Live de Python #21

beautifulSoup / Web Scraping #2

Roteiro

- O que é o BeautifulSoup?
- Porque o BeautifulSoup é legal?
- Sintaxe básica
- CODE
  - Primeiros passos com uma página simples
  - Raspando dados do reclame aqui

O que é o BeautifulSoup?

“Desde 2004, tem economizado horas ou dias de trabalho de desenvolvedores em projetos de raspagem de dados”

Três coisas tornam o BS poderoso:

1. Métodos simples e pythonicos para parsear árvores (html, xml)
2. Converte tudo para unicode (UTF-8)
3. Permite diferentes ferramentas para parsear (lxmx, html5lib, html.parse)

Porque o BeautifulSoup é legal?

- Encontra elementos por tags (a, b, p, div)
- Combina ela com os dados agrupados com tags (classes, por exemplo)
- Retona as buscas em novos objetos BS
- Corrige falhas no html/xml
- Transforma as strings em objetos (obj.title)
- FACILITA A VIDAAAAAAA

Sintaxe básica (o que realmente vai ser útil hoje)

```
from bs4 import BeautifulSoup

page = BeautifulSoup(<string>, <parse>)

page.title  # teste

# Encontra a primeira tag a na string
page.find('a')

# Encontra todas as tags 'p' com uma classe
page.find_all('p', class_='<classe>')
```

```
.find() # busca um elemento

.find_all() # busca todos os elementos

.find_parent() # busca um elemento e o bloco

....
```

CODE !!