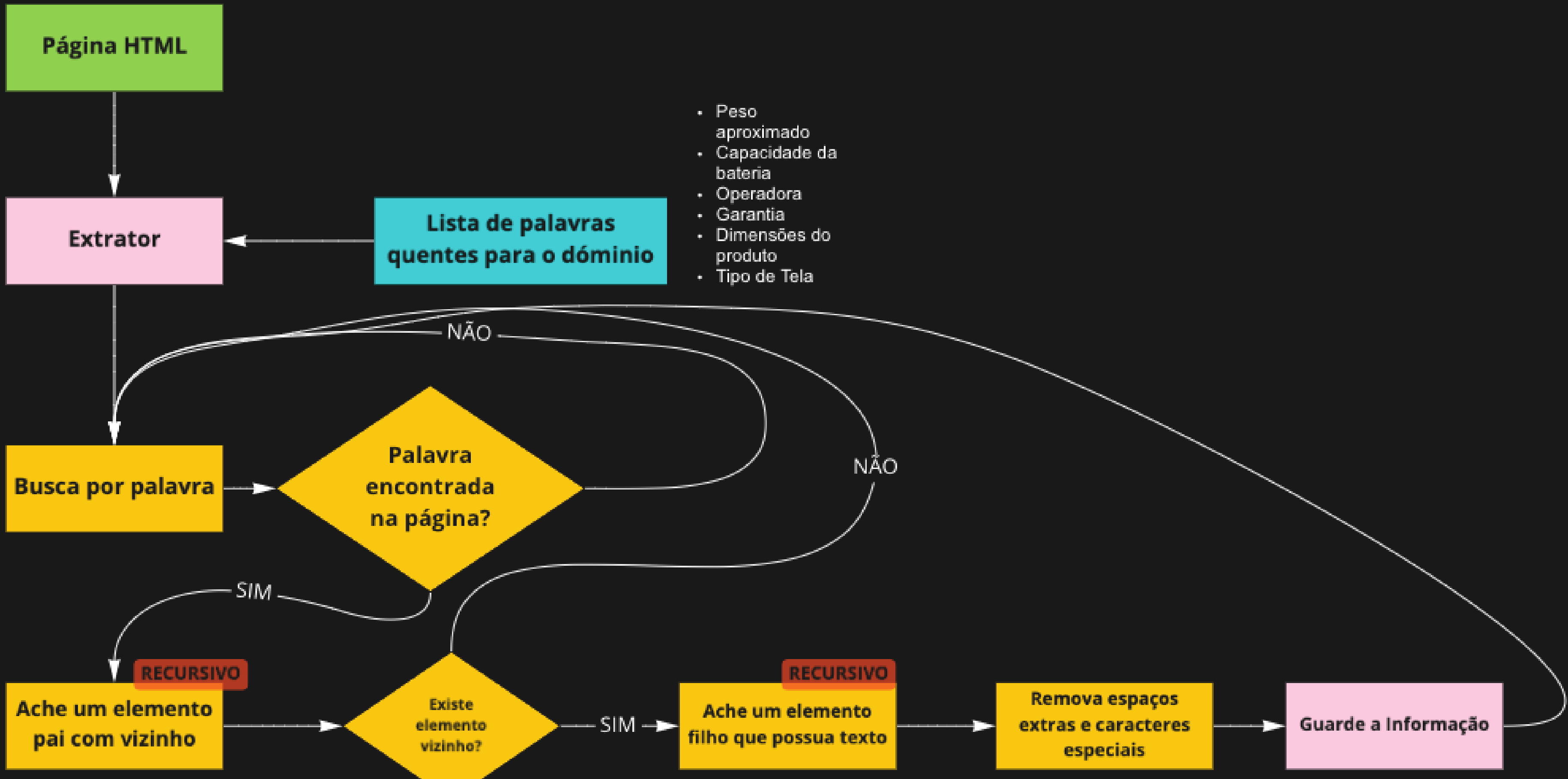
Nosso coletor generalista de domínio funciona através de diversas técnicas de localização de informação, somado a regras de navegação em páginas HTML (DOM Tree), somado a técnicas de processamento de texto para obtenção do resultado final.

Segue o fluxo

COLETOR GENERALISTA

Nosso coletor generalista de domínio funciona através de diversas técnicas de localização de informação, somado a regras de navegação em páginas HTML (DOM Tree), somado a técnicas de processamento de texto para obtenção do resultado final.

Segue o fluxo




```

▼<table class="src__SpecsCell-sc-70o4ee-5 gYhGqJ">
  ▼<tbody>
    ▼<tr class="spec-drawer__View-sc-jcvy3q-4 eHkstE">
      <td class="spec-drawer__Text-sc-jcvy3q-5 fMwSYd">Código</td> == $0
      <td class="spec-drawer__Text-sc-jcvy3q-5 fMwSYd">4533175065</td>
    </tr>
    ▼<tr class="spec-drawer__View-sc-jcvy3q-4 eHkstE">
      <td class="spec-drawer__Text-sc-jcvy3q-5 fMwSYd">Código de barras
      </td>
      <td class="spec-drawer__Text-sc-jcvy3q-5 fMwSYd">7892509121217</td>
    </tr>
  
```

```

▼<div data-testid="ux-layout-section" class="ux-layout-section ux-l
yout-section--features">
  ▼<div data-testid="ux-layout-section__item" class="ux-layout-sect
on__item ux-layout-section__item--table-view">
    ▼<div class="ux-layout-section__row">
      ▼<div class="ux-labels-values__labels">
        ▼<div class="ux-labels-values__labels-content">
          ▼<div>
            <span class="ux-textspans">Estado do item:</span>
          </div>
        </div>
      </div>
      ▼<div class="ux-labels-values__values">
        ▼<div class="ux-labels-values__values-content">
          ▼<div> == $0
            <span class="ux-textspans">Excelente - recondicionado
            </span>
          </div>
        </div>
      </div>
    </div>
    ▶<div class="ux-labels-values__labels">...</div>
    ▶<div class="ux-labels-values__values">...</div>
  
```


RESULTADOS 1/2

AMAZON	100%
AMERICANAS	91.6%
SUBMARINO	91.6%
MAGAZINE LUIZA	69.6%
CASAS BAHIA	63.6%

RESULTADOS 2/2

WEBFONE	100%
EBAY	100%
COLOMBO	88.23%
FASTSHOP	44.4%
GIRAFA	0%

