PROJETO 1 – GITHUB

(AlexandreRodel/README.md at main · alerodel/AlexandreRodel · GitHub)

Nome do Projeto

WebscrapingProj1

Autor
Alexandre Rodel de Almeida

Objetivo

Ler notícias da página inicial do site G1 de forma estática usando a biblioteca “Beatifulsoup” no Python e armazená-las para que as informações possam ser consultadas posteriormente.

Introdução

O projeto está dividido em partes e será executado necessariamente da seguinte forma:

- entendimento de como funciona a “inspeção” de um site para entender quais são as estruturas que usaremos;
- entendimento de como “comunicar” com o site e qual biblioteca usar para isso;
- entendimento de como será feito o encapsulamento das informações retidas do site e como serão manipuladas dentro do Python;
- entendimento de como será realizado o armazenamento do arquivo e a comunicação com os bancos de dados utilizados;
- por fim, pretende-se usar um trecho de código que verifique quais são as palavras mais repetidas dentro do arquivo de notícias e apresentar um gráfico de barras e/ou hierárquico radial com as palavras mais repetidas.

Desenvolvimento do Projeto

1) Entendendo o site, inspecionando, comunicando e realizando uma raspagem

a) primeiro entramos no site e verificamos como são as estruturas das notícias, notaremos que em geral elas possuem um Header, um Body e um Link. Então podemos ir direto na “div” que contém a classe que precisamos e retiramos a informação de cada notícia de forma similar;

b) no Python podemos usar a biblioteca “requests” para comunicar com o site e verificar o status dessa comunicação. Se for Status 200 significa que está Ok;

c) para fazer uma raspagem simples usamos a biblioteca “BeatifulSoup”. Com ela é possível fatiar as informações e códigos contidos no site a partir de certos trechos (as div);

d) aqui, o Pandas é usado caso queira criar uma lista vazia e preenchê-la com informações obtidas da raspagem no site de forma organizada, facilitando a manipulação dessas informações no futuro;

e) após a raspagem e a organização é possível salvar o arquivo em formato .csv ou enviá-lo para um banco de dados (relacional ou não-relacional);

2) Após o armazenamento de dados dentro da base de dados ou do arquivo .csv

a) agora, já é possível fazer alguma análise com base nos dados armazenados. Um exemplo de estudo possível de se fazer é uma contagem de palavras que se repetem e posteriormente fazer um gráfico, tal como:
- treemap
- packed circles
- sunburst
- hierarquical bar
- radial

b) no caso das palavras que mais se repetem, é possível que ográfico hierárquico radial seja a melhor opção para um site de notícias.

Na Prática

Cada passo desse projeto está contido na pasta “WebscrapingProj1” no Github.

Conclusão

Era esperado nomes de personalidades políticas estivessem entre as palavras que mais se repetem, bem como conectivos de frases, artigos e outros elementos de textos.

Após analisar os resultados mostrados nos gráficos, surgiu a ideia de inserir algumas melhorias interessantes. São elas:
- incluir a data e hora das notícias na raspagem de dados;
- após gravar os dados brutos, realizar uma limpeza eliminando notícias com informações incompletas (sem estar com todas as colunas preenchidas). Embora a escolha de realizar a tarefa da forma como foi feita tenha vindo do fato de a minoria das virem com algum campo vazio;
- eliminar palavras que não precisaremos analiasr, deixando apenas palavras que fazem sentido ser analisadas.

Em geral, esse projeto foi útil porque nos ajudou a entender como funciona uma página de notícias e sua dinâmica. Algumas das etapas e parte do código que usamos aqui podem ser usadas para realizar raspagem de dados em sites de E-commerce e em outras sites e situações similares. Um próximo passo talvez envolva o uso de uma biblioteca que acompanhe o dinamismo de determinadas páginas de internet, como é o caso da biblioteca Selenium do python.