

# Scrapy

Projeto Crawler Telelista

Com tudo que aprendemos até agora temos conhecimento para analisar um determinado site e começar a retirar informações do mesmo.

Neste projeto prático vamos pegar informações do site da Telelista.

O que vamos fazer não é ilegal pois estamos retirando informações que estão disponíveis para qualquer um na internet.

Como em qualquer mineração de dados na internet, uma mudança no site fará com que nosso exemplo pare de funcionar. Mas fica registrado o processo que tive que seguir para conseguir retirar informações que me interessavam. Tal processo pode lhe ajudar em outros casos.

Primeiro vamos criar nosso projeto.

Crie uma pasta no seu computador chamada CrawlerTelelista e acesse a mesma via prompt de comando/terminal para criarmos o projeto.

# Scrapy

## Projeto Crawler Telelista

```
Administrator: Command Prompt
P:\Temp\CrawlerTelelista>scrapy startproject crawlertelelista
New Scrapy project 'crawlertelelista', using template directory 'c:\\evaldo\\ferramentasdesenvolvimento\\python\\python36\\lib\\site-packages\\scrapy\\templates\\project', created in:
    P:\Temp\CrawlerTelelista\crawlertelelista

You can start your first spider with:
    cd crawlertelelista
    scrapy genspider example example.com

P:\Temp\CrawlerTelelista>
```

# Scrapy

## Projeto Crawler Telelista

Antes de começar a escrever o spider vamos analisar o HTML do site e as requisições que vamos executar.

O site é:

`https:\\www.telelistas.net`



# Scrapy

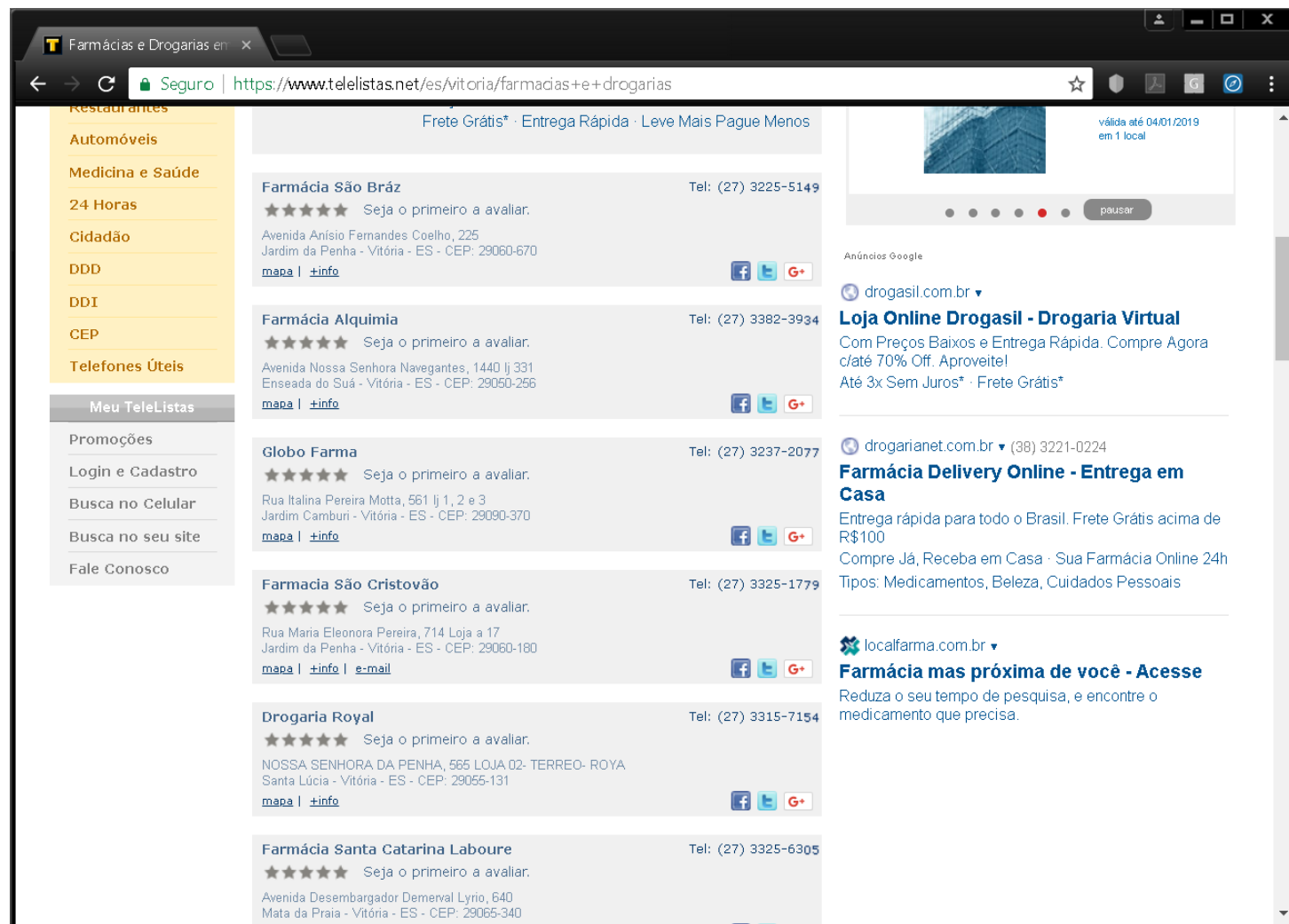
## Projeto Crawler Telelista

Pesquise a palavra-chave “farmacias”, UF igual a “ES” e cidade igual a “Vitória”, por exemplo.

Será exibida uma lista de produtos relacionados à farmácias. Clique em farmácias e Drogarias.



Será exibida uma lista de farmácias com endereço e telefone.





# Scrapy

## Projeto Crawler Telelista

Ao clicar com o botão direito e exibir o código-fonte, podemos avaliar no HTML da página se existe algum padrão para apresentação dos dados.



```
1179 </td>
1180 <td width="460" align="left" valign="middle">
1181 <table width="459" border="0" cellspacing="0" cellpadding="0">
1182 <tr>
1183 <td height="5" colspan="3" class="text_medio">
1184 </td>
1185 </tr>
1186 <tr>
1187 <td width="324" colspan="2" align="left" valign="top" class="nome_resultado_ag">
1188 <a
1189 href="https://www.telelistas.net/loais/es/vitoria/farmacias+e+drogarias/3013121
1190 87/farmacia+sao+braz">
1191 Farmácia São Bráz</a>
1192 </td>
1193 <td width="135" rowspan="3" align="right" valign="top"
1194 class="text_resultado_ib">
1195 Tel: (27) 3225-51
1197 </td>
1198 </tr>
1199 <tr>
1200 <td height="1" colspan="3" align="left" width="90%" align="left" valign="top">
1201 <span class="starRow_AG" id="starRow">
1202 <div class='star_AG star0_AG'>
1203 </div>
1204 </span>
1205 </td>
1206 </tr>
1207 </table>
1208 </td>
1209 </tr>
1210 </table>
```

Ao analisar o HTML observamos que existem algumas tabelas com suas linhas e colunas contendo as informações que precisamos.

# Scrapy

## Projeto Crawler Telelista

Nome do estabelecimento:

O nome do estabelecimento está no texto de uma tag “td” (coluna) cuja classe é “nome\_resultado\_ag”.

```
<td width="324" colspan="2" align="left" valign="top" class="nome_resultado_ag">  
<a href="https://www.telelistas.net/locais/es/vitoria/farmacias+e+drogarias/301312187/farmacia+sao+braz">  
Farmácia São Bráz</a>  
</td>
```

Desta forma podemos pegar todos os nomes apresentados na página e colocá-los em uma lista. Faça o teste com o *scrapy shell*.

```
scrapy shell "https://www.telelistas.net/es/afonso+claudio/supermercados+e+hipermercados"
```

```
nomes = response.xpath('//td[@class="nome_resultado_ag"]//a/text()').extract()
```

```
In [2]: nomes
```

```
Out[2]:
```

```
['\nComercial Guifer', '\nSupermercados Mirante', '\nSupermercado Giestas', '\nBar e  
Mercearia Meira', '\nSupermercado Schwambach E Tesch', '\nLevi Tesch']
```

# Scrapy

## Projeto Crawler Telelista

Telefone:

O telefone do estabelecimento está no texto de uma tag “td” cuja classe é “text\_resultado\_ib”.

```
<td width="135" rowspan="3" align="right" valign="top" class="text_resultado_ib">  
Tel: (27) 3225-51  
</td>
```

# Scrapy

## Projeto Crawler Telelista

Particularidades do telefone:  
Existem tags “td” cuja classe é  
“text\_resultado\_ib” e não são tags  
relacionadas ao telefone.

```
<td colspan="2" width="294" align="left" valign="top" class="text_resultado_ib">  
Nossa missão é satisfazer às necessidades de saúde, bem-estar, higiene e  
conforto da população, ...  
</td>
```

# Scrapy

## Projeto Crawler Telelista

### Particularidades do telefone:

É um padrão em todos os registros de telefone aparecer o texto “Tel:” ou “PABX:”, já na tag abaixo não existe esta informação no texto da tag “td”. Sendo assim, podemos verificar se no text da tag existe uma dessas palavras, por exemplo.

```
<td colspan="2" width="294" align="left" valign="top" class="text_resultado_ib">  
Nossa missão é satisfazer às necessidades de saúde, bem-estar, higiene e  
conforto da população, ...  
</td>
```

```
<td width="135" rowspan="3" align="right" valign="top" class="text_resultado_ib">  
Tel: (27) 3225-51  
</td>
```

# Scrapy

## Projeto Crawler Telelista

Particularidades do telefone:

Os dois últimos números do telefone não são exibidos, sendo substituídos por uma imagem gerada pelo sistema. Esta é uma particularidade importante. Foi feito desta forma exatamente para evitar o uso de robôs.

```
1 <td width="135" rowspan="3" align="right" valign="top" class="text_resultado_ib">  
2 Tel: (27) 3225-51  
4 </td>
```



## Projeto Crawler Telelista

Particularidades do telefone:

Após gerarmos uma lista de nomes, vamos gerar uma lista de telefones da seguinte maneira:

```
telefones = response.xpath('//*[td[@class="text_resultado_ib" and contains(text(),"Tel")] /text()').extract()
```

Este será o resultado:

Teremos que “limpar” este resultado.

```
In [4]: telefones
Out[4]:
['\nTel: (27) 3735-25',
'\n',
'\nTel: (27) 3735-20',
'\n',
'\nTel: (27) 3735-21',
'\n',
'\nTel: (27) 3735-16',
'\n',
'\nTel: (27) 3735-12',
'\n',
'\nTel: (27) 3735-13',
'\n']
```

# Scrapy

## Projeto Crawler Telelista

Navegação para próxima página:

Observe no final da página que existe o link “próxima” para navegar para a próxima página (caso o resultado retorne mais de uma página). Esse botão está na tag link com a propriedade “rel” igual a “next”.



## Projeto Crawler Telelista

Navegação para próxima página:

Teste com:

```
scrapy shell "https://www.telelistas.net/es/vitoria/supermercados+e+hipermercados"
```

```
In [1]: proxima_pagina = response.xpath('//link[contains(@rel, "next")]/@href').extract_first()
```

```
In [2]: proxima_pagina
```

```
Out[2]: 'https://www.telelistas.net/es/vitoria/supermercados+e+hipermercados?pagina=2'
```

# Scrapy

## Projeto Crawler Telelista

Endereço:

O endereço do estabelecimento está na propriedade text de uma tag “td” com a classe igual a “text\_endereco\_ib”.

```
<td colspan="2" width="294" align="left" valign="top" class="text_endereco_ib">  
Avenida Maruípe, 1259<br>  
Maruípe - Vitória - ES - CEP: 29043-215  
</td>
```

Podemos pegar uma lista de endereços com:

```
enderecos = response.xpath('//td[@class="text_endereco_ib"]/text()).extract()
```

“Mais telefones” ou “Ver telefone”:

Para estabelecimentos que possuem mais de um telefone o site não mostra o número na primeira página, ou mostra apenas um número exibindo um link para ver os outros números.

# Scrapy

## Projeto Crawler Telelista

“Mais telefones” ou “Ver telefone”:  
Ao procurarmos no “ES” pela cidade “Vila Velha”  
teremos um exemplo de “Mais telefones”.

Rede Farmes - Unidade Cobilândia

★★★★★ Seja o primeiro a avaliar.

Rua Papa João Xxiii, 436  
Cobilândia - Vila Velha - ES - CEP: 29111-400

[mapa](#) | [site](#) | [+info](#) | [e-mail](#)

Tel: (27) 3226-8197

[mais telefones](#)

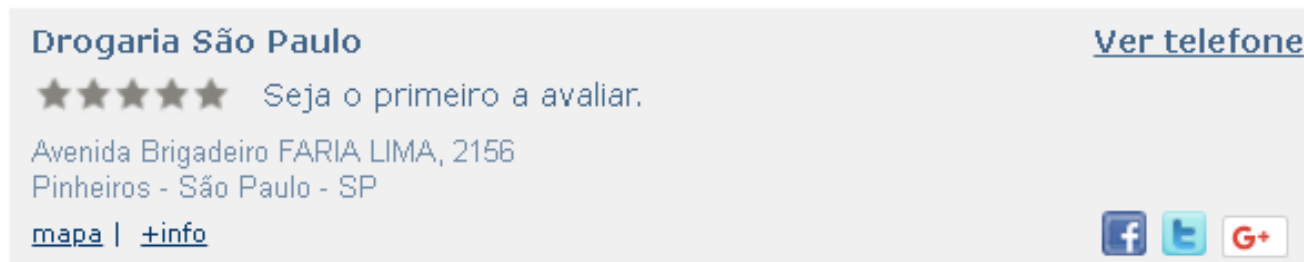
  

```
<td width="135" rowspan="3" align="right" valign="top" class="text_resultado_ib">  
Tel: (27) 3226-81 <div class='ib_ser'>  
<a href='https://www.telelistas.net/locais/es/vila+velha/  
farmacias+e+drogarias/336846896/rede+farmes+unidade+cobilandia'  
style="cursor:hand;">mais telefones</a></div>  
</td>
```

# Scrapy

## Projeto Crawler Telelista



“Mais telefones” ou “Ver telefone”:  
Ao procurarmos em “SP” pela cidade “São Paulo”  
teremos um exemplo de “Ver telefone”.



```
<td width="135" rowspan="3" align="right" valign="top" class="text_resultado_ib">  
<a href="https://www.telelistas.net/locais/sp/sao+paulo/  
farmacias+e+drogarias/298350651/drogaria+sao+paulo"  
onclick="doLogPerformance('IG', 13, 298350651, 11407838);"><u>  
<b>Ver telefone</b></u></a></td>
```

“Mais telefones” ou “Ver telefone”:

Nos dois casos temos uma tag “a” dentro da tag “td”, o que vai servir como referência para ignorarmos o telefone da primeira página e entrar no link exibido para pegarmos todos os telefones.

**Rede Farmes - Unidade Cobilândia**  
★★★★★ [Seja o primeiro a avaliar. Faça seu login para prosseguir](#)  
 Tel: (27) 3226-8197  
Tel: (27) 3226-6835  
Tel: (27) 3359-4412  
 Rua Papa João Xxiii, 436  
Cobilândia - Vila Velha - ES



Página individual:

À partir deste ponto teremos uma página específica diferente da página geral.

## Projeto Crawler Telelista

```
<h1 class="nome_anun item">Epa Boa Praça Supermercados</h1>

<div id="telInfo" class="infoplus_text2 telInfo">
<div style="float: left;">

</div>
<div style="float: left;">
<span>Tel: (27) 3332-47
<img src='https://www.telelistas.net/ImgFactory.ashx?t=6E79&s=4' style='vertical-align: bottom; margin-bottom: 1px;' alt='...'
/></span><br>
<span>Tel: (27) 3223-14
<img src='https://www.telelistas.net/ImgFactory.ashx?t=6E7D&s=4' style='vertical-align: bottom; margin-bottom: 1px;' alt='...' />
</span><br>
<span>Tel: (27) 3322-47
<img src='https://www.telelistas.net/ImgFactory.ashx?t=6E78&s=4' style='vertical-align: bottom; margin-bottom: 1px;' alt='...'
/></span></div></div>
<input type="hidden" id="enderecoreg" value="Avenida Alberto Torres" />
```

## Projeto Crawler Telelista

Página individual:

Com o código abaixo teremos quatro listas, sendo uma com o nome, outra com os telefones, outra com os links das imagens que complementam os telefones e outra com o endereço.

```
nome = response.xpath('//h1[contains(@class,"nome_anun")]/text()').extract()

telefones = response.xpath('//div[@id="telInfo"]//span/text()').extract()

imagens = response.xpath('//div[@id="telInfo"]//span//img/@src').extract()

endereco = response.xpath('//input[contains(@id,"enderecoreg")]/@value').extract()
```

### Tratando o endereço:

```
endereco = response.xpath('//input[contains(@id,"enderecoreg")]/@value').extract()
```

A tag usada retorna somente o nome da rua. Não temos uma tag com o endereço completo identificada. A tag div contendo o endereço completo é assim:

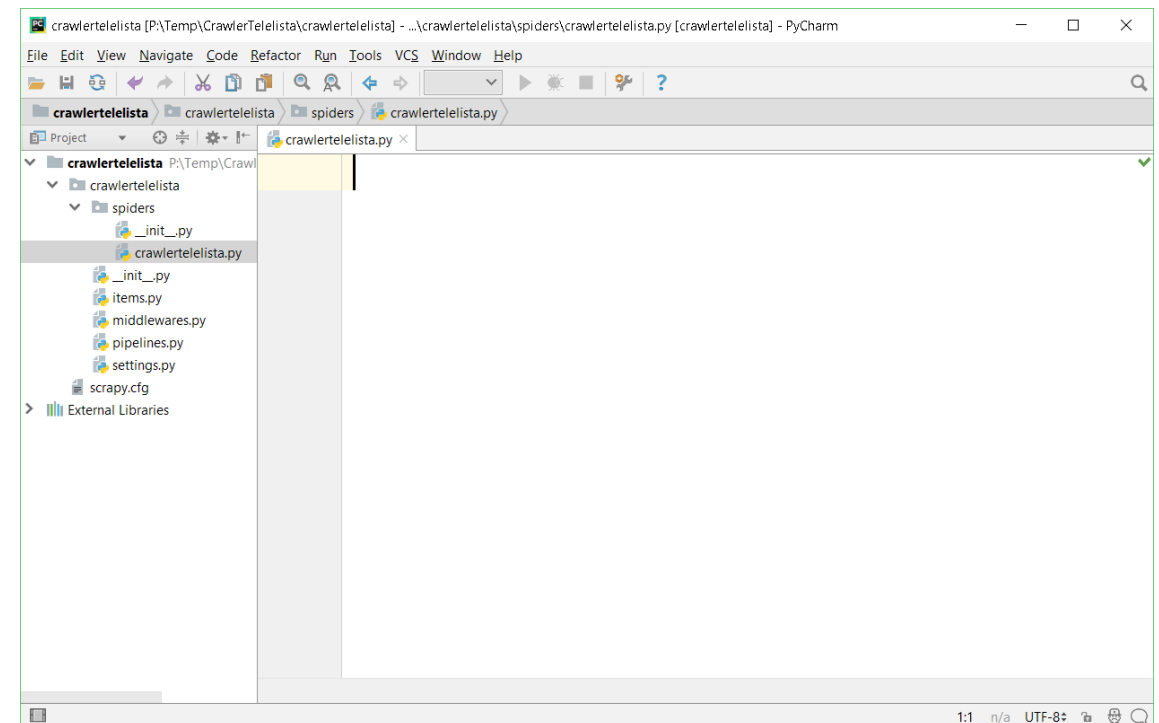
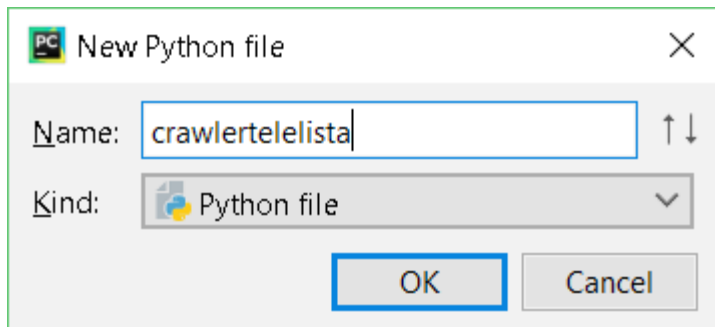
```
<div>Avenida Alberto Torres, 423<br>Jucutuquara - Vitória - ES</div>
```

Vou usar o endereço localizado com xpath para retornar o texto de uma div que contenha o texto do endereço encontrado. Ou seja, vou pegar o texto da div que contenha “Avenida Alberto Torres”, baseado neste exemplo.

# Scrapy

## Projeto Crawler Telelista

Abra o projeto no PyCharm e vamos começar a escrever o spider.



## Projeto Crawler Telelista

Para simplificar o exemplo, vamos gravar os dados em um arquivo de texto e vamos fazer download da imagem que representa os dois últimos números do telefone.

O exemplo pode ser evoluído para gravar os dados e a imagem em um banco de dados.

Vou disponibilizar junto à aula os fontes utilizando banco de dados Microsoft SQL Server. A biblioteca utilizada para conexão com o banco de dados é a pyodbc. Neste programa os dados capturados são gravados no banco de dados e é gravado o endereço da imagem com os dois números do telefone salva em disco.

# Scrapy

## Projeto Crawler Telelista

Vou citar algumas partes do programa que merecem atenção. Eu vou passar a URL por parâmetro na chamada da função, vou salvar em uma variável e depois passar para a requisição. Vou salvar também o tipo de estabelecimento na variável `self.ramo_atividade`.

```
scrapy crawl telelista -a url="https://www.telelistas.net/es/afonso+claudio/supermercados+e+hipermercados"
```

```
def __init__(self, url=''):
    super(SpiderTelelista, self).__init__()

    url_desmontada = url.split("/")
    self.ramo_atividade = url_desmontada[5]
    self.url = url
```

```
def start_requests(self):
    yield scrapy.Request(url=self.url, callback=self.parse)
```

Temos dois estilos de URL que serão acessadas pelo sistema. Um que busca por estado, cidade e tipo:

<https://www.telelistas.net/es/afonso+claudio/supermercados+e+hipermercados>





## Projeto Crawler Telelista

E outro que busca dados de um estabelecimento específico:

<https://www.telelistas.net/locais/es/vitoria/supermercados+e+hipermarcados/335521454/epa+boa+praca+supermercados>

The screenshot displays a web page from telelistas.net. The main content area features a yellow-bordered box for 'Epa Boa Praça Supermercados'. Inside this box, there are five stars, a link to 'Seja o primeiro a avaliar. Faça seu login para prosseguir', and three phone numbers: (27) 3332-4714, (27) 3223-1410, and (27) 3322-4715. Below the phone numbers is the address 'Avenida Alberto Torres, 423 Jucutuquara - Vitória - ES' and a category link 'Supermercados e Hipermercados'. To the right of the yellow box, there is a small section titled 'Epa Boa Praça Supermercados'. The page also includes a top navigation bar with the URL, a sidebar with social media links (Facebook, Twitter), and a right sidebar with various promotional banners and links.

## Projeto Crawler Telelista

Em cada caso o HTML gerado é diferente.

Sendo assim fiz uma validação para avaliar qual URL está sendo processada.

A validação foi simples, contando o número de “/” na URL, caso tenha 5 barras

(<https://www.telelistas.net/es/afonso+claudio/supermercados+e+hipermercados>)

é uma URL geral, caso tenha um número diferente é uma URL de um estabelecimento específico

(<https://www.telelistas.net/locais/es/vitoria/supermercados+e+hipermercados/335521454/epa+boa+praca+supermercados>) •

Como o estado, a cidade e o tipo de estabelecimento estão na URL, vou desmontar a URL para pegar estas informações e vou trocar o “+” da cidade por um espaço.

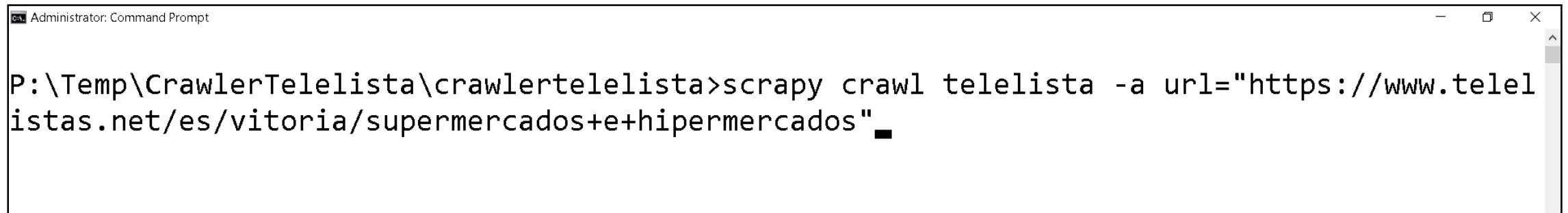
```
link_desmontado = response.url.split("/")
uf_busca = link_desmontado[3]
cidade_busca = link_desmontado[4]
cidade_busca = cidade_busca.replace("+", " ")
```

# Scrapy

## Projeto Crawler Telelista

Executando o sistema:

\$ crawl telelista -a url="URL"

A screenshot of a Windows Command Prompt window titled "Administrator: Command Prompt". The window shows the command `scrapy crawl telelista -a url="https://www.telelistas.net/es/vitoria/supermercados+e+hipermercados"` being entered at the prompt. The path `P:\Temp\CrawlerTelelista\crawllertelelista>` is visible before the command. The cursor is at the end of the command line.

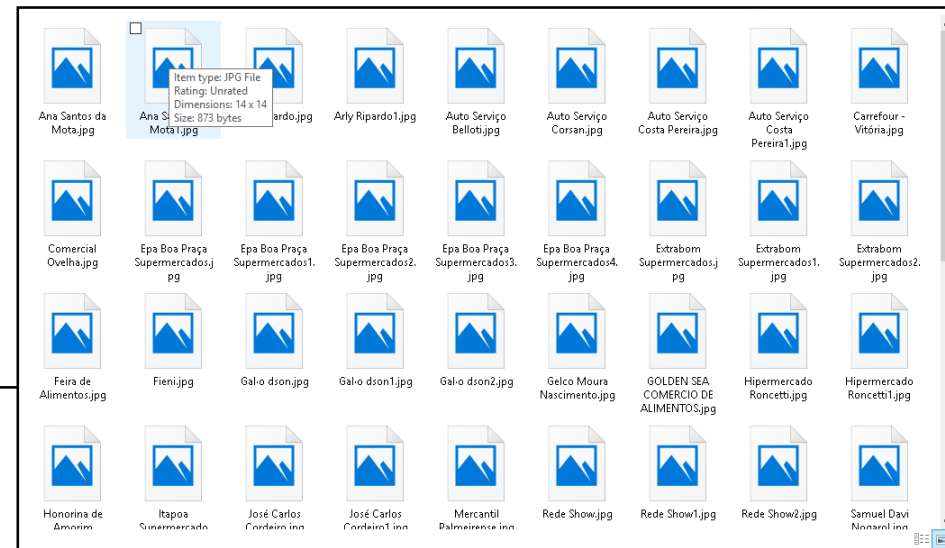
```
Administrator: Command Prompt

P:\Temp\CrawlerTelelista\crawllertelelista>scrapy crawl telelista -a url="https://www.telelistas.net/es/vitoria/supermercados+e+hipermercados"
```

## Projeto Crawler Telelista

### Resultado:

```
es - vitoria -  
Supermercado São José -  
Avenida Marechal Mascarenhas Moraes, 1304  
Vitória - ES  
- supermercados+e+hipermercados -  
Tel: (27) 3019-50 - C:\Temp\imagens_capturadas\es\Supermercado São José.jpg  
es - vitoria -  
Epa Boa Praça Supermercados -  
Praça Regina Frigeri Furno, 210  
Jardim da Penha - Vitória - ES - CEP: 29060-200  
- supermercados+e+hipermercados -  
Tel: (27) 3314-32 - C:\Temp\imagens_capturadas\es\Epa Boa Praça Supermercados.jpg  
es - vitoria -  
Epa Boa Praça Supermercados -  
Rua Eugênio Netto, 631  
Santa Lúcia - Vitória - ES - CEP: 29055-270  
- supermercados+e+hipermercados -  
Tel: (27) 3325-10 - C:\Temp\imagens_capturadas\es\Epa Boa Praça Supermercados1.jpg
```



# FIM