Desafio RPA  
Kauanna Veira Silva

06/07/2021

# Visão geral

## Descrição do projeto

Este projeto tem como objetivo criar uma web scraping, que extraia do site dos correios, as informações de localidade das faixas de CEP informadas mediante base de dados enviada via avaliador. Todos os dados encontrados são coletados e armazenados em formato *.xlsx*.

## Estrutura do projeto

* DesafioRPA – Nome da solução e nome do projeto;
* Helper
  + HelperData – Classes com propriedades e métodos auxiliares para melhor organização do código
    - Propriedade
      * directoryPath – Armazena o diretório principal do projeto;
      * fileNameSouce – Caminho da planilha enviada como base de dados;
      * fileNameTarget – Caminho da planilha do resultado final;
      * fileNameLog – Caminho do arquivo de log principal
      * fileNameCepResquest – Caminho do arquivo de log onde ficará armazenado os números de cep utilizados em cada requisição
    - Métodos
      * Log – Método que recebe o nome do arquivo e a mensagem que deve ser armazenada;
      * WriteExceptionLog – Método auxiliar que rebe o tipo de exceção e a mensagem da exceção gerada.
      * ExtractData – Método que ler a planilha enviada como base da dados e retorna uma lista povoada com os dados extraídos;
      * RemoveDuplicateCep – Método que recebe uma lista e devolve-a ordenada e sem duplicações;
      * CreateResultFile – Cria o arquivo *.xlsx* solicitado;
      * WriteResultFile - Escreve no arquivo de resultado as informações extraídas do site dos correios;
  + InfoLocation – Classe para estruturar as informações extraída do site;

## Funcionamento do robô

Para implantação do robô basta apenas dar duplo clique do arquivo dentro do diretório do projeto ..\bin\Debug\net5.0\DesafioRPA.exe. Ao executar o robô irá inicialmente é verificado de se já existe os arquivos de log caso exista esses são removidos, para não ocorrer nenhum. Depois disso o arquivo de resultado é criado. Feito isso o robô está pronto para começar a registrar as rotinas realizadas, afim de armazenar tudo aquilo que está sendo feito e as exceções que podem vir a acontecer.

Após este processo de manipulação de arquivos, o robô começa a extrair as informações da planilha armazenando a mesma em uma lista de dois inteiros, onde estarão presentes a faixa inicial e final do Cep. Logo após a extração é realizada a ordenação e remoção das duplicações percebidas ao analisar a planilha enviada. Este processo se faz necessário, devido ao fato de a lista possuir várias faixas repetidas, o que ocasionaria em uma maior quantidade de dados a serem processados.

Para manipulação da planilha foi utilizado um pacote que bom desempenho nesta tarefa, no entanto para ordenar uma planilha com esta quantidade de arquivos este pacote demora entorno de 5 min o que não é viável para este projeto, para melhorar a performance desse processo, as manipulações foram realizadas em uma lista de dois inteiros reduzindo assim para em média de 18s para tal ação.

Quando os dados estão prontos para serem avaliados o robô realiza uma estrutura de repetição para ler cada faixa de Cep. O cep ele possui uma estrutura fixa formada de 5 dígitos que definem a região, sub-região e etc. do endereço especificado, acompanhado de 3 dígitos referentes ao sufixo, partindo desta informação cada faixa terá um intervalo dos 5 primeiros dígitos da faixa inicial até os 5 primeiros dígitos da faixa final, e cada um desses intervalos vão possuir de 0 a 999 números de possíveis CEPs.

Devido a quantidade de informações requisitadas se fez necessários forçar com 3 tentativas a reconexão em caso de serviço indisponível, no entanto, ainda pode não ser o suficiente devido a alguns fatores: falta de conexão com a internet, lentidão no acesso a internet, crash de memória, em caso do PC que o robô está sendo executado não possuir memória suficiente disponível para executá-lo. Este robô ocupa em média 55MB de memória consumindo 2,4% da CPU.

Cada CEP é avaliado individualmente no site dos correios, simulando uma interação do usuário, digitando o número do CEP no local indicado e clicando em buscar, o site dos correios apresenta as informações respectivas ao CEP informado, logradouro, cidade, estado e cep em caso de o CEP não existir o site não apresenta nada. Quando é encontrada informações o robô extrai do site estas informações e salvo do arquivo especificado para esta ação.

## Características técnicas

Para que o robô funcione foram utilizados alguns pacotes externos, todos eles disponíveis no gerenciador de pacotes *NuGet*:

* EPPlus – Este pacote possibilita uma manipulação mais robusta de planilhas Excel, ele possui um melhor desempenho se comparado ao Microsoft.Office.Interop.Excel e é de fácil utilização. Outra característica que se faz essencial a utilização desde pacote é que ao utiliza-lo a maquina que estiver rodando o robô não tem a necessidade de ter o Office Excel instalado.
* Selenium – Este pacote é um ambiente integrado de desenvolvimento para scripts de testes automatizados permitindo facilmente e rapidamente que seja gravado e reproduzido teste em ambiente real, uma vez que ele permite simular o comportamento do usuário utilizando o navegador. Sua sintaxe é bem fácil de usar o que o torna muito popular.
* Selenium WebDriver ChormeDriver– Este pacote permite a escrita de forma mais produtiva e organizada. Ele faz chamadas diretamente ao navegador utilizando o suporte à automação nativa de cada navegador. Desta forma, seus testes se tornam bastante realistas, além de ter suporte para praticamente todos os navegadores. No entanto neste projeto foi utilizado o ChromeDriver por preferência pessoal.

## Dificuldades

O site fornecido como url base do robô, https://www.buscacep.correios.com.br/sistemas/buscacep/buscaFaixaCep.cfm, não está mais viável de utilização. Pois há instabilidade no site informado sendo retornado em sua maioria os resultados Request timed out ou 404 not found, tornando inviável utiliza-lo. Como não houve nenhuma restrição de utilizar o site oficial dos correios ativo no momento, a url base foi modificada para https://buscacepinter.correios.com.br/app/endereco/index.php

Outro ponto importante foi que no site atual dos correios não tem mais uma pesquisa por faixa de cep, sendo então necessário fazer uma busca com uma quantidade de registro enorme. Como este acesso depende do servidor dos correios, sendo este um órgão que não pode haver ataques que geram indisponibilidade, devido a isso não será possível entregar de forma satisfatória os resultados.