Captura

A parte de captura dos dados foi realizada através da implementação de um código em python com algumas bibliotecas auxiliares. O script tem 178 linhas (incluindo comentários) e funciona da seguinte forma:

1. Dada a url de busca
   https://www.in.gov.br/consulta/-/buscar/dou?q="deferir+os+registros+e+as+petições+dos+produtos+saneantes"&s=todos&exactDate=personalizado&sortType=0&publishFrom=01-05-2021&publishTo=30-06-2021

   é realizado o mapeamento dos links de cada publicação resultante.
2. Posteriormente, é realizada a coleta dos dados de cada página de publicação advinda do passo anterior. Essa coleta se dá pela leitura do html de cada página.
3. Os dados coletados são utilizados para a construção de um dataframe que é, então, gravado em um arquivo excel (.xlsx).

No passo 1, é utilizado a url resultante da busca pelas publicações no Diário Oficial da União pelo termo de busca "deferir os registros e as petições dos produtos saneantes" no período entre 01/05/2021 e 30/06/2021. O resultado são 9 publicações que são passadas para a coleta, sendo elas:

- RESOLUÇÃO RE Nº 2.495, DE 24 DE Junho DE 2021 (28/06/2021)
- RESOLUÇÃO RE Nº 2.404, DE 17 DE Junho DE 2021 (21/06/2021)
- RESOLUÇÃO RE Nº 2.314, DE 10 DE Junho DE 2021 (14/06/2021)
- RESOLUÇÃO RE Nº 2.186, DE 2 DE Junho DE 2021 (07/06/2021)
- RESOLUÇÃO RE Nº 2.115, DE 27 DE Maio DE 2021 (31/05/2021)
- RESOLUÇÃO RE Nº 2.037, DE 20 DE Maio DE 2021 (24/05/2021)
- RESOLUÇÃO RE Nº 1.951, DE 13 DE Maio DE 2021 (17/05/2021)
- RESOLUÇÃO RE Nº 1.842, DE 6 DE Maio DE 2021 (10/05/2021)
- RESOLUÇÃO RE Nº 1.754, DE 29 DE Abril DE 2021 (03/05/2021)

A implementação do script durou um total de 4 horas, incluindo prototipação e reorganização do código. A execução do script dura cerca de 12.5 segundos, mas esse tempo pode variar de acordo com a velocidade de resposta apresentada pelo acesso às páginas.

O script foi implementado de forma a ser facilmente reaproveitado, sendo necessário apenas a alteração do link de pesquisa. Portanto, a alteração do termo de busca e das datas das publicações pode ser realizada de forma prática.

Ferramentas Utilizadas

- Python 3
  - Selenium
  - BeautifulSoup
  - Pandas

Transformação

Os dados coletados são gravados diretamente em uma planilha excel (.xlsx) de acordo com o layout apresentado. Antes da gravação, os dados são transportados para um dataframe, o que facilitaria uma possível transformação ou visualização durante a coleta.

A seguir, apresento possíveis sugestões de tratamento de transformação e/ou qualidade dos dados coletados:

- Possível alteração do formato do campo *Vencimento* para "<mês por extenso> de <ano>", de acordo com a preferência do cliente.
- No caso de ser uma planilha de uso imediato e por tempo curto, substituir os valores do campo *Vencimento* por valores que indiquem o prazo até a data de vencimento em meses e/ou anos.
- Ajustar o campo *Validade Produto* para conter todos os valores em meses (ou anos).
- No campo *Categoria*, retirar o código da categoria que precede o nome da categoria.
- No campo *Assunto Petição*, retirar o código do assunto que precede o nome do assunto.
- Substituir o campo *Resolução* desmembrando-o em *Número Resolução* e *Data Resolução*.