Web Scrapping com Python

Cristian Muñoz Villalobos

Sumario

Parte 1: Seu primeiro web scraper

- Conectado-se
- Introduçã ao BeautifulSoup

Parte 2: Análise de HTML avançada

- find e findAll
- Lidar com filhos , descendentes e pais

Exemplo 1

Exemplo 2

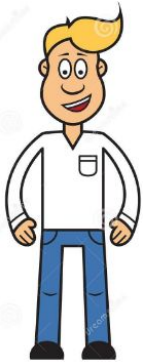
Conectando-se

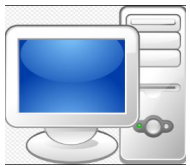
Que acontece quando abrimos o navegador e escrevemos www.google.com ?

Web scraping: camada de interface (

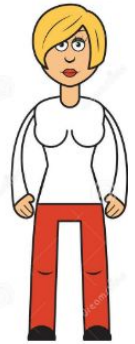
- nível do navegador (HTML, CSS e JavaScript) e
- nível de conexão de rede)

Bob





Alice



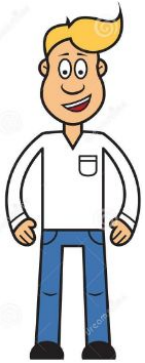
Conectando-se

Que acontece quando abrimos o navegador e escrevemos www.google.com ?

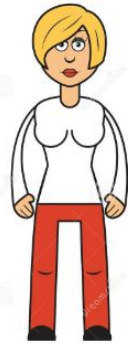
Web scraping: camada de interface (

- nível do navegador (HTML, CSS e JavaScript) e
- nível de conexão de rede)

Bob



Alice

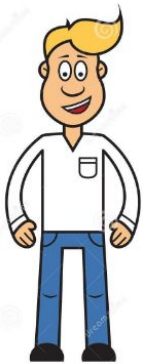


Conectando-se

Que acontece quando abrimos o navegador e escrevemos www.google.com ?

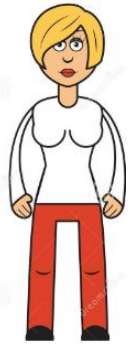
Web scraping: camada de interface (

- nível do navegador (HTML, CSS e JavaScript) e
- nível de conexão de rede)



Roteador local

Alice



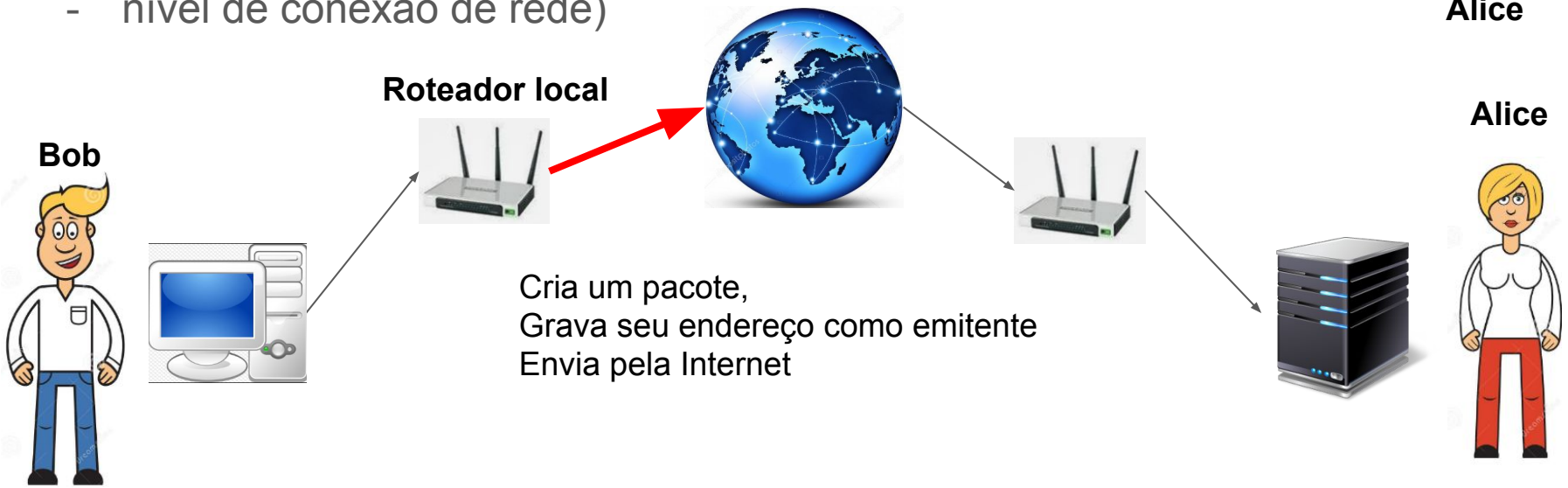
Cabeçalho:
Endereço imediato: MAC do roteador local
Endereço Final: IP da máquina de Alice
Corpo:
Solicitação feita ao servidor de Alice

Conectando-se

Que acontece quando abrimos o navegador e escrevemos www.google.com ?

Web scraping: camada de interface (

- nível do navegador (HTML, CSS e JavaScript) e
- nível de conexão de rede)

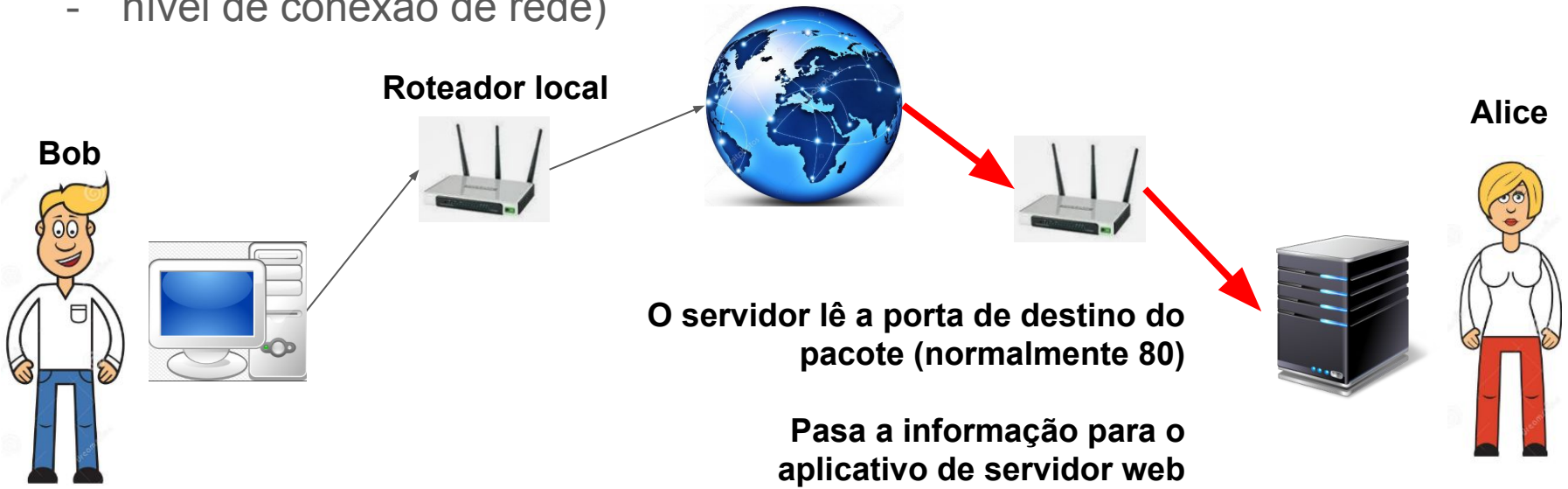


Conectando-se

Que acontece quando abrimos o navegador e escrevemos www.google.com ?

Web scraping: camada de interface (

- nível do navegador (HTML, CSS e JavaScript) e
- nível de conexão de rede)

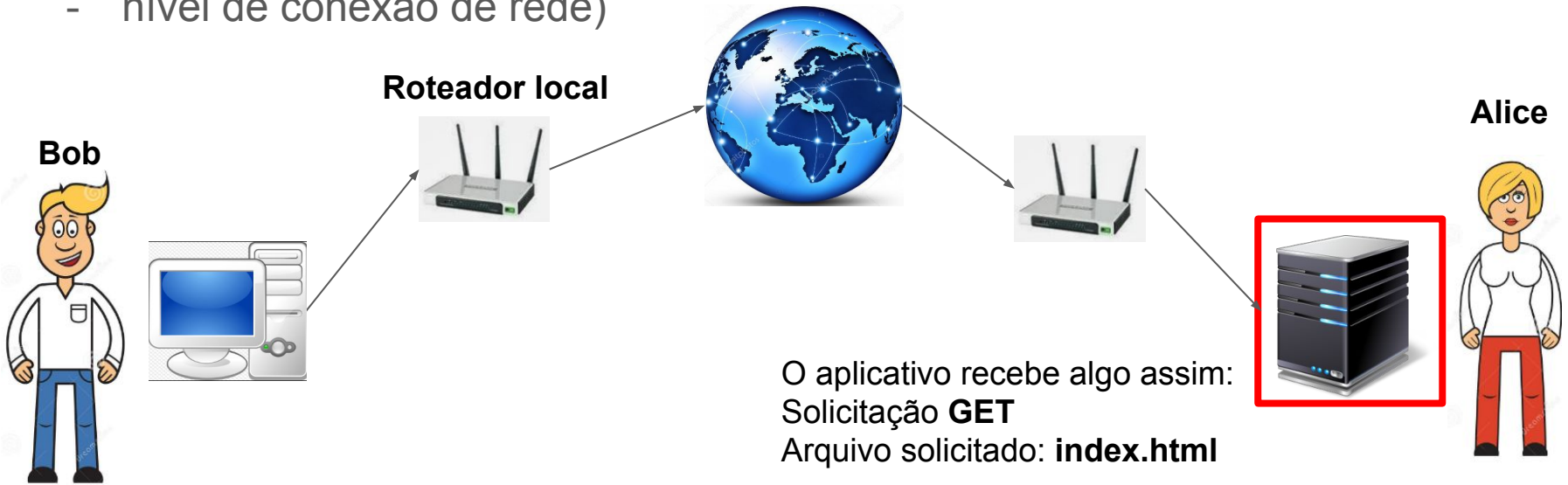


Conectando-se

Que acontece quando abrimos o navegador e escrevemos www.google.com ?

Web scraping: camada de interface (

- nível do navegador (HTML, CSS e JavaScript) e
- nível de conexão de rede)



Conectando-se

Que acontece quando abrimos o navegador e escrevemos www.google.com ?

Web scraping: camada de interface (

- nível do navegador (HTML, CSS e JavaScript) e
- nível de conexão de rede)

Introduçã ao BeautifulSoup

O nome da biblioteca BeautifulSoup vem de um poema de mesmo nome de Lewis Carroll encontrado em Alice’s Adventures in Wonderland, cantado por Mock Turtle.

"Linda Sopa, tão rica e verdinha,
Assentada em uma quente terrina!
Quem não se entregaria a tamanha iguaria?
Sopa noturna, sopa tão linda!"

Instalação: pip install beautifulsoup4

Análise de HTML avançada

Ao perguntar a Michelangelo como ele conseguiu esculpir uma obra de arte tão majestosa como seu Davi, dizem que ele respondeu: “Fácil. É só remover as partes de pedra que não se parecem com Davi.” (***It is easy. You just chip away the stone that doesn’t look like David.)***

Análise de HTML avançada

find e findAll

findAll(tag, attributes, recursive, text, limit, keywords)

find(tag, attributes, recursive, text, keywords)

Lidar com filhos e descendentes

….

Lidar pais

….

Exemplo 1:

Baixar os dados da tabela de ranking de filmes fornecidos no site:
http://www.imdb.com/chart/moviemeter?ref_=nv_mv_mpm_8

Info: Title, Ano, Ranking,....

Exemplo 2

?