Leonardo F. Nascimento[1] Eric Brasil[2] Gabriel Andrade[3]
Tarssio Barreto[4] Vítor Mussa[5] Daniel Mendes[6]

2021-04-28

[1]UFBA/ICTI/LABHDUFBA/PPGCS, leofn@ufba.br
[2]UNILAB - LABHDUFBA, profericbrasil@unilab.edu.br
[3]LABHDUFBA, gabriel.andrad4@gmail.com
[4]LABHDUFBA, tarssioesa@gmail.com
[5]UFRJ/PPGSA/DTA - LABHDUFBA, vtrmussa@gmail.com
[6]UFRJ/PATHS, daniel\protect_mnds34@hotmail.com

Contents

Chapter 1

Apresentação

A ideia desta obra foi reunir esforços de diferentes pesquisadores e instituições na elaboração de scripts para coletar - de modo automatizado - a produção intelectual dos principais congressos e eventos das áreas das humanidades.

Além disso, nós tivemos como objetivo mais amplo enfatizar a importância do desenvolvimento de habilidades computacionais por parte dos pesquisadores em todos os campos das humanidades.

Os scripts, as bases de dados e todos os documentos estão disponíveis e poderão ser baixados com apenas um clique. O acervo servirá para a realização de investigações sobre os mais variados aspectos e ampliar, com isso, o conhecimento sobre a produção acadêmica, científica e intelectual do Brasil das ciências humanas e sociais ao longo de décadas.

Para o lancamento, nós escolhemos o Dia Internacional das Humanidades Digitais em 29/04/2021.

Ao compartilhar nas redes, pedimos que usem a hashtag **#dayofdh21**

Figure 1.1: Símbolo do #dayofdh21

Chapter 2

2 Webscraping e ciências sociais

2.1 2.1 Por que automatizar?

A dataficação e a digitalização tornaram-se fenomenos massivos das sociedades contemporâneas. Ao interargirmos com as tecnologias digitais nós deixamos traços de dados que podem ser usados para a pesquisa sobre a sociedade. O desafio colocado para os pesquisadores das humanidades está em acessar e manipular tais dados:

> "Como uma técnica de extração de dados online, o [webscraping] parece de interesse especial para nós porque é uma parte importante do que torna a pesquisa social digital praticamente possível." (MARRES, N. & WELTEVREDE, E. Scraping the Social? Journal of Cultural Economy, v. 6, n. 3, p. 313–335, 1 ago. 2013, p.317)

O volume, quantidade e qualidade dos dados digitais e digitalizados nunca foi tão grande. O acesso à fontes digitalizadas através de mecanismos de busca por palavras-chave, por assuntos, por metadatos em geral, os milhares de dados produzidos a cada segundo nas redes sociais ou o volume de publicações acadêmicas têm impactado as pesquisas e a própria construção do conhecimento nas ciências humanas e sociais.

Assim, é urgente a necessidade de enfrentarmos os desafios metodológicos e teóricos colocados por esse cenário. A automatização na coleta de dados na Web não é apenas uma forma de acelarar essa relação do pequisador com os dados, mas de qualificar e potencializar a tarefa heurística de seleção dos mesmos.

2.2 2.2 Como começar?

É preciso aprender algum tipo de linguagem de programação (geralmente R ou Python), além de conhecimentos em HTML, CSS e XPATH. Sabemos que, à primeira vista, parecem ser termos complicados para quem vem "das humanas", mas o entendimento destas coisas é relativamente mais simples que muitas das leituras que nós fazemos.

Portanto, talvez o primeiro passo seja buscar compreender a estrutura da página que abriga os dados que você pretende coletar. Para isso, é preciso conhecer o mínimo de HTML.

Em seguida é importante definir quais dados e informações você pretende coletar e qual a estrtura de organização você pretende construir como resultado. Esse é um procedimento metodológico fundamental para a pesquisa e demanda do pesquisador o mesmo rigor acadêmico do trabalho com dados de outra natureza.

Por fim, a escrita do código, utilizando a linguagem que melhor atenda aos seus interesses.

Todos esses processos demandam um empenho de tempo e formação técnicas específicas, sem dúvida. Entretanto, acreditamos que os retornos possíveis justificam o investimento de tempo. Além disso, amplia as possibilidades de trabalho interdisciplinar, colaborativo e aberto.

2.3 2.3 Webscraping enquanto técnica das humanidades

Ao realizarmos um webscrapig é preciso atentar para os procedimentos não apenas "técnicos" envolvidos na raspagem mas, também, para os aspectos analíticos e epistemológicos. Cada plataforma, website ou API possui características particulares que vão, juntamente com o código que vamos contruir, determinar o tipo e natureza dos dados coletados.

> A raspagem, entretanto, não é apenas uma técnica, mas também envolve uma forma particular de lidar com a informação e o conhecimento: é também uma prática analítica.(MARRES, N. & WELTEVREDE, E. Scraping the Social? Journal of Cultural Economy, v. 6, n. 3, p. 313–335, 1 ago. 2013, p.317)

Erros no código de raspagem podem produzir dados distorcidos, com lacunas ou mesmo em duplicidade. Podemos, então, considerar que um erro no código torna-se um erro metodológico.

Chapter 3

Linguagens de programação

3.1 R

3.2 Python

Alguns dos códigos que compõe o Redhbr foram escritos em Python 3.8. Esta é uma linguagem de programação que permite ao programados trabalhar rapidamente e integrar diferentes sistemas com maior eficiência.

> Foi lançada por Guido van Rossum em 1991. Atualmente, possui um modelo de desenvolvimento comunitário, aberto e gerenciado pela organização sem fins lucrativos Python Software Foundation.[1]

Parte da filosofia da linguagem está resumida no poema *Zen of Python*, escrito por Tim Peters em 1999.

> Bonito é melhor que feio Explícito é melhor que implícito Simples é melhor que complexo Complexo é melhor que complicado Linear é melhor do que aninhado Esparso é melhor que denso Legibilidade conta Casos especiais não são especiais o bastante para quebrar as regras. Ainda que praticidade vença a pureza Erros nunca devem passar silenciosamente. A menos que sejam explicitamente silenciados Diante da ambiguidade, recuse a tentação de adivinhar Deveria haver um — e preferencialmente apenas um — modo óbvio para fazer algo. Embora esse modo possa não ser óbvio a princípio a menos que você seja holandês Agora é melhor que nunca Embora

[1] Python - Wikipedia.org

> nunca freqüentemente seja melhor que já Se a implementação é difícil de explicar, é uma má ideia Se a implementação é fácil de explicar, pode ser uma boa ideia Namespaces são uma grande ideia — vamos ter mais dessas![2]

Para executar um arquivo .py é preciso instalar o Python3 em seu computador.

Clique aqui para um tutorial de instalação do Python no Windows, clique aqui para Linux e clique aqui para Mac.

Após a instalação, vc pode executar o arquivo .py direto do prompt de comando do Windows ou pelo terminal do Linux, ou utilizar as diversas IDE disponíveis.

Segue um exemplo de como executar utilizando o terminal do Linux, após instalar o Python3.8:

1. Acesse o diretório em que o arquivo .py está salvo:

sh `$ cd "caminho do diretório"`

1. Instale as bibliotecas requeridas:

sh `$ pip3 install -r requirements.txt`

1. Execute o arquivo usando Python3.8

sh `$ python3 script-anais-anpuh.py`

[2] Zen of Python - Wikipedia.org

Chapter 4

ANPUH

4.1 O que é ANPUH?

A Associação Nacional de História, Anpuh, fundada em 1961, inicialmente destinada aos docentes de cursos de graduação e pós-graduação. Em 1993, a ANPUH ampliou sua base para todoa os profissionais de história.

> A cada dois anos, a ANPUH realiza o Simpósio Nacional de História, o maior e mais importante evento da área de história no país e na América Latina[1].

Desenvolvemos scripts diferentes para dois tipos de conjuntos de dados relacionados à Associação Nacional de História.

- Anais-Anpuh: script para raspagem de todos os trabalhos publicados nos Anais dos Simpósio Nacionais de História entre 1963 e 2017, disponíveis no site da Anpuh.
- anpuh-scraper: script para raspagem dos resumos (e demais informações) de todos os trabalhos aprovados para todos os simpósios temáticos dos SNH nos aos de 2013, 2015, 2017 e 2019.

4.2 Scripts de raspagem

4.2.1 Anais em pdf da ANPUH

Clique aqui para acessar o repositório no Github

[1] Anpuh-Quem somos

Esse script realiza a raspagem dos trabalhos em PDF de todos os Simpósios Nacionais da Anpuh entre 1963 até 2017, disponíveis atualmente na site da associação, que podem ser acessados aqui.

Escrito em Python 3.8, o script utiliza as seguintes bibliotecas e módulos

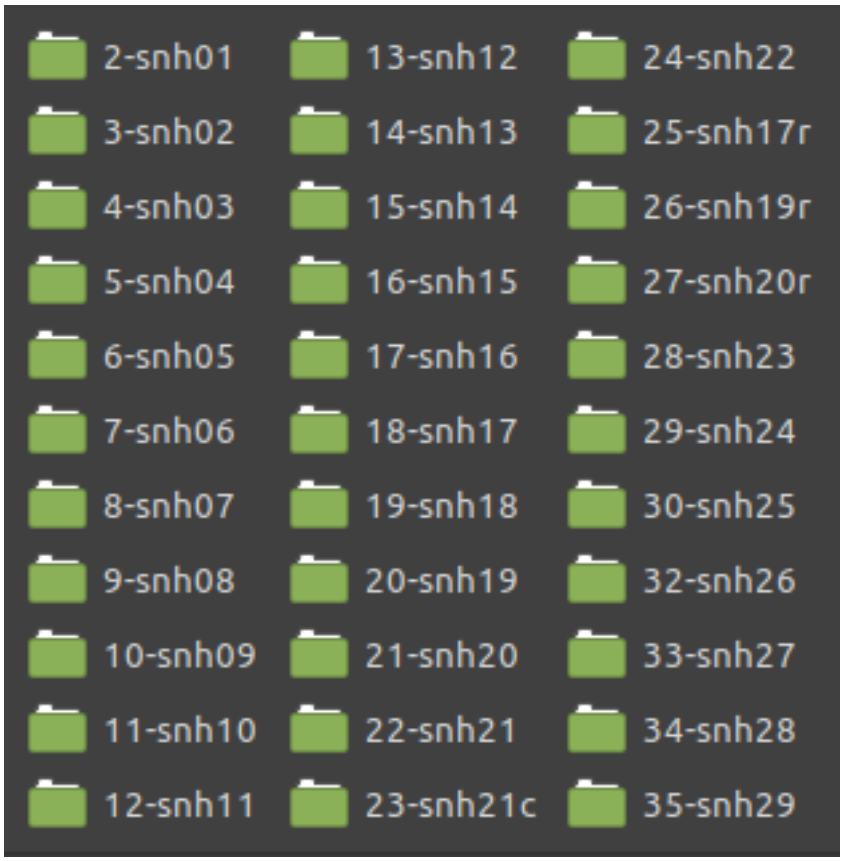
- **urllib.requests**: módulo do Python para acessar urls. Saiba mais.
- **os**: módulo do Python que permite manipular funções do sistema operacional. Saiba mais.
- **bs4**: Beautiful Soup é uma biblioteca Python para extrair dados de arquivos HTML e XML.
- **re**: Regular Expressions é um módulo do Python para operar com expressões regulares.
- **pandas**: Pandas é uma biblioteca escrita em Python para manipulação e análise de dados.
- **wget**: Wget é uma biblioteca escrita em Python para realizar downloads.

O script tem o seguinte funcionamento quando executado:

- Cria pasta para salvar os PDFs, após verificar se a mesma não existe no local: `Anais Anpuh> pdf` utilizando módulo `os`.
- Acessa a URL dos Anais com a biblioteca `urllib` e realiza a análise do HTML da mesma com a biblioteca `BeautifulSoup`;
- Cria uma lista de eventos a partir da página principal;
- Acessa as páginas de cada evento contidas na lista criada anteriormente através de uma iteração;
- Em cada item da lista de eventos, o script busca todos os papers da primeira página e cria uma nova lista. Nessa lista de papers de uma dada página o script realizará as seguintes ações:
  - encontrar as informações de cada paper;
  - inclui essas informações em uma lista (que depois gerará um CSV com os dados);
  - busca se há pdf disponível e se ele não é repetido faz download do PDF
  - Após realizar essas ações para todos os itens de uma página, busca a próxima página de papers do evento, se não houver, passa para o próximo evento e repete as ações em um *loop* até o último evento disponível.

4.2.2 Dados

O script retorna para o usuário **todos os pdfs disponíveis em todas as páginas de todos os Simpósios Nacionais da Anpuh desde 1963 até 2017**. São criadas pastas com o número de cada evento para o armazenamento dos arquivos em PDF.



É importante notar que muitos papers não estão com pdf disponível no site, assim como nas edições mais antigas encontramos arquivos que contém vários papers num único PDF.

O script também gera um arquivo **CSV** (*comma-separated values*) contendo os seguintes valores para cada paper: Autor(es)/Instituições,Título, Tipo, Evento, Ano, Link do Arquivo. Esse arquivo pode ser aberto como uma planilha e trabalhado em banco de dados.

	Autor(es)/Instituições	Título	Tipo	Evento	Ano	Link do Arquivo
0	Maria Helena Capelato	os desafios da anpuh frente à crise brasileira: a luta pela preservação da demo	Conferência	XXIX Simpós	2017	https://anpuh.org.br/
1	Abner Neemias da Cruz (Univ	percursos políticos independentistas:a primeira geração diplomática, o impéri	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
2	A. Ricardo Abdalla (Pontifíci	comensalidade e memória árabe na área central da cidade de são paulo	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
3	Acácia Regina Pereira (UERJ	currículo de história como responsável pela formação da identidade e memóri	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
4	Adalberto Paranhos (UNIVER	à flor da pele: pulsações do desejo feminino na música popular brasileira dos a	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
5	Adalberto Coutinho de Araúj	democracia e socialismo: propostas de políticas econômicas do operariado em	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
6	Adna Gomes Oliveira (UERJ)	o itamaraty, o corpo diplomático e a onu no início da guerra fria: a atuação bra	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
7	Adriana Aparecida Pinto (Uni	a imprensa e discurso sobre mulheres na imprensa rondonopolitana nos anos f	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
8	Adriana Maria de Souza Ziere	elementos religiosos da ascensão de d. joão i ao poder: o messias, o povo e a ci	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
9	Adriana Barreto de Souza (UI	josé narcizo de magalhães e menezes, a profissão militar e as milícias de home	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
10	Aécio Lessa Macedo (Univers	as potencialidades do estudo da história local em sala de aula: análise de uma	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
11	Adriane Piovezan (FIES Facul	rituais fúnebres militares: o túmulo dos fuzileiros navais mortos na intentona i	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
12	Afonso Henrique Sant Ana B.	labaredas do rio - abordagens histórico-sociais do rio e do corpo de bombeiros	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
13	Agnaldo Kupper (Universidad	futebol e contextos	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
14	Agenor Sarraf Pacheco (UNI\	a folia da cidade-floresta: patrimônio do afeto e lutas pela memória em melga	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
15	Agostinho Júnior Holanda Cc	a santa casa da misericórdia do maranhão e a intervenção dos presidentes-pro	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
16	Alan Christian de Souza Sant	lauro sodré em revista: textos, traços e retratos do senador paraense no perio	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
17	Airton José Cavenaghi (Anhe	mapas turísticos e sua representação cultural: marcel mauss e seu "ensaio sob	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
18	Aiman Jorge Henrique Franc	ivan ribeiro e a "via prussiana" no mundo rural brasileiro	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
19	Ailton José Morelli (Universi	a importância dos espaços nas memórias de infância	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
20	Alana Cavalcanti Cruz (UNIVE	memórias de moradores da rua duque de caxias, joão pessoa-pb, da primeira m	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
21	Alan Ricardo Duarte Pereira	trajetórias militares no império português: notas introdutórias sobre a família	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
22	Alberto Dias Mendes (UERJ)	jânio quadros e as influências de bandung	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
23	Alanna de Jesus Teixeira (UFI	a representação do passado em anatole france: literatura e história	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
24	Alec Ichiro Ito (USP)	a sucessão episcopal no bispado e diocese de congo e angola entre 1607 e 161	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
25	Aldina da Silva Melo (UNIVEF	a áfrica no ensino de história no brasil	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/
26	Alessandra Gonzalez de Carv	da "zona de contato" ao "limite antagônico": disputas entre criollos e indígena	Comunicação	XXIX Simpós	2017	https://anpuh.org.br/

4.2.3 Resumos dos trabalhos da ANPUH

Clique aqui para acessar o repositório no Github.

Raspador dos resumos dos Simpósios Nacionais de História da Associação Nacional de História - Anpuh. O programa raspa todos os resumos dos SNH 27, 28, 29 e 30, respectivamente dos anos de 2013, 2015, 2017 e 2019 Escrito em Python 3.8, o script utiliza as seguintes bibliotecas e módulos

- **urllib.requests**: módulo do Python que ajuda a acessar urls. Saiba mais.
- **bs4**: Beautiful Soup é uma biblioteca Python para extrair dados de arquivos HTML e XML.
- **pandas**: Pandas é uma biblioteca escrita em Python para manipulação e análise de dados.

O script tem o seguinte funcionamento quando executado:

Pergunta ao usuário que ano pretende raspar e se deseja incluir um novo ano à lista. Após a criação da lista com os anos escolhidos pelo usuário, o script acessa cada uma das páginas com as listas dos STs nos sites de cada evento; Acessa cada ST, encontra os dados de todos os resumos e passa para o ST seguinte; Após terminar um ST, passa para o próximo evento e executa as mesmas função; Todos os dados são inseridos em um DataFrame em Pandas e ao final são salvos no formato CSV.

4.2.4 Dados

O script retorna para o usuário um **CSV (*comma-separated values*) com os dados de todos os trabalhos aceitos nos Simpósio Temáticos dos SNH 27, 28, 29 e 30**.

O CSV contém as seguintes variáveis para cada resumo:

`Ano, Evento, Cidade, ST, Coordenadores, Autor(es)/Instituições, Título, Resumo`

Esse arquivo pode ser aberto como uma planilha e trabalhado em banco de dados.

	Ano	Evento	Cidade	ST	Coordenadores	Autor(es)/Instituições	Título	Resumo
0	2013	XXVII	Natal	001. "	ALEXANDRE VIEIRA RIB	Letícia Cristina Fonseca Destro	Cristãos, mouros e gentios: os africa	Quando as caravelas portuguesas se encontraram em terras ignotas ale
1	2013	XXVII	Natal	001. "	ALEXANDRE VIEIRA RIB	Ariane Carvalho da Cruz	A africanização da guerra em Angol	O presente trabalho analisa as formas de organização das tropas milita
2	2013	XXVII	Natal	001. "	ALEXANDRE VIEIRA RIB	Érika Melek Delgado	A Expedição Vitoriana para a África	Este artigo tem como objetivo dar espaço para uma apresentação siste
3	2013	XXVII	Natal	001. "	ALEXANDRE VIEIRA RIB	VANICLEIA SILVA SANTOS	O colonialismo inquisitorial: O Sant	O objetivo dessa apresentação é mostrar o funcionamento da inquisiç
4	2013	XXVII	Natal	001. "	ALEXANDRE VIEIRA RIB	Cristiana Ferreira Lyrio Ximenes	ANGOLA E BAHIA NAS REDES E ROTA	O texto apresenta algumas conexões constituídas nas rotas de comérci
5	2013	XXVII	Natal	001. "	ALEXANDRE VIEIRA RIB	Bruno Rafael Véras de Morais e Silva	Representação e literatura de viage	Os relatos e depoimentos de viajantes, atualmente, vêm sendo objeto O conceito de Representação é analisado como ferramenta metodológ Por tanto, são questões centrais na reflexão histórica sobre os relatos
6	2013	XXVII	Natal	001. "	ALEXANDRE VIEIRA RIB	Nielson Rosa Bezerra	Brasil e Serra Leoa: uma perspectiv	Durante o século XIX, aproximadamente 3.3 milhões de africanos foram
7	2013	XXVII	Natal	001. "	ALEXANDRE VIEIRA RIB	Daniela Carvalho Cavalheiro	Angola e Brasil: Tráfico Ilegal e Afric	Este trabalho tem como objetivo identificar os indivíduos apresados na
8	2013	XXVII	Natal	001. "	ALEXANDRE VIEIRA RIB	Luís Frederico Dias Antunes	Intérpretes, intermediários, e funci	Por mais paradoxal que possa parecer, no segundo quartel do século X Esta comunicação, tendo como objecto de estudo as respostas ao «inq mas também a sua inter-relação com os intérpretes neste processo, pr
9	2013	XXVII	Natal	001. "	ALEXANDRE VIEIRA RIB	MARIA CRISTINA CORTEZ WISSENBACH	Entre o porto de Ambriz, as minas d	Tendo como base relatos de expedicionários europeus (em especial jo
10	2013	XXVII	Natal	001. "	ALEXANDRE VIEIRA RIB	Márcia Cristina Pacito Fonseca Almeida	Comércio, bens de prestígio e insig	Ao longo da segunda metade do século XIX, a região da África centro-c
11	2013	XXVII	Natal	001. "	ALEXANDRE VIEIRA RIB	Francisco Aimara Carvalho Ribeiro	A Senegâmbia e a construção do di	Nessa comunicação, analiso o relato de André Almada sobre a região d
12	2013	XXVII	Natal	001. "	ALEXANDRE VIEIRA RIB	Roquinaldo Ferreira	A institucionalização dos Estudos A	Este artigo analisa o advento, consolidação e transformação dos estud
13	2013	XXVII	Natal	001. "	ALEXANDRE VIEIRA RIB	Marcia Guerra Pereira	A pesquisa em História da África na	Este trabalho integra o mapeamento do estado da arte da disciplina Hi
14	2013	XXVII	Natal	001. "	ALEXANDRE VIEIRA RIB	LUCIANA REGINA POMARI	Possibilidades e limites de inventar	O objetivo deste trabalho está focado na possibilidade de estabelecer
15	2013	XXVII	Natal	002. "	LEILA MARIA G. L. HER	Priscila Henriques Lima	Literatura de guerrilha: a ideologia	Mesmo diante de todo o apoio oferecido que tornava o MPLA o primeir Com a palavra de ordem Todos para o interior!", Agostinho Neto chama Com este movimento, Agostinho Neto não almejava apenas o avar Em 1972, Pepetela é transferido para a região da Frente Leste, ondi