

# Análise Clima-Tempo

Semana 12 | Web Scraping, Regex e Pandas

## Objetivo do miniprojeto

A partir de sua escolha de uma cidade qualquer, faça a raspagem de uma página e descubra a temperatura média da próxima semana nesta cidade. Utilize o serviço [Tempo](#) na sua solução. (Desejável incluir gráficos do Plotly).

## Projeto-piloto

Aqui, importamos as bibliotecas que utilizaremos para fazer a raspagem e análise de dados.

```
In [53]: import requests
from bs4 import BeautifulSoup
import pandas as pd
import numpy as np
import plotly.express as px
```

### Seleção do serviço de clima-tempo

Apenas para a ilustração dos passos, vamos fazer a raspagem do [National Weather Service](#), um serviço de climatologia norte-americano. A cidade indicada na `url` pela latitude e longitude aponta para a cidade de *Ciudad Juarez* (MX), na fronteira dos EUA.

**Para que estes passos sejam aplicáveis, é importante verificar se o serviço de clima-tempo apresenta as informações utilizando HTML ou Javascript. Estes passos só servem para respostas em HTML.**

```
In [5]: # Localizamos uma cidade usando o próprio serviço de clima-tempo
url = "https://forecast.weather.gov/MapClick.php?lat=31.604&lon=-106.2511#.Xo"
```

Download da página. Uma resposta 200 indica que o retorno da página foi bem-sucedido.

```
In [6]: # usamos a biblioteca requests para fazer o download da página web
page = requests.get( url )
page
```

```
Out[6]: <Response [200]>
```

A biblioteca `BeautifulSoup` permite que manipulemos o HTML da página.

```
In [10]: # criamos um objeto beautiful soup para poder fazer a raspagem dos dados
soup = BeautifulSoup(page.content, 'html.parser')
#soup
#print(soup.prettify())
```

Utilizando o modo de inspeção do navegador (ex: Chrome, Firefox, Safari...), identifique o elemento HTML em que deseja extrair as informações.

```
In [12]: # após investigar, com um navegador web, a estrutura da página e localizar o
seven_day = soup.find(id="seven-day-forecast")
seven_day
```

```
Out[12]: <div class="panel panel-default" id="seven-day-forecast">
```

```

<div class="panel-heading">
<b>Extended Forecast for</b>
<h2 class="panel-title">
2 Miles NW Clint TX </h2>
</div>
<div class="panel-body" id="seven-day-forecast-body">
<div id="seven-day-forecast-container"><ul class="list-unstyled" id="seven-day-forecast-list"><li class="forecast-tombstone">
<div class="tombstone-container">
<p class="period-name">Today<br/><br/></p>
<p></p><p class="short-desc">Sunny</p><p class="temp temp-high">High: 95 °F</p></div></li><li class="forecast-tombstone">
<div class="tombstone-container">
<p class="period-name">Tonight<br/><br/></p>
<p></p><p class="short-desc">Mostly Clear</p><p class="temp temp-low">Low: 62 °F</p></div>
</li><li class="forecast-tombstone">
<div class="tombstone-container">
<p class="period-name">Wednesday<br/><br/></p>
<p></p><p class="short-desc">Mostly Sunny</p><p class="temp temp-high">High: 96 °F</p></div>
</li><li class="forecast-tombstone">
<div class="tombstone-container">
<p class="period-name">Wednesday Night<br/><br/></p>
<p></p><p class="short-desc">Partly Cloudy</p><p class="temp temp-low">Low: 64 °F</p></div></li><li class="forecast-tombstone">
<div class="tombstone-container">
<p class="period-name">Thursday<br/><br/></p>
<p></p><p class="short-desc">Mostly Sunny</p><p class="temp temp-high">High: 96 °F</p></div></li><li class="forecast-tombstone">
<div class="tombstone-container">
<p class="period-name">Thursday Night<br/><br/></p>
<p></p><p class="short-desc">Mostly Clear</p><p class="temp temp-low">Low: 64 °F</p></div></li><li class="forecast-tombstone">
<div class="tombstone-container">
<p class="period-name">Friday<br/><br/></p>
<p></p><p class="short-desc">Sunny</p><p class="temp temp-high">High: 97 °F</p></div></li><li class="forecast-tombstone">
<div class="tombstone-container">
<p class="period-name">Friday Night<br/><br/></p>
<p></p><p class="short-desc">Clear</p><p class="temp temp-low">Low: 64 °F</p></div></li><li class="forecast-tombstone">

```

```

<div class="tombstone-container">
<p class="period-name">Saturday<br/><br/></p>
<p></p><p class="short-desc">Sunny</p><p class="temp temp-high">High: 100 °F</p></div></li></ul></div>
<script type="text/javascript">
// equalize forecast heights
$(function () {
    var maxh = 0;
    $(".forecast-tombstone .short-desc").each(function () {
        var h = $(this).height();
        if (h > maxh) { maxh = h; }
    });
    $(".forecast-tombstone .short-desc").height(maxh);
});
</script> </div>
</div>

```

Dentro do elemento HTML `seven-day-forecast` , pegamos todos os elementos da classe `tombstone-container` .

```

In [19]: forecast_items = seven_day.find_all(class_="tombstone-container")
         #forecast_items

```

O primeiro elemento da lista `forecast_items` representa a manhã de hoje.

```

In [20]: morning = forecast_items[0]
         morning

```

```

Out[20]: <div class="tombstone-container">
<p class="period-name">Today<br/><br/></p>
<p></p><p class="short-desc">Sunny</p><p clas s="temp temp-high">High: 95 °F</p></div>

```

```

In [21]: # Impressão na tela formatado
         print(morning.prettify())

```

```

<div class="tombstone-container">
  <p class="period-name">
    Today
    <br/>
    <br/>
  </p>
  <p>
    
  </p>
  <p class="short-desc">
    Sunny
  </p>
  <p class="temp temp-high">
    High: 95 °F
  </p>
</div>

```

O segundo elemento da lista `forecast_items` representa o próximo período.

```

In [22]: afternoon = forecast_items[1]
         print(afternoon.prettify())

```

```

<div class="tombstone-container">

```

```

<p class="period-name">
    Tonight
    <br/>
    <br/>
</p>
<p>
    
    </p>
    <p class="short-desc">
        Mostly Clear
    </p>
    <p class="temp temp-low">
        Low: 62 °F
    </p>
</div>

```

### Extraindo informações da página

A etiqueta representada pela variável `morning` contém toda a informação de que precisamos:

- O nome do item da previsão - neste caso, Morning
- A descrição das condições - está localizado no título da propriedade de `img`
- Uma descrição breve das condições - neste caso Patchy Blowing Dust then Cloudy
- A temperatura mínima - neste caso 81 °F

In [30]: `morning`

Out[30]: 

```
<div class="tombstone-container">
<p class="period-name">Today<br/><br/></p>
<p></p><p class="short-desc">Sunny</p><p clas
s="temp temp-high">High: 95 °F</p></div>
```

In [29]: `#morning.find(class_="period-name").get_text()`

In [27]: 

```
period = morning.find(class_="period-name").get_text()
period
```

Out[27]: `'Today'`

In [31]: 

```
short_desc = morning.find(class_="short-desc").get_text()
short_desc
```

Out[31]: `'Sunny'`

In [32]: 

```
temp = morning.find(class_="temp").get_text()
temp
```

Out[32]: `'High: 95 °F'`

Uma opção seria pegar as informações contidas na tag `<img>` .

In [34]: 

```
img = morning.find("img")
img
```

```
Out[34]: 
```

```
In [35]: # Obtem o título (title) da descrição da img
desc = img['title']
desc
```

```
Out[35]: 'Today: Sunny, with a high near 95. Calm wind becoming west southwest 5 to 9 mph in the morning. '
```

## Extraindo toda a informação da página

Agora que sabemos como extrair a informação de um elemento, vamos obter todos os outros.

```
In [37]: seven_day
```

```
Out[37]: <div class="panel panel-default" id="seven-day-forecast">
<div class="panel-heading">
<b>Extended Forecast for</b>
<h2 class="panel-title">
2 Miles NW Clint TX </h2>
</div>
<div class="panel-body" id="seven-day-forecast-body">
<div id="seven-day-forecast-container"><ul class="list-unstyled" id="seven-day-forecast-list"><li class="forecast-tombstone">
<div class="tombstone-container">
<p class="period-name">Today<br/><br/></p>
<p></p><p class="short-desc">Sunny</p><p class="temp temp-high">High: 95 °F</p></div></li><li class="forecast-tombstone">
<div class="tombstone-container">
<p class="period-name">Tonight<br/><br/></p>
<p></p><p class="short-desc">Mostly Clear</p><p class="temp temp-low">Low: 62 °F</p></div>
</li><li class="forecast-tombstone">
<div class="tombstone-container">
<p class="period-name">Wednesday<br/><br/></p>
<p></p><p class="short-desc">Mostly Sunny</p><p class="temp temp-high">High: 96 °F</p></div>
</li><li class="forecast-tombstone">
<div class="tombstone-container">
<p class="period-name">Wednesday<br/>Night</p>
<p></p><p class="short-desc">Partly Cloudy</p><p class="temp temp-low">Low: 64 °F</p></div></li><li class="forecast-tombstone">
<div class="tombstone-container">
<p class="period-name">Thursday<br/><br/></p>
<p></p><p class="short-desc">Mostly Sunny</p><p class="temp temp-high">High: 96 °F</p></div></li><li class="forecast-tombstone">
<div class="tombstone-container">
<p class="period-name">Thursday<br/>Night</p>
```

```

<p></p><p class="short-desc">Mostly Clear</p><p class="temp temp-low">Low: 64 °F</p></div></li><li class="forecast-tombstone">
<div class="tombstone-container">
<p class="period-name">Friday<br/><br/></p>
<p></p><p class="short-desc">Sunny</p><p class="temp temp-high">High: 97 °F</p></div></li><li class="forecast-tombstone">
<div class="tombstone-container">
<p class="period-name">Friday<br/>Night</p>
<p></p><p class="short-desc">Clear</p><p class="temp temp-low">Low: 64 °F</p></div></li><li class="forecast-tombstone">
<div class="tombstone-container">
<p class="period-name">Saturday<br/><br/></p>
<p></p><p class="short-desc">Sunny</p><p class="temp temp-high">High: 100 °F</p></div></li></ul></div>
<script type="text/javascript">
// equalize forecast heights
$(function () {
    var maxh = 0;
    $(".forecast-tombstone .short-desc").each(function () {
        var h = $(this).height();
        if (h > maxh) { maxh = h; }
    });
    $(".forecast-tombstone .short-desc").height(maxh);
});
</script> </div>
</div>

<div class="tombstone-container">
    <p class="period-name">

    </p>
</div>
<div class="tombstone-container">
    <p class="period-name">

    </p>
</div>
<div class="tombstone-container">
    <p class="period-name">

    </p>
</div>

```

```
In [36]: seven_day.select(".tombstone-container .period-name")
```

```
Out[36]: [<p class="period-name">Today<br/><br/></p>,
<p class="period-name">Tonight<br/><br/></p>,
<p class="period-name">Wednesday<br/><br/></p>,
<p class="period-name">Wednesday<br/>Night</p>,
<p class="period-name">Thursday<br/><br/></p>,
<p class="period-name">Thursday<br/>Night</p>,
<p class="period-name">Friday<br/><br/></p>,
```

```
<p class="period-name">Friday<br/>Night</p>,
<p class="period-name">Saturday<br/><br/></p>]
```

```
In [41]: # Usa o seletor CSS para extrair todos os period-name dentro de tombstone-con
period_tags = seven_day.select(".tombstone-container .period-name")

# Usando list comprehensions (não vimos ainda)
#periods = [pt.get_text() for pt in period_tags]

periods = []
for pt in period_tags:
    periods.append(pt.get_text())
periods
```

```
Out[41]: ['Today',
'Tonight',
'Wednesday',
'WednesdayNight',
'Thursday',
'ThursdayNight',
'Friday',
'FridayNight',
'Saturday']
```

Vamos usar a mesma técnica para os outros três campos.

```
In [42]: # Usando list comprehensions (não vimos ainda)
#short_descs = [sd.get_text() for sd in seven_day.select(".tombstone-containe

short_descs_tags = seven_day.select(".tombstone-container .short-desc")
short_descs = []
for sd in short_descs_tags:
    short_descs.append(sd.get_text())
short_descs
```

```
Out[42]: ['Sunny',
'Mostly Clear',
'Mostly Sunny',
'Partly Cloudy',
'Mostly Sunny',
'Mostly Clear',
'Sunny',
'Clear',
'Sunny']
```

```
In [43]: # Usando list comprehensions (não vimos ainda)
# temps = [t.get_text() for t in seven_day.select(".tombstone-container .temp

temps_tags = seven_day.select(".tombstone-container .temp")
temps = []
for tp in temps_tags:
    temps.append(tp.get_text())
temps
```

```
Out[43]: ['High: 95 °F',
'Low: 62 °F',
'High: 96 °F',
'Low: 64 °F',
'High: 96 °F',
'Low: 64 °F',
'High: 97 °F',
'Low: 64 °F',
'High: 100 °F']
```

```
In [45]: # Usando list comprehensions (não vimos ainda)
# desc = [d["title"] for d in seven_day.select(".tombstone-container img")]
```



```

descs_tags = seven_day.select(".tombstone-container img")
descs = []
for d in descs_tags:
    descs.append(d['title'])
descs

```

```

Out[45]: ['Today: Sunny, with a high near 95. Calm wind becoming west southwest 5 to 9
mph in the morning. ',
'Tonight: Mostly clear, with a low around 62. West wind 5 to 10 mph becoming
light and variable after midnight. ',
'Wednesday: Mostly sunny, with a high near 96. South wind 5 to 15 mph becomin
g west southwest in the afternoon. ',
'Wednesday Night: Partly cloudy, with a low around 64. West southwest wind 5
to 15 mph. ',
'Thursday: Mostly sunny, with a high near 96. Light south southwest wind beco
ming west 11 to 16 mph in the morning. ',
'Thursday Night: Mostly clear, with a low around 64. West wind 11 to 16 mph d
ecreasing to 5 to 10 mph after midnight. ',
'Friday: Sunny, with a high near 97. West wind 5 to 14 mph. ',
'Friday Night: Clear, with a low around 64. West wind 7 to 14 mph becoming so
uth southwest after midnight. ',
'Saturday: Sunny, with a high near 100. South wind 7 to 13 mph becoming west
in the afternoon. ']

```

## Pandas

Um DataFrame é um objeto que pode armazenar dados tabulares, facilitando a análise.

Vamos instanciar uma classe DataFrame e passar a lista de itens que temos. Vamos passar como parte de um dicionário. Cada chave do dicionário irá virar uma coluna no DataFrame e os elementos da lista irão se tornar os valores das colunas.

```

In [46]: import pandas as pd
df_clima_tempo = pd.DataFrame({
    "Período": periods,
    "Minidescrição": short_descs,
    "Temperatura": temps,
    "Descrição": descs
})
df_clima_tempo

```

```

Out[46]:

```

	Período	Minidescrição	Temperatura	Descrição
0	Today	Sunny	High: 95 °F	Today: Sunny, with a high near 95. Calm wind b...
1	Tonight	Mostly Clear	Low: 62 °F	Tonight: Mostly clear, with a low around 62. W...
2	Wednesday	Mostly Sunny	High: 96 °F	Wednesday: Mostly sunny, with a high near 96. ...
3	WednesdayNight	Partly Cloudy	Low: 64 °F	Wednesday Night: Partly cloudy, with a low aro...
4	Thursday	Mostly Sunny	High: 96 °F	Thursday: Mostly sunny, with a high near 96. L...
5	ThursdayNight	Mostly Clear	Low: 64 °F	Thursday Night: Mostly clear, with a low aroun...
6	Friday	Sunny	High: 97 °F	Friday: Sunny, with a high near 97. West wind ...
7	FridayNight	Clear	Low: 64 °F	Friday Night: Clear, with a low around 64. Wes...



	Período	Minidescrição	Temperatura	Descrição
8	Saturday	Sunny	High: 100 °F	Saturday: Sunny, with a high near 100. South w...

Com a tabela anterior, podemos fazer a análise do clima em uma cidade, incluindo as brasileiras.

### Desafio

Podemos utilizar [expressões regulares](#) e o método `Series.str.extract` para extrair o valor numérico das temperaturas. Esta [página](#) oferece uma introdução sobre o tópico.

```
In [47]: df_clima_tempo["Temperatura"]
```

```
Out[47]: 0      High: 95 °F
1      Low: 62 °F
2      High: 96 °F
3      Low: 64 °F
4      High: 96 °F
5      Low: 64 °F
6      High: 97 °F
7      Low: 64 °F
8      High: 100 °F
Name: Temperatura, dtype: object
```

Utilizamos o método `Series.str.extract`, que utiliza expressões regulares para extrair padrões do texto.

```
In [48]: # Neste caso, a expressão regular para extrair os dígitos da string
temp_nums = df_clima_tempo["Temperatura"].str.extract(r"(\d+)", expand=False)
temp_nums
```

```
Out[48]: 0      95
1      62
2      96
3      64
4      96
5      64
6      97
7      64
8     100
Name: Temperatura, dtype: object
```

```
In [23]: # O tipo resultante de Series.str.extract
type(temp_nums)
```

```
Out[23]: pandas.core.series.Series
```

```
In [50]: # Converte os valores da coluna temp_num em inteiro
df_clima_tempo["Temperatura"] = temp_nums.astype('int')
df_clima_tempo["Temperatura"]
```

```
Out[50]: 0      95
1      62
2      96
3      64
4      96
5      64
6      97
7      64
8     100
Name: Temperatura, dtype: int64
```

```
In [51]: df_clima_tempo
```

Out[51]:

	Período	Minidescrição	Temperatura	Descrição
0	Today	Sunny	95	Today: Sunny, with a high near 95. Calm wind b...
1	Tonight	Mostly Clear	62	Tonight: Mostly clear, with a low around 62. W...
2	Wednesday	Mostly Sunny	96	Wednesday: Mostly sunny, with a high near 96. ...
3	WednesdayNight	Partly Cloudy	64	Wednesday Night: Partly cloudy, with a low aro...
4	Thursday	Mostly Sunny	96	Thursday: Mostly sunny, with a high near 96. L...
5	ThursdayNight	Mostly Clear	64	Thursday Night: Mostly clear, with a low aroun...
6	Friday	Sunny	97	Friday: Sunny, with a high near 97. West wind ...
7	FridayNight	Clear	64	Friday Night: Clear, with a low around 64. Wes...
8	Saturday	Sunny	100	Saturday: Sunny, with a high near 100. South w...

Calcula a média para as temperaturas altas e baixas.

In [52]:

```
df_clima_tempo["Temperatura"].mean()
```

Out[52]:

82.0