BUSCA_DOU

Script em R para monitoramento do Diário Oficial da União

Escrito por Henrique A. Castro

Doutorando na Faculdade de Direito da Universidade de São Paulo, pesquisador visitante na Harvard Kennedy School

Contato: henrique.almeida.castro@usp.br

LATTES: http://buscatextual.cnpq.br/buscatextual/visualizacv.do?id=K4992004E4

*Esse script é gratuito e vem sem garantia alguma. É direcionado a fins acadêmicos. Jamais pague por ele.

Sumário

APRESENTAÇÃO

O BUSCA_DOU é um script de programação que desempenha a função de buscar e separar, em relação aos dias em que for executado, publicações no Diário Oficial da União (DOU) que correspondam a parâmetros de pesquisa definidos pelo usuário. Não tem como propósito buscar publicações em edições passadas, mas sim monitorar diariamente o DOU. Ele emprega a linguagem de programação R, comumente usada no meio acadêmico para a obtenção e processamento de dados.

O script foi escrito com a automação em mente. Ou seja, é possível, com as ferramentas dos sistemas operacionais de computador, fazer com que o script rode diariamente de maneira automática (sem a necessidade de ações do usuário) – desde que o computador esteja ligado e conte com acesso à internet. Pode-se, ainda, utilizar um sistema de notificações por celular, o qual avisa o usuário sobre os resultados da pesquisa (avisando, inclusive, a ocorrência de erros).

Escrevi o script para minhas próprias pesquisas, mas tentei o tornar o mais acessível possível a usuários sem conhecimento da linguagem R (eu mesmo estou muito longe de um grande entendido no assunto). De qualquer maneira, como não é um programa com interface de usuário própria, diversos aspectos de seu funcionamento podem ser contraintuitivos para a maior parte dos pesquisadores. Por isso, recomenda-se atenção às instruções contidas neste manual.

Se tiver sugestões para melhorar o script, ou identificar erros sistemáticos no seu funcionamento, pode me contatar em henrique.almeida.castro@usp.br.

PROCESSO E PRODUTO DO SCRIPT

O BUSCA_DOU executa suas funções através de uma série de passos:

1. Acessa o site endereço https://inlabs.in.gov.br/, o qual distribui diariamente o DOU em formato .xml.
2. Realiza login a partir do e-mail e da senha do usuário.
3. A partir das definições das edições do DOU (regular e/ou extra) e das partes (1, 2, e/ou 3) a serem buscadas, faz o download os arquivos .xml (cada qual contendo um ato) relativos...
    a. ...à edição regular do DOU do dia em que for executado, e
    b. ...à edição extra do DOU *do dia anterior* àquele em que for executado.
4. Busca, entre os arquivos .xml, aqueles cujos textos correspondam aos parâmetros de pesquisa.
5. Extrai os textos dos arquivos .xml selecionados, colocando-os em arquivos .txt
6. Deleta os arquivos .xml

O produto (se tudo ocorrer bem) são pastas contendo os arquivos .txt produzidos, salvas no mesmo diretório onde se encontrar o script. As pastas mais altas tomam designações relativas ao dia da busca, no formato "AAAA-MM-DD", com a designação da pasta relativa à edição extra contendo um "E" no final. Em seguida, há pastas indicando a parte do diário oficial de onde saíram os arquivos, nomeadas no formato "DO#" (as designações das edições extras contêm um "E" ao final). Por fim, há os arquivos .txt contendo os atos selecionados, nomeados a partir da identificação do ato.

Como exemplo, uma busca rodada no dia 14/08/2020, a depender dos parâmetros de pesquisa, poderia gerar um resultado com o seguinte aspecto:

- /2020-08-13E
  - /DO1E
    - /PORTARIA Nº 468, DE 13 DE AGOSTO DE 2020.txt
    - /PORTARIA Nº 650, DE 13 DE AGOSTO DE 2020.txt
  - /DO2E
    - /PORTARIA Nº 674 DE 13 DE AGOSTO DE 2020.txt
- /2020-08-14
  - /DO1
    - /PORTARIA Nº 278, DE 12 DE AGOSTO DE 2020.txt
  - /DO2
    - /aApQv7EJxJ0wYT6GC6a1nPcWj7tqWD.txt

IMPORTANTE 1: note que o último arquivo .txt tem um nome estranho. Isso não é um erro. Os nomes são retirados de uma etiqueta de identificação contida nos arquivos .xml. Em alguns casos, por algum motivo, essa etiqueta vem vazia. Quando isso ocorre, o script nomeia o arquivo .txt com uma série de 30 caracteres gerados aleatoriamente.

IMPORTANTE 2: é normal que o código demore um pouco na primeira execução. Isso porque ele precisa baixar e instalar alguns pacotes.

COMO USAR

1. Baixe e instale os programas “R” e “R Studio”. Ambos são gratuitos.
   a. “R” está disponível em: https://cran.r-project.org/.
   b. “R Studio” está disponível em: https://rstudio.com/products/rstudio/.
2. Registre-se no site inlabs: https://inlabs.in.gov.br/acessar.php.
3. Baixe o pacote do BUSCA_DOU em seu repositório no GitHub.
   a. O pacote está disponível em: https://github.com/HENRCAST/BUSCA_DOU.
   b. Para baixar, clique em “CODE -> Dowload ZIP”

4. Extraia o arquivo .zip, e coloque a pasta resultante onde desejar.
5. Preencha o arquivo “config.R” com os parâmetros de busca, conforme as instruções contidas no manual.
6. Agende a execução do script utilizando os recursos de seu sistema operacional (no caso do Windows, o “Agendador de Tarefas”, conforme as instruções contidas no manual).
7. OPCIONAL: caso queira utilizar o sistema de notificações no celular, instale o app “Pushbullet”, cadastre-se no serviço, e preencha o arquivo “config.R”, conforme as instruções contidas no manual.

CONFIGURAÇÕES DO SCRIPT

O arquivo “config.R” contém as variáveis utilizadas pelo script durante a sua execução. Seus valores precisam ser definidos pelo usuário. São elas:

- Identificação:
  - “user”: e-mail cadastrado no site https://inlabs.in.gov.br/
  - “pass”: senha cadastrada no site https://inlabs.in.gov.br/

- Termos de busca:
  - “buscar_por”: contém os termos a serem buscados. O script seleciona publicações que contenham em seu texto qualquer um dos termos.
  - “nao_incluir”: contém termos de exclusão. O script exclui publicações que contenham em seu texto qualquer um dos termos, mesmo que essas publicações contenham algum dos termos definidos em “buscar_por”

- Seções de edições do Diário Oficial incluídas na busca:
  - “dou”: contém os números (1, 2, e/ou 3) das seções do Diário Oficial nos quais a busca será realizada.
  - “dou_e”: contém os números (1, 2, e/ou 3) das seções da edição extra Diário Oficial nos quais a busca será realizada.

- Chave de acesso do Push Duo:
  - “chave_pushbullet”: contém a chave de acesso necessária para que o sistema de notificações funcione. É opcional (mas recomendada).

Como as configurações estão no formato de um script de programação, sem uma interface gráfica para o usuário, é preciso seguir algumas regras na definição dos valores das variáveis. Caso desobedecidas, o script pode resultar em um erro:

- Não altere as designações das variáveis.
- Toda variável é seguida do símbolo “<- “. Isso faz indica que os valores à direita do símbolo são assumidos pela variável. Não delete esse símbolo.
- Todo valor deve estar entre aspas.
    - Ex.1: user <- “exemplo@gmail.com”
    - Ex.2: buscar_por <- "Conselho Nacional"
- Quando houver mais de um valor, eles precisam estar conjuntamente contidos no termo “c()” e separados por vírgula.
    - Ex.1: buscar_por <- c("Conselho Nacional", "portaria", "ministério público")
    - Ex.2:dou <- c("1", "2")
- Mesmo que escolha não designar um valor para determinada variável, mantenha as aspas.
    - Ex.: dou_e <- ""
- Nota: as variáveis de busca “buscar_por” e “nao_incluir” desconsideram a caixa dos caracteres ("CNPq" e "cnpq" dão na mesma).

Um arquivo de configuração devidamente preenchido tem o seguinte aspecto:

user <- “exemplo@gmail.com”

pass <- “senhaindevassavel333”

buscar_por <- c("Conselho Nacional", "Estrutura Regimental e o Quadro Demonstrativo")

nao_incluir <- c("CONFAZ", "Conselho Nacional do Ministério Público", "CNPq", "Conselho Nacional de Justiça")

dou <- c(“1”, “2”)

dou_e <- “”

chave_pushbullet <- "a.asdffsdIHFSAD198scfF2Gdfg234GvFVYP"

AUTOMATIZAÇÃO DE TAREFA

Utilizando os recursos dos sistemas operacionais de computador, é possível automatizar a execução do script. Esse manual lida apenas com o Windows 10, o sistema a que tenho acesso.

Não posso garantir que funcione, mas um tutorial para MacOs está disponível em: https://www.r-bloggers.com/how-to-source-an-r-script-automatically-on-a-mac-using-automator-and-ical/

1. Adicione o "R" às suas variáveis de ambiente. Para tanto, vá em:
    a. Painel de controle
    b. Contas de usuário
    c. Contas de usuário (outra vez)
    d. Selecione, no lado esquerdo da tela, "Alterar as variáveis do meu ambiente".
    e. Selecione a variável "Path"
    f. Clique em "Novo"
    g. Escreva no campo aberto "C:\Program Files\R\R-X.X.X\bin\x64", sendo que os "X" devem ser substituídos pelos números da versão do "R" instalada em seu computador (no meu caso, fica "C:\Program Files\R\R-4.0.2\bin\x64").
    h. Feche o Painel de Controle
2. Abra o programa "Agendador de Tarefa".
3. Selecione "Criar Tarefa Básica".
4. Dê o nome que quiser à Tarefa.
5. Na aba "disparador", selecione:
    a. "Semanalmente".
    b. O horário em que deseja que o script rode.
    c. Os dias em que deseja que o script rode. Lembre-se de que o script busca a edição extra do DOU publicada no dia anterior a sua execução. Assim, é recomendável que selecione, além de todos os dias úteis, também o sábado.
6. Na aba "ação", precisamos dizer ao Agendador para rodar o arquivo "exec_script.bat". Para isso,
    a. Selecione "iniciar um programa"
    b. No campo "programa/script", clique em "selecionar".
    c. Vá até o diretório onde se encontram os arquivos do "BUSCA_DOU", e selecione o "exec_script.bat".
    d. No campo "adicione argumentos", insira o caminho do diretório. Isso vai variar conforme o lugar onde guardou o script. No meu caso, fica: C:\Users\henri\OneDrive\Documentos\R\ArquivosR\BUSCA_DOU-

master.
Para obter o caminho do script, copie-o do navegador.

7. Na aba “Concluir”, revise as informações, selecione “Abrir a caixa propriedades da tarefa”, e clique em “Concluir”
8. Na aba “Condições”, selecione “Iniciar somente se a seguinte conexão de rede estiver disponível:”, e depois, “Qualquer conexão”. Isso fará com que a tarefa se inicie apenas quando o computador tiver acesso à internet, o que é necessário para que o script conclua sua tarefa.
9. Na aba “Configurações”, selecione “Executar o mais cedo possível após uma inicialização agendada ter sido perdida”. Assim, se por algum motivo a tarefa tiver perdido seu agendamento, ela será executada quando o computador puder.

Enquanto o script estiver sendo executado, ficará aberto de forma minimizada o “prompt de comando” do Windows (uma tela preta com um monte de palavras). Não feche essa janela, ou a execução será interrompida.

EXECUTANDO O SCRIPT MANUALMENTE

Em alguns casos, pode ser necessário executar o script manualmente. Isso ocorrerá, em especial, quando o script já tiver sido executado, mas sem sucesso na realização de suas funções. Nessa hipótese, o Agendador de Tarefas não irá operar novamente. Por exemplo: pode ser que, no momento da execução agendada, o site https://inlabs.in.gov.br/ esteja apresentando instabilidades, e que isso impeça o dowload dos arquivos necessários.

Executar o script manualmente não é complicado. Simplesmente clique no arquivo “exec_script.bat”.

NOTIFICAÇÕES POR CELULAR COM “PUSHBULLET”

É opcional, mas recomendado, ligar o script ao sistema de notificação por celular "PushBullet". Isso tem duas vantagens:

- Você receberá uma notificação assim que a busca for concluída, informando o número de resultados encontrados.
- E, mais importante, caso ocorra algum erro que impeça o funcionamento do script, receberá um aviso, acompanhado de uma mensagem de erro gerada pelo R.

É importante notar que, caso o script seja executado sem que o computador tenha acesso à internet, ele não concluirá sua função e não enviará a notificação de erro.

Para usar o PushBullet:

1. Baixe o app PushBullet na loja virtual do seu aparelho.
2. Cadastre-se no aplicativo.
3. Entre, preferencialmente pelo seu computador, no site https://www.pushbullet.com.
4. Faça login utilizando seu cadastro.
5. Na aba "Settings":
    a. Selecione "Account"
    b. Selecione "Create Access Token"
6. Copie a chave gerada.
7. No arquivo "config.R", designe-o à variável "chave_pushbullet". Lembre-se: como todos os demais valores no "config.R", a chave deve ficar entre aspas.

Log de mudanças

- Ver. 1.0.1. = correções no agendamento de tarefas