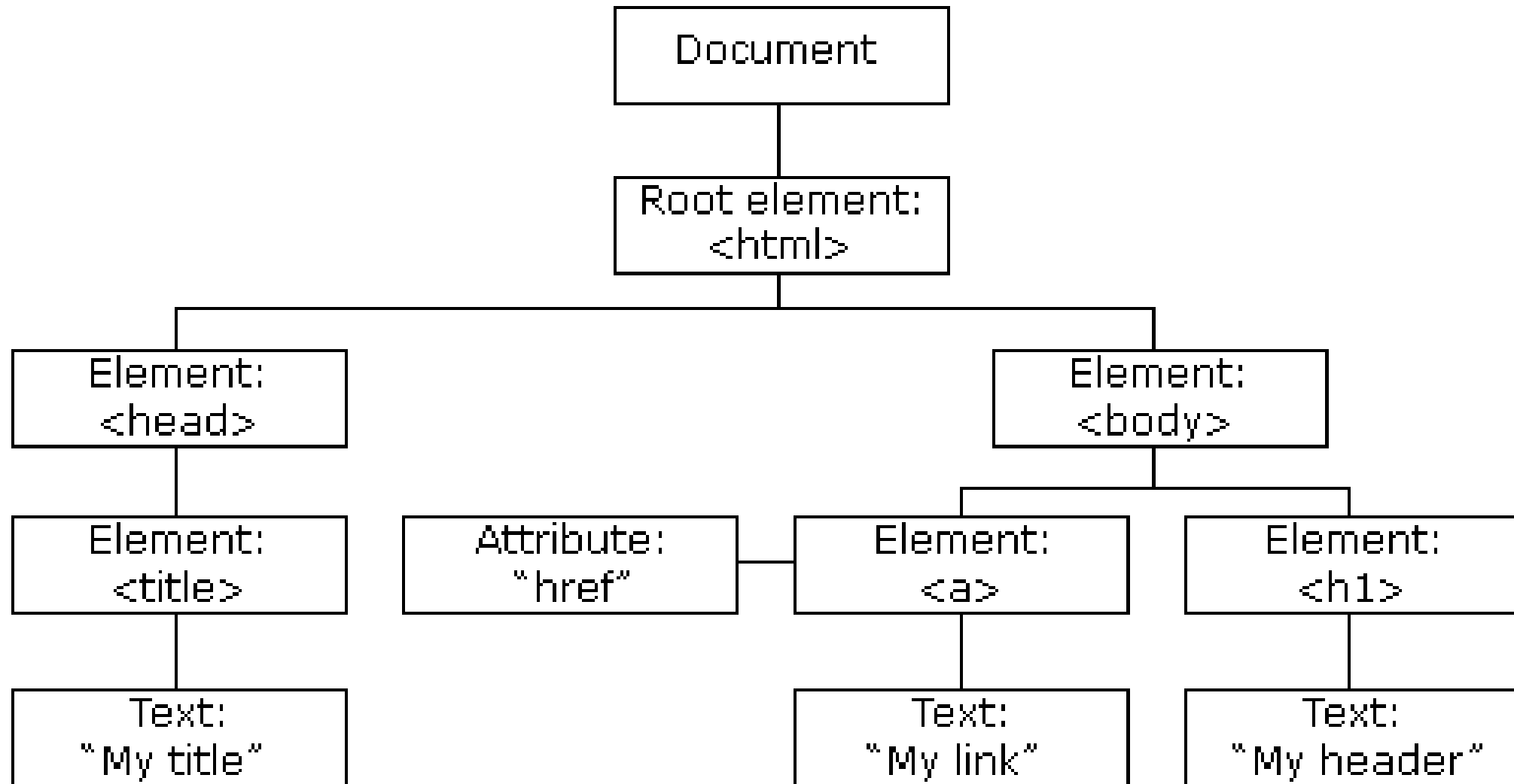


Pensamento Computacional e Lógica de Programação

Prof. Dr. Álvaro Campos
Ferreira

HTML e suas tags

DOM - Document Object Model



As principais tags

As principais tags (resumidamente):

- `<html>`
- `<head>`
- `<body>`
- `<p>`
- `<h1>`
- `<a>`
- ``
- `<div>`
- `` e ``

Argumentos

Algumas tags aceitam argumentos que contém informações necessárias como a localização de uma imagem ou link.

```
<a href="google.com">Google</a>
```

```

```

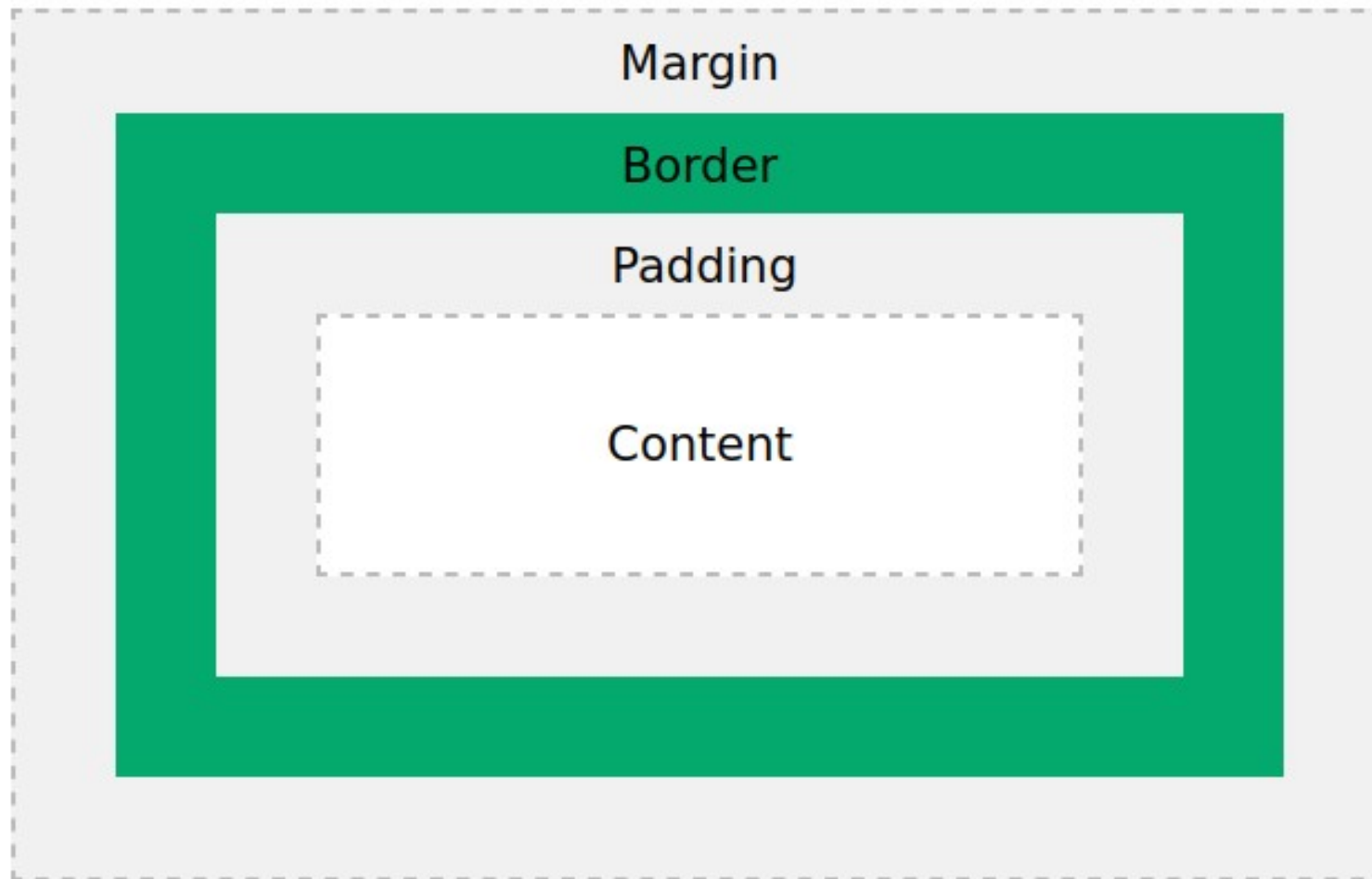
Deixando tudo melhor com CSS

Propriedades

Algumas das propriedades mais básicas:

- background
- width e height
- margin
- padding
- border
- color
- font
- text-align

Modelo caixa (box model)



Webscrapping

Webscrapping

É legal ou ilegal fazer webscrapping?

- É possível acidentalmente realizar muitos requests e isso é visto como um ataque
- Seu IP pode ser banido para algum site ou serviço

Webscrapping

É perfeitamente legal se feito de forma responsável!

- Robots.txt

<https://twitter.com/robots.txt>

- Requests
- Delays

Webscrapping

Existem dados em tabelas que podem ser utilizados diretamente pelo Pandas.

```
import pandas as pd  
url = 'https://pt.wikipedia.org/wiki/COVID-19'  
html = pd.read_html(url)
```

Webscrapping

Para selecionar apenas a tabela que queremos, usamos o argumento match na função read_html().

```
import pandas as pd
```

```
url = 'https://pt.wikipedia.org/wiki/COVID-19'
```

```
html = pd.read_html(url, match='Frequência')
```

Webscrapping

Como obter os dados da página que desejamos estudar?

```
import requests  
res = requests.get('https://www.google.com/')
```

Webscrapping

Esses dados não estão estruturados. Para estrutura-los, usamos BeautifulSoup.

```
from bs4 import BeautifulSoup  
soup = BeautifulSoup(res.text,"lxml")
```

Webscrapping

Dados obtidos da página estão na linguagem HTML, que divide os dados no que chama de tags.

Para explorar as tags de um site, usa-se a ferramenta de desenvolvimento do navegador, F12 no Firefox.

Webscrapping

Uma vez identificadas as tags de interesse, filtramos o resultado:

```
soup.findAll('div')  
for tag in soup.findAll('div'):  
    print(tag.text)
```



INSTITUTO BRASILEIRO DE ENSINO,
DESENVOLVIMENTO E PESQUISA