Capítulo 11 - Web Scraping

Douglas Almeida

25 de Maio de 2017

Resumo

Neste capítulo do livro *Automate the Boring Stuff with Python* de Al Sweigart é introduzida a possibilidade de usar um programa Python para baixar uma página da internet e processar o conteúdo da mesma – procedimento chamado de *web scraping*. A fim de exercitar os novos conhecimentos adquiridos foram desenvolvidos três programas com base nos exemplos apresentados pelo autor do livro ao mesmo tempo em que conceitos aprendidos em capítulos anteriores, como importação de módulos e manipulação de arquivos foram revisitados.

1 Recursos utilizados

Todos os módulos adicionais necessários foram baixados e instalados usando a ferramenta pip3 do Python.

- Módulo *webbrowser*, para uso da função:
  - *webbrowser.open()* – Abre uma página da internet no navegador padrão do computador a partir de seu endereço fornecido através de uma *string*.
- Módulo *pyperclip*, para uso da função:
  - *pyperclip.paste()* – Recupera o conteúdo textual armazenado na área de transferência do sistema operacional.
- Módulo *requests*, para uso das funções:
  - *requests.get()* – Transfere o conteúdo de uma página da internet para um objeto do tipo *requests.models.Response*. O endereço da página é fornecido como uma *string*.

  - *raise_for_status()* – Método de um objeto *requests.models.Response* invocado para gerar uma exceção durante a execução do programa caso a transferência da página tenha sido mal sucedida.
  - *iter_content()* – Método de um objeto *requests.models.Response* invocado para iterar ao longo do conteúdo da página transferida alguns bytes de cada vez.

- Módulo *Beatiful Soup (bs4)*, para uso das funções:
  - *bs4.BeautifulSoup()* – Cria um objeto que consiste na representação de uma página transferida da internet.
  - *select()* – Método de um objeto *bs4.BeautifulSoup* invocado para gerar uma lista com todas as intâncias de uma determinada etiqueta HTML presentes na página transferida.

2 Programas criados

Os programas foram disponibilizados em arquivos individuais com extensão *.py* e acompanham este resumo. Os arquivos podem ser executados sem nenhuma modificação através de um terminal de texto nos sistemas operacionais Linux ou macOS desde que a versão 3.0 ou superior do Python e os módulos adicionais tenham sido devidamente instalados.

2.1 mapIt

Mostra no Google Maps um endereço fornecido via linha de comando ou obtido da área de transferência.

2.2 fileDownloader

Armazena no computador uma cópia de um arquivo presente na internet. Os argumentos são a URL do arquivo remoto e o caminho (incluindo nome e extensão) do arquivo a ser criado no computador com o conteúdo do arquivo da remoto.

2.3 latestNews

Busca as últimas notícias publicadas pelo site Inovação Tecnológica e as exibe, individualmente, usando o navegador de internet padrão.