

Requires support

spec/support

run as spec

in _spec.rb

run twice. It

run with _spec.

end with _spec.

end with _spec.

and only on a spec.

and for mongoid



Texto base

9

Web scraping

Osvaldo Kotaro Takai & Ana Cristina dos Santos

Resumo

O objetivo desta parte é aprender a extrair informações de páginas web (scalping).

1. Introdução

Web scraping é uma técnica usada para extrair dados de sites por meio de um processo automatizado. Essa técnica pode ser utilizada para atender a vários propósitos, desde coletar a cotação do dólar em reais até para aplicações mais complexas como o monitoramento de determinadas ações na bolsa de valores.

Esta lição pretende exemplificar a técnica e não necessariamente desenvolver uma aplicação que possa agregar valor a alguém. O objetivo é passar a noção de como as informações podem ser extraídas de uma página web. Para tanto, será utilizada a seguinte página do governo brasileiro:

 $\underline{\text{http://blog.politicos.org.br/fundao-eleitoral-quem-foram-os-263-deputados-que-aprovaram-aumento/}\\$

Ranking dos Políticos Início Vídeos Entrevistas Acesse o Ranking Contato Ranking na Mídia Fatos e dados > Marcelo Álvaro: como foi o mandato do Ministro do Turismo de Bolso Quem foram os 263 deputados que alteraram (pra pior) as regras eleitorais O QUE VOCÊ PROCURA? Quem foram os 263 deputados que alteraram (pra pior) as regras eleitorais 4 SETEMBRO 2019 | RANKING DOS POLÍTICOS LEIA MAIS A Câmara dos Deputados aprovou, por 263 a 144, projeto de lei que abre caminho para ampilar o bilionário fundo eleitoral, recría a propaganda partidária na TV (que foi extinta justamente para custear as campanhas políticas), autoriza o uso de dinheiro público para pagar advogados de investigados e institui regras mais flexíveis para contratação de dirigentes partidários. Indicação de Kássio Nunes ao STF: análise da Confira abaixo a lista com os nomes dos deputados que votaram A FAVOR e CONTRA o projeto. Veja também a lista organizada por estados, e confira os nomes e votos dos parlamentares do seu estado. Confira quem são os 125 parlamentares que querem concorrer às eleições municipais; maioria vai mal no Ranking dos Políticos Deputados que votaram A FAVOR do afrouxamento das regras eleitorais: Confira o Ranking dos estados de acordo com a NOME PARTIDO ESTADO DEM AC Jéssica Sales Perpétua Almeida PCdoB

Figura 1.1 - Página web usada para extrair dados.

Fonte: Blog Ranking dos Políticos, 2021.

O bot irá (Figura 1.2):

- 1) Abrir o navegador na página web acima.
- 2) Extrair:
 - a) O título da matéria.

Dra. Vanda Milani SOLIDARIEDADE AC

- b) A tabela de deputados que votaram a favor do afrouxamento das regras eleitorais contendo três colunas: Nome, Partido e Estado.
- 3) Salva o conteúdo dessa tabela num arquivo CSV.

11 - Obter dados da Web

3
Quem foram os 263 deputados que alteraram (pra pior) as regras eleitorais

Figura 1.2 - Esquema do Bot.

Fonte: do autor, 2021.

2. Preparação

Antes de iniciar a criação do bot, recomenda-se que o tamanho do texto seja 100% e que a resolução de tela seja selecionada na opção marcada como (Recomendável). Importante manter essas mesmas configurações durante durante a criação do bot e durante a sua execução.

3. Criação do Bot

Desta vez, o bot será criado linha a linha para demonstrar a captura individual de ações do recorder.

3.1. Abrir página web

Será usada a ação **Browser: Launch website** para abrir a página web usando o navegador **Chrome** (Figura 2.1).

Figura 2.1- Configuração da ação Browser: Launch website.



Fonte: Control Room do Automation 360 CE.



No campo URL deve ser digitar o endereço da página que deverá ser aberta com um parâmetro "--incognito" sem as aspas no final do endereço:

http://blog.politicos.org.br/fundao-eleitoral-quem-foram-os-263-deputados-que-aprovaram-aume nto/ --incognito

Esse parâmetro permite que o Chrome seja aberto numa nova janela com perfil anônimo.

Salve o bot e execute-o. O bot deverá abrir uma nova janela do Chrome na página web similar ao exibido na Figura 2.2.



Figura 2.2 - Página web aberta com a opção "--incognito".

Fonte: Blog Ranking dos Políticos, 2021.

3.2. Wait for window

Esta ação garante que a página web tenha sido aberta antes de executar a próxima ação. Logo após a ação ser arrastada deve-se:

- 1) Selecionar Application;
- 2) Atualizar o combobox; e
- 3) Selecionar o navegador Chrome com a página anteriormente aberta.
- 4) Salvar a configuração.

A figura 2.3 ilustra os detalhes desta configuração.



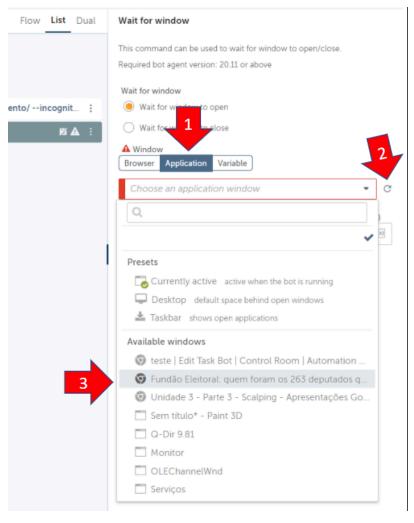


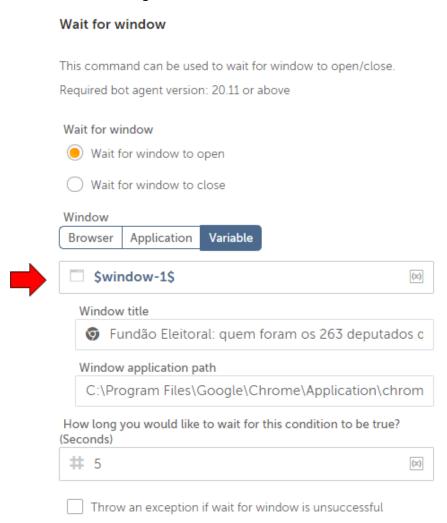
Figura 2.3 - Configuração da ação Wait for window.

Fonte: Control Room do Automation 360 CE.

Após salvar a configuração, o bot terá criado a variável **window-1** e que foi associada à ação (Figura 2.4).



Figura 2.4 - Variável window-1 criada



Fonte: Control Room do Automation 360 CE.

3.3. Simulate keystrokes

O objetivo agora é posicionar a página para que ela possa exibir todos os itens que se deseja capturar. Esses itens estão próximos ao texto: "Deputados que votaram A FAVOR do afrouxamento das regras eleitorais:".

Assim, esta ação irá simular a busca desse texto na página web que é representada, agora, pela variável **window-1**.

Normalmente, para fazer essa busca, utiliza-se a tecla de atalho Ctrl+F; entretanto, o bot não reconhece essa tecla de atalho. Por isso, a tecla F3 foi utilizada como alternativa Ctrl+F, pois possui a mesma finalidade.

A Figura 2.5 exibe a configuração desta ação.



Select window Browser Application Variable (x) \$window-1\$ Window title Fundão Eleitoral: quem foram os 263 deputados que aprovaram aumento - Google Chro Window application path C:\Program Files\Google\Chrome\Application\chrome.exe Resize window May improve bot accuracy Keystrokes Enter keystrokes here or use the on-screen keyboard 99 [F3] Deputados que votaram A FAVOR do afrouxamento das regras eleitorais: (x) Delay between each keystroke in ms (optional) # 10 A value greater than 2 is recommended.

Figura 2.5 - Configuração da ação Simulate keystrokes.

Fonte: Control Room do Automation 360 CE.

O passo 1 associa a variável **window-1** à ação e o passo 2 insere [F3] (equivalente ao Ctrl+F) seguido do texto sem as aspas: "Deputados que votaram A FAVOR do afrouxamento das regras eleitorais:".

Para verificar se está tudo funcionando, é importante salvar a configuração; fechar a janela do Chrome aberta anteriormente executar o bot.

Após a execução, o aspecto da janela aberta do Chrome aberta será similar ao exibido na Figura 2.6.



Quem foram os 263 deputados que alteraram (pra pior) as regras eleitorais 4 SETEMBRO 2019 | RANKING DOS POLÍTICOS A Câmara dos Deputados aprovou, por 263 a 144, projeto de lei que abre caminho para ampliar o bilionário fundo eleitoral, recria a propaganda partidária na TV (que foi extinta justamente para custear as campanhas políticas), autoriza o uso de dinheiro público para pagar advogados de investigados e institui regras mais flexíveis para contratação de dirigentes partidários. Indicação de Kássio Nunes ao STF: análise da Celso Russomanno: como foi o mandato do polêmico deputado federal Por estado: os 42 senadores que votaram para permitir aumentos para o funcionalismo durante a crise Confira abaixo a lista com os nomes dos deputados que votaram A FAVOR e CONTRA o projeto. Veja também a lista organizada por estados, e confira os nomes e votos dos parlamentares do seu estado. Confira quem são os 125 parlamentares que querem concorrer às eleições municipais; maioria vai mal no Ranking dos Políticos Confira o Ranking dos estados de acordo com a pontuação de seus parlamentares NOME PARTIDO ESTADO Jéssica Sales MDB PCdoB Perpétua Almeida Dra. Vanda Milani SOLIDARIEDADE Isnaldo Bulhões Jr. MDB Arthur Lira

Figura 2.6 - Aspecto página web aberta até a linha 3.

Fonte: Blog Ranking dos Políticos, 2021.

3.4. Recorder: Capture (capturar título da matéria)

O objetivo agora é extrair o texto "Quem foram os 263 deputados que alteraram (pra pior) as regras eleitorais" da página aberta.

Para tanto, deve-se:

- 1) Figura 2.7: Associar a variável window-1 à ação e salvar; e
- 2) Figura 2.7: Selecionar Capture object.



Application Variable

Swindow-1\$

Window title

Fundão Eleitoral: quem foram os 263 deputados que aprov

Window application path

C:\Program Files\Google\Chrome\Application\chrome.exe

Capture object

Cick "Capture object" to choose an object from the window.

Wait for control

15

Assign the output to variable (optional)

Choose a variable

(x)

Choose a variable

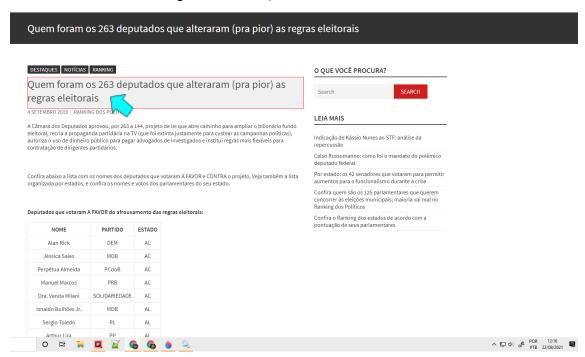
Figura 2.7 - Configuração inicial da ação Recorder: Capture.

Fonte: Control Room do Automation 360 CE.

Ao acionar a captura do objeto (**Capture object**), a janela do Chrome será exibida. É importante movimentar o ponteiro do mouse sobre a página até que itens da página comecem a ser contornados em vermelho para então, posicionar o ponteiro do mouse sobre o texto (título da matéria) que se deseja selecionar e aguardar até que ele seja completamente contornado por uma linha vermelha para dar um clique (Figura 2.8).



Figura 2.8 - Seleção do título da matéria.



Fonte: Blog Ranking dos Políticos, 2021.

Após o clique da seleção do texto, a configuração da ação será similar ao da Figura 2.9.



Recorder: Capture (disabled) This command can be used to simulate a object's click event. Object detail Browser Application Variable swindow-1\$ (x) Window title Fundão Eleitoral: quem foram os 263 deputados que apr Window application path $C:\label{lem:condition} C:\label{lem:condition} C:\l$ Recapture object Preview NOTÍCIAS RANKING ram os 263 deputados que alteraram (pra pio eitorais RANKING DOS POLÍTICOS Cannot be empty Action N Get all children names Get all children values Set text Click Left click Right click Double click zuem jorani os zoo ueputados due Wait for control # 15

Figura 2.9 - Configuração da ação Recorder: Capture após a captura do Título.

Fonte: Control Room do Automation 360 CE.

Continuando a configuração da ação, deve-se selecionar no combobox a ação **Get property**. Com isso, abrirá um novo campo, **Property name**, que permite indicar o nome da propriedade que contém a informação que se deseja capturar (Figura 3.1).



Recorder: Capture (disabled) provider application Telegraph \$window-1\$ Window title Fundão Eleitoral: quem foram os 263 deputados que apr Window application path C:\Program Files\Google\Chrome\Application\chrome.exe Recapture object Preview NOTÍCIAS RANKING ram os 263 deputados que alteraram pra pio eitorais Action Get property Run in the background A Property name (x) Object properties (3 of 16) Value Name Control Type CLIENT Technology Type MS_ACTIVE_ACCESSIBILITY 99 Quem foram os 263 deputados que a Name Wait for control

Figura 3.1 - Configuração da ação Recorder: Capture após seleção Get Property.

Fonte: Control Room do Automation 360 CE.

Para descobrir qual é o nome dessa propriedade, deve-se abrir a opção **Object properties (3 of 16)** (Figura 3.2).



Recorder: Capture (disabled) Run in the background Property name (x) ▼ Object properties (16) Name Value Control Type CLIENT ✓ Technology Type MS_ACTIVE_ACCESSIBILITY Path 77 Description Left **99** 374 Parent 99 Class Name 99 Chrome_WidgetWin_1 Index 99 2 Default Action 99 Quem foram os 263 deputados q ✓ Name States Тор **99** 203 Value Height 99 87 □ ID Width 99 771

Figura 3.2 - Configuração da ação Recorder: Capture após abrir a opção Object properties (3 of 16).

Fonte: Control Room do Automation 360 CE.

Rolando-se a tela, verifica-se que a propriedade que possui o valor desejado é Name. Assim, deve-se digitar "Name" em **Property name** (Figura 3.3).



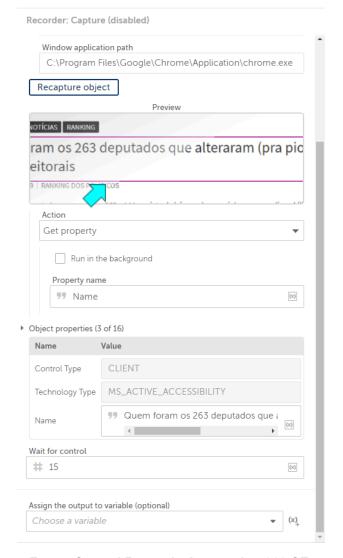


Figura 3.3 - Configuração da ação Recorder: Capture com Name.

Fonte: Control Room do Automation 360 CE.

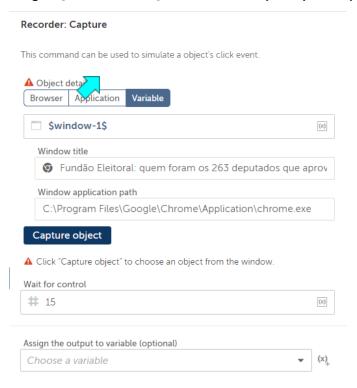
Por fim, uma variável String, **vTitulo**, deve ser criada e associada ao campo "Assign the output to variable (optional)". É importante salvar a configuração neste momento.

3.5. Recorder: Capture (capturar tabela)

O objetivo agora é extrair a tabela que contém os 263 deputados. A ação atuará também na janela window-1, porém, a captura será de uma tabela (Figura 3.4).



Figura 3.4 - Configuração inicial da ação Recorder: Capture para capturar tabela.



Fonte: Control Room do Automation 360 CE.

Ao clicar em Capture object, a janela do Chrome será aberta. O ponteiro do mouse deve agora, ser posicionado sobre a tabela, por exemplo, na célula NOME, aguardando ser contornado pela linha vermelha antes de clicar (Figura 3.5).



O QUE VOCÊ PROCURA? Quem foram os 263 deputados que alteraram (pra pior) as regras eleitorais 4 SETEMBRO 2019 | RANKING DOS POLÍTICOS A Câmara dos Deputados aprovou, por 263 a 144, projeto de lei que abre caminho para ampliar o bilionário fund eleitoral, recria a propaganda partidária na TV (que foi extinta justamente para custear as campanhas políticas), autoriza o uso de inichero pólitico para pagar advogados de investigados e institui regras mais flexíveis para contratação de dirigentes partidários. Indicação de Kássio Nunes ao STF: análise da Por estado: os 42 senadores que votaram para permitir aumentos para o funcionalismo durante a crise Confira quem são os 125 parlamentares que querem concorrer às eleições municipais; maioria vai mal no Ranking dos Políticos Confira o Ranking dos estados de acordo com a pontuação de seus parlamentares ESTADO Jéssica Sales Perpétua Almeida Dra. Vanda Milani O # # <u>P</u> <u>P</u> <u>O</u> <u>O</u> <u>O</u> <u>O</u> <u>O</u> <u>O</u> <u>O</u> ^ 🖫 🕬 🦸 POR 1401 🖪 Digite aqui para pesquisar

Figura 3.5 - Página do Chrome aberta para seleção da tabela.

Fonte: Blog Ranking dos Políticos, 2021.

Após o clique sobre o elemento contornado, a tabela terá sido capturada (Figura 3.6).



Recorder: Capture Object detail Browser Application swindow-1\$ (x) Fundão Eleitoral: quem foram os 263 deputados que apr Window application path C:\Program Files\Google\Chrome\Application\chrome.exe Recapture object Preview Cannot be empty Action Ob Get property Get cell Text By Index Get cell Index By Text Get total rows Get total column Set cell by index Set cell by text Wa Click Left click Right click Double click Assign the output to variable (optional) → (x) Choose a variable

Figura 3.6 - Configuração inicial da ação Recorder: Capture após capturar a tabela.

Fonte: Control Room do Automation 360 CE.

Agora, deve ser selecionado **Get table** no campo **Action**. Uma variável **vTabela** do tipo "Any" deve ser criada para receber a tabela que será capturada (Figura 3.7). Uma variável do tipo Any permite receber qualquer tipo de informação.



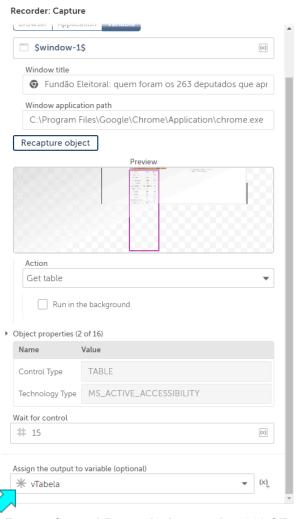


Figura 3.7 - Configuração final da ação Recorder: Capture.

Fonte: Control Room do Automation 360 CE.

3.6. Message box

Para verificar que o bot está funcionando corretamente, a ação **Message box** será usada. Se estiver tudo certo, ela exibirá a quantidade de colunas capturada.

Para tanto, no campo de mensagem deve-se colocar o seguinte:

\$vTabela.DataTable:columnCount.Number:toString\$

- vTabela é a variável que contém a tabela capturada.
- DataTable:columnCount.Number obtém a quantidade de colunas da tabela contida em vTabela.
- toString\$ converte o número em String.

Salvar as configurações, fechar a janela do Chrome e executar o bot. O resultado da execução deverá ser a exibição da mensagem com o número 3 (Figura 3.8).



Figura 3.8 - Mensagem exibida

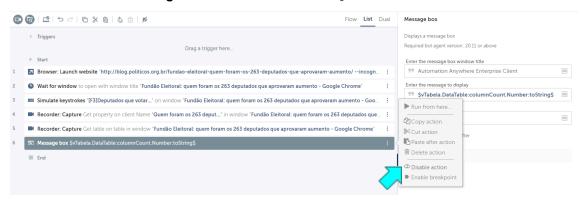


Fonte: Control Room do Automation 360 CE.

Se aparecer qualquer outro número diferente de 3 é porque a tabela não foi capturada corretamente. Nesses casos, é importante repetir as ações da captura da tabela observando todos os detalhes.

Estando certo, pode-se desabilitar a linha 6 (Figura 3.9).

Figura 3.9 - Desabilitando a ação da linha 6.



Fonte: Control Room do Automation 360 CE.

3.7. Data Table: Write to file

Esta ação permite salvar uma tabela num arquivo CSV. O vTitulo será o nome do arquivo e o conteúdo será o vTable (Figura 4.1).

Data Table: Write to file Writes a data table into a specified file Required bot agent version: 20.11 or above Data table name * vTabela Enter file name \$vCaminho\$/\$vTitulo\$.csv Browse.. Required extensions: ".csv", ".txt" ✓ Create folders/files if it doesn't exist When writing Append to the existing file Overwrite existing file Row delimiter new line Column delimiter comma Encoding UTF-8

Figura 4.1 - Configuração da ação Data Table: Write to file.

Fonte: Control Room do Automation 360 CE.

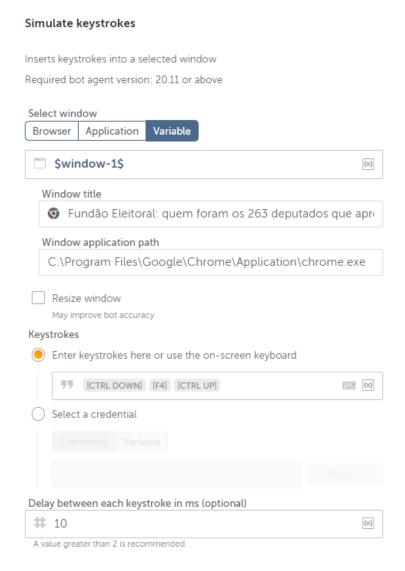
Deve-se seguir os sete passos indicados na Figura 4.1 para finalizar a configuração desta ação.

3.8. Simulate keystrokes (para fechar o chrome)

Para finalizar, deve-se fechar o navegador Chrome. Para isso deve-se simular o pressionamento das teclas Ctrl+F4 (Figura 4.2).



Figura 4.2 - Configuração da ação Simulate keystrokes (para fechar o chrome).



Fonte: Control Room do Automation 360 CE.

Salvar a configuração.

4. Considerações finais

Fechar o navegador Chrome antes de executar o bot. Se tudo estiver correto, o bot terá gerado um arquivo com o nome:

Quem foram os 263 deputados que alteraram (pra pior) as regras eleitorais.csv

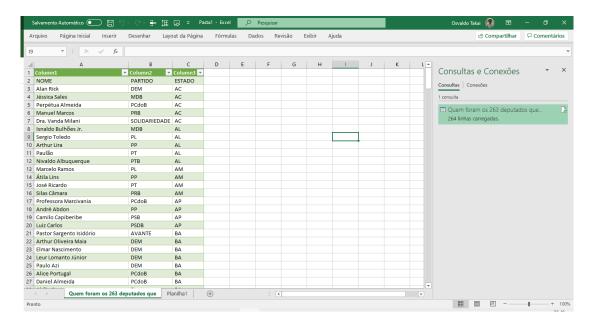
Para visualizar corretamente o arquivo CSV no excel, siga os seguintes passos:

- 1) Abrir o Excel com uma planilha vazia.
- 2) Selecionar a opção **De Text/CSV** do menu **Dados**.
- 3) Da caixa de diálogo que será aberta, selecione o arquivo CSV que deseja abrir.
- 4) Clicar no botão Importar.
- 5) Na nova caixa de diálogo exibida, deve-se selecionar Carregar.



6) A planilha com os dados terão sido carregados (Figura 5.1).

Figura 5.1 - Excel com a tabela extraída da Página Web.



Fonte: do autor, 2021.



Referências

AUTOMATION ANYWHERE IQ BOT. **Como usar IQ Bot**; 2020. Disponível em https://docs.automationanywhere.com/pt-BR/bundle/iq-bot-v6.5/page/iq-bot/topics/iq-bot/user/overview.html, acesso em 15/08/2021.

AUTOMATION ANYWHERE AUTOMATION 360. Using dictionary variable for email properties. Disponível em https://docs.automationanywhere.com/bundle/enterprise-v2019/page/enterprise-cloud/topics/aae-client/bot-creator/commands/cloud-using-email-properties.html>, acesso em 21 de julho de 2021.

AUTOMATION ANYWHERE COMMUNITY EDITION. Formulário para obtenção de acesso à versão Community Edition do Automation Anywhere Gratuita. São José – EUA. Disponível em https://www.automationanywhere.com/products/enterprise/community-edition, acesso em 21 de junho de 2021.

AUTOMATION ANYWHERE UNIVERSITY. **Introdução ao Automation Anywhere**. São José – EUA. Disponível em https://apeople.automationanywhere.com/s/getting-started>, acesso em 21 de junho de 2021a.

AUTOMATION ANYWHERE UNIVERSITY. **Trilhas de Aprendizagem**. São José – EUA. Disponível em: https://university.automationanywhere.com/training/rpa-learning-trails/>, acesso em 21 de junho de 2021b.

AUTOMATION ANYWHERE UNIVERSITY. **Email server setting.** Disponível em https://docs.automationanywhere.com/bundle/enterprise-v2019/page/enterprise-cloud/topics/aae-client/bot-creator/commands/cloud-configuring-mail-server.html >, acesso em 23 de julho de 2021c.

BANIN, S. L. **Python 3: conceitos e aplicações: uma abordagem didática**. São Paulo; Érica, 2018. ISBN 978-85-365-3025-3. https://integrada.minhabiblioteca.com.br/#/books/9788536530253/

CHICONI, N. O que é ASCII, UNICODE e UTF-8. CCM, 2020. Disponível em https://br.ccm.net/faq/9956-o-que-e-ascii-unicode-e-utf-8>. Acesso em 21 de julho de 2021.

CHANDRA, R. V.; VARANASI, B. S. **Python Requests Essentials**: Learn how to integrate your applications seamlessly with web services using Python Requests; Packt Publishing, 2015.

DIGICERT. THE ULTIMATE GUIDE: What is SSL, TLS and HTTPS? Disponível em https://www.websecurity.digicert.com/security-topics/what-is-ssl-tls-https>, acesso em 23 de julho de 2021.

ELMAN, J.;LAVIN, M. **Django Essencial - Usando REST, websockets e Backbone**. 1.ed. São Paulo: Novatec, 2015.



GOOGLE. Ajuda do Administrador do Google Workspace: Controle o acesso a apps menos seguros. Disponível em https://support.google.com/a/answer/6260879?hl=pt-BR>. Acesso em 21 de julho de 2021.

JARMUL, K.; LAWSON, R. Python Web Scraping. 2nd. Birmingham: Packt Publishing, 2017.

LOPES, M. D. e LIMA, W. R. **Análise do Índice de Massa Corporal de funcionários de uma instituição de ensino superior**; EFDeportes.com, Revista Digital. Buenos Aires - Año 18 - Nº 181 - Junio de 2013. Disponível em https://www.efdeportes.com/efd181/analise-do-indice-de-massa-corporal-de-funcionarios.htm, acesso em 21 de julho de 2021.

MCINGVALE, FRANK. **All about Python and Unicode**; trad. Menezes, Nilo; PythonBrasil, 2007. Disponível em https://wiki.python.org.br/TudoSobrePythoneUnicode>, acesso em 29 de julho de 2021.

MICROSOFT. O que são IMAP e POP? Disponível em https://support.microsoft.com/pt-br/office/o-que-s%C3%A3o-imap-e-pop-ca2c5799-49 f9-4079-aefe-ddca85d5b1c9>, acesso em 23 de julho de 2021.

PYTHON BRASIL. **Instalando o Python 3 no Windows**. Disponível em https://python.org.br/instalacao-windows/>, acesso em 24 de julho 2021.

NOTEPAD++, **What is Notepad++**. Disponível em <<u>https://notepad-plus-plus.org/</u>>, acesso em 28 de junho de 2021.

WDG AUTOMATION – AN IBM COMPANY. **7 pilares essenciais para projetos de RPA bem-sucedidos**. São Paulo: Newsletter WDG. Disponível em: https://www.wdgautomation.com/7-pilares-essenciais-para-projetos-de-rpa-bem-sucedidos/>, acesso em 21 de junho de 2021.