



Portugal WordPress Fingerprint

Criptografia e Segurança na Comunicação

Alunos: Marco da Luz 26476
Francisco Oliveira 22252
Romilson Monteiro 28891

Orientação: Prof. Hugo Almeida e Pedro Pinto

**ipvc**estg

ERSC

ENGENHARIA DE REDES E
SISTEMAS DE COMPUTADORES

Julho de 2023

Introdução

- ❑ Este projeto consiste em um web crawler que busca por sites em português construídos com WordPress. Ele utiliza características específicas em URLs e conteúdos das páginas para identificar esses sites.
- ❑ Além disso, há um segundo crawler que verifica se um domínio específico utiliza WordPress.
- ❑ Também será desenvolvido uma interface web para facilitar a interação do utilizador com os crawlers, permitindo obter uma lista de sites em português construídos com WordPress e filtrar os resultados por critérios como a versão do WordPress e a popularidade do site. O objetivo do projeto é fornecer insights sobre o uso e distribuição do WordPress no cenário web em português.



Objectivos

Os objetivos desse mini-projeto, são:

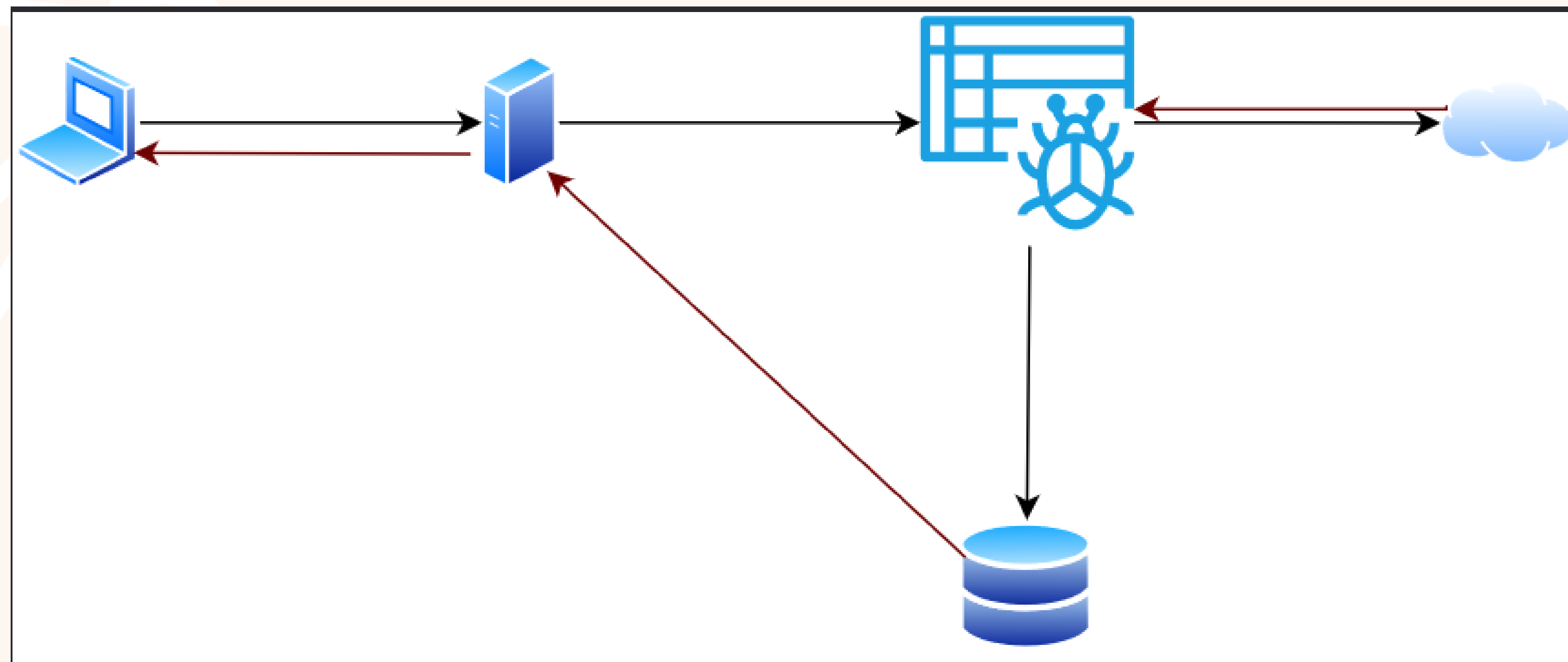
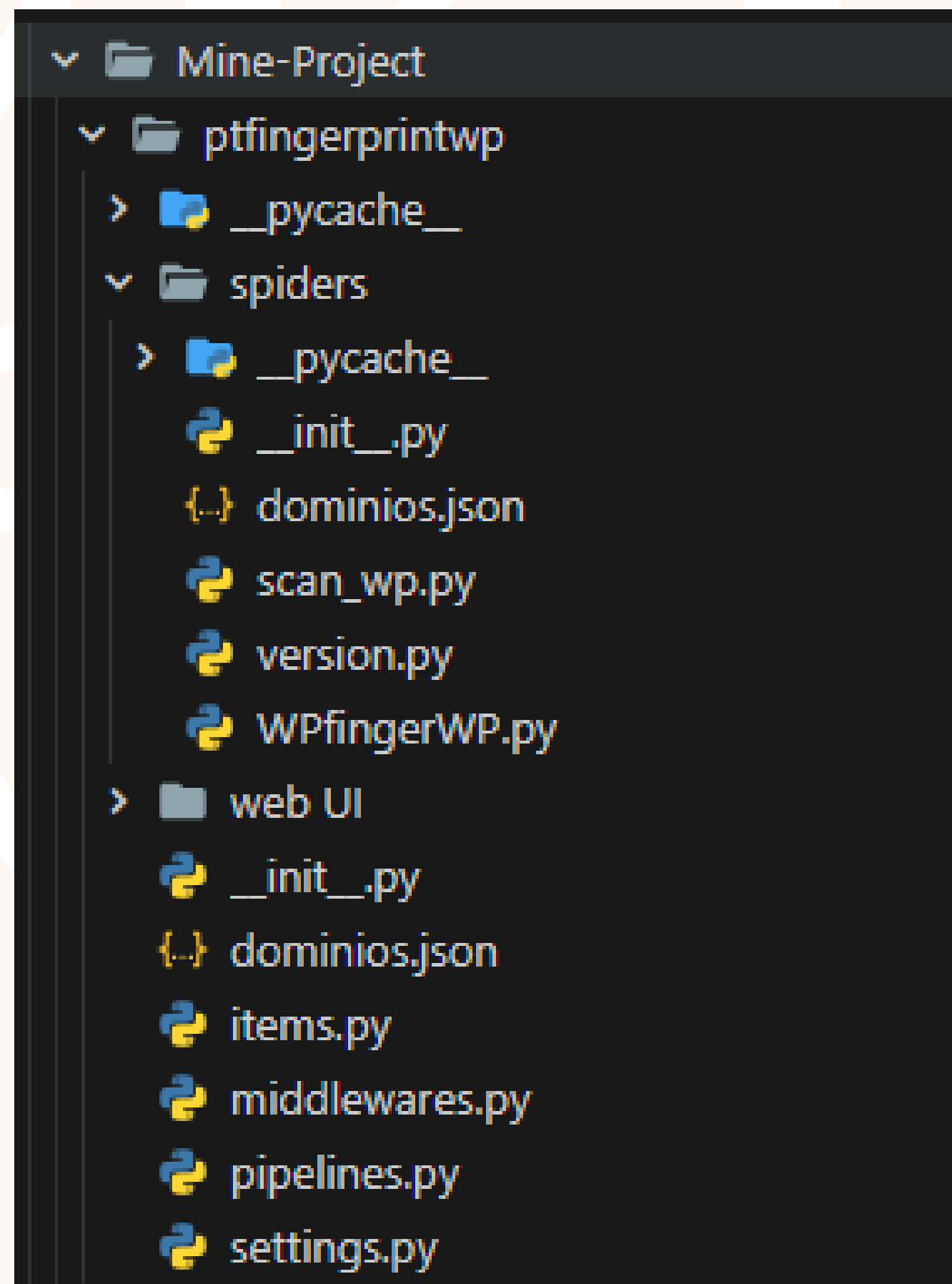
1. Desenvolver um crawler em Python utilizando técnicas de "Google Hacking" para identificar websites portugueses (.pt) desenvolvidos com WordPress.
2. Identificar passivamente a versão do WordPress de cada site e armazenar essas informações em uma base de dados.
3. Criar um crawler adicional para verificar se um website específico está utilizando o WordPress e qual é a sua versão.
4. Desenvolver uma interface web para os crawlers.

Conceitos importantes

Antes de explicarmos o desenvolvimento do projeto, vamos apresentar alguns conceitos importantes que serão utilizados:

- ❑ **Crawler:** É um programa automatizado que percorre a internet de forma sistemática, coletando informações ao visitar diferentes páginas da web. Seu objetivo é descobrir, indexar e armazenar dados.
- ❑ **WordPress:** É um sistema de gerenciamento de conteúdo amplamente utilizado para criar e gerenciar sites e blogs. É um CMS de código aberto que permite criar, editar e publicar conteúdo na web sem conhecimento avançado de programação ou design.
- ❑ **Google Hacking:** : Técnica de pesquisa avançada do Google que permite encontrar informações ocultas, por meio de operadores de pesquisa específicos. Neste projeto, utilizaremos exemplos como *'site:.pt inurl:wp-content'*, *'site:.pt inurl:wp-login.php'* e *'site:.pt inurl:wp-admin'*.
- ❑ **Scrapy:** Framework de web scraping em Python que oferece uma maneira eficiente e flexível de extrair dados de sites de forma automatizada, facilitando a colheita de informações da web.

Arquitetura do projeto

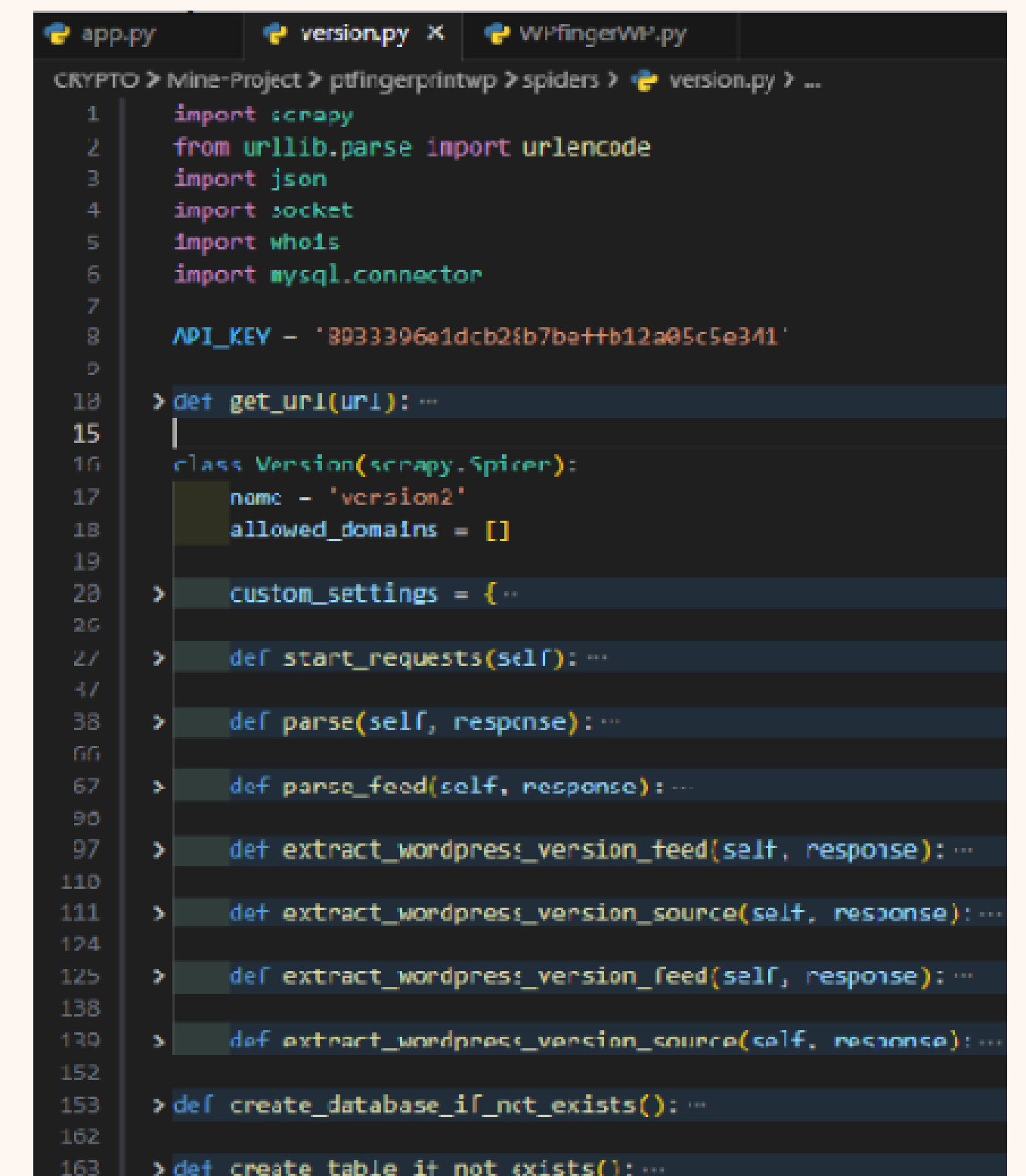


Desenvolvimento do crawler PT WordPress Fingerprint

Para implementar o crawler, dividimos o processo em duas partes principais: o Spider pt_fingerprint_wp e o Spider version.

Spider pt_fingerprint_wp:

- ❑ Envia uma solicitação ao mecanismo de busca do Google com consultas específicas para filtrar os resultados relacionados a sites em português que utilizam o WordPress.
- ❑ Utiliza técnicas de "Google hacking" como 'site:.pt inurl:wp-content', 'site:.pt inurl:wp-login.php', 'site:.pt inurl:wp-admin' e 'site:.pt "Powered by WordPress"'.
- ❑ Extrai os domínios dos sites encontrados e armazena os resultados para uso posterior.

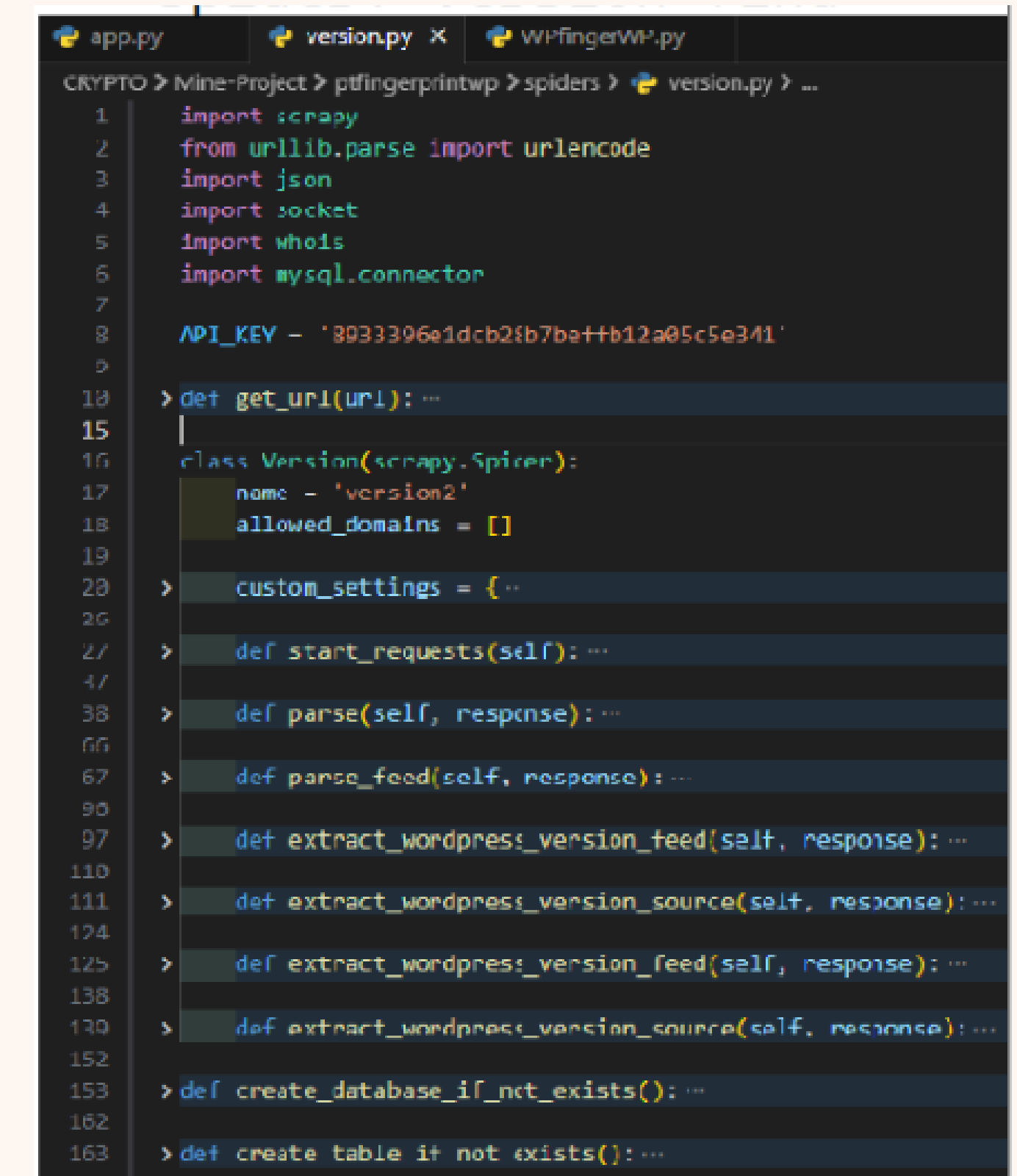


```
app.py version.py x WPfingerprintWP.py
CRYPTO > Mine-Project > ptfingerprintwp > spiders > version.py > ...
1 import scrapy
2 from urllib.parse import urlencode
3 import json
4 import socket
5 import whois
6 import mysql.connector
7
8 API_KEY = '8033396e1dcb28b7be4b12a05c5e341'
9
10 > def get_url(url): ...
11
12
13
14
15 class Version(scrapy.Spider):
16     name = 'version2'
17     allowed_domains = []
18
19
20 > custom_settings = {...
21
22
23
24
25
26 > def start_requests(self): ...
27
28
29
30
31
32
33
34
35
36
37
38 > def parse(self, response): ...
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57 > def parse_feed(self, response): ...
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97 > def extract_wordpress_version_feed(self, response): ...
98
99
100
101
102
103
104
105
106
107
108
109
110
111 > def extract_wordpress_version_source(self, response): ...
112
113
114
115
116
117
118
119
120
121
122
123
124
125 > def extract_wordpress_version_feed(self, response): ...
126
127
128
129
130
131
132
133
134
135
136
137
138
139 > def extract_wordpress_version_source(self, response): ...
140
141
142
143
144
145
146
147
148
149
150
151
152
153 > def create_database_if_not_exists(): ...
154
155
156
157
158
159
160
161
162
163 > def create_table_if_not_exists(): ...
```


Desenvolvimento dos crawlers PT WordPress Fingerprint

Spider version:

- ❑ Obtém os resultados guardados pelo spider pt_fingerprint_wp.
- ❑ Com esses domínios, envia solicitações individuais para cada domínio encontrado a fim de confirmar se o site realmente utiliza o WordPress.
- ❑ Caso seja confirmado o uso, extrai a versão específica do WordPress em uso.
- ❑ O spider também, coleta outras informações relevantes, como endereço IP e name_servers associados ao domínio.
- ❑ Armazena todas as informações coletadas numa base de dados para análise posterior.



```
app.py version.py x WPfingerWP.py
CRYPTO > Mine-Project > ptfingerprintwp > spiders > version.py > ...
1 import scrapy
2 from urllib.parse import urlencode
3 import json
4 import socket
5 import whois
6 import mysql.connector
7
8 API_KEY = '3D33396e1dcb2f6b7ba1b12a05c5e341'
9
10 > def get_url(url): ...
15 |
16 class Version(scrapy.Spider):
17     name = 'version2'
18     allowed_domains = []
19
20 > custom_settings = { ...
26
27 > def start_requests(self): ...
47
38 > def parse(self, response): ...
66
67 > def parse_feed(self, response): ...
90
97 > def extract_wordpress_version_feed(self, response): ...
110
111 > def extract_wordpress_version_source(self, response): ...
124
125 > def extract_wordpress_version_feed(self, response): ...
138
139 > def extract_wordpress_version_source(self, response): ...
152
153 > def create_database_if_not_exists(): ...
162
163 > def create table if not exists(): ...
```

Desenvolvimento do crawler find WordPress in website

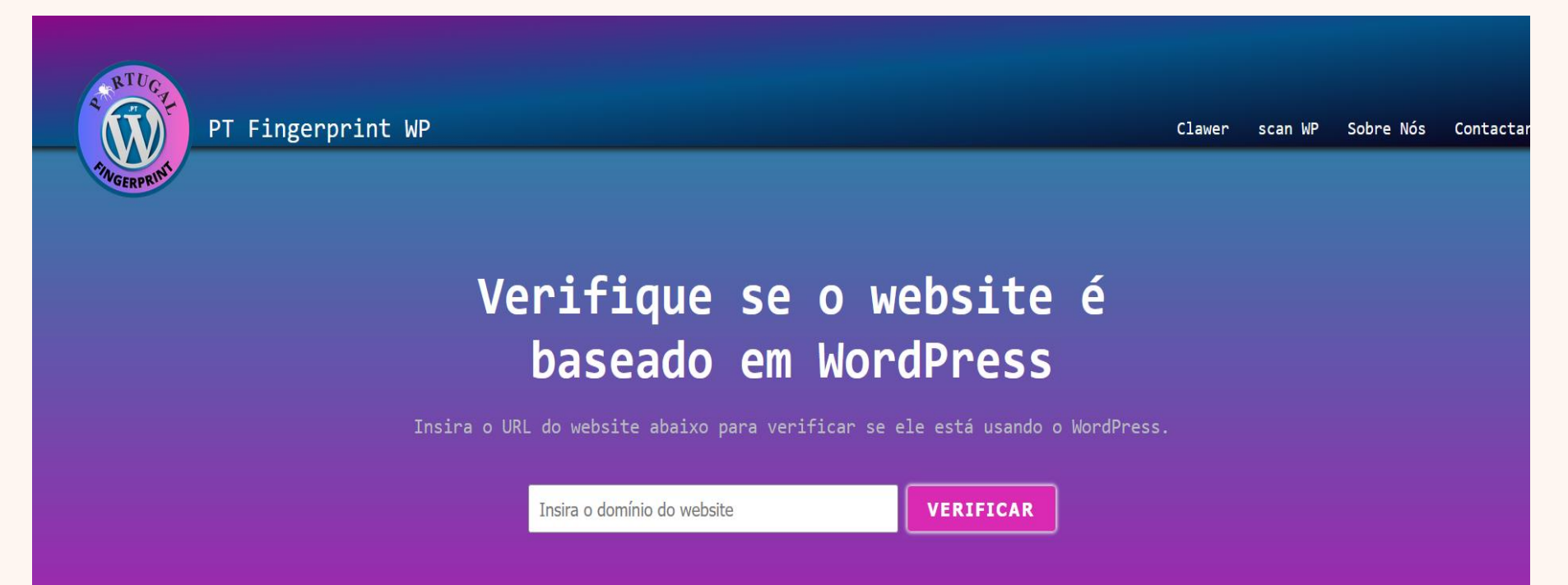
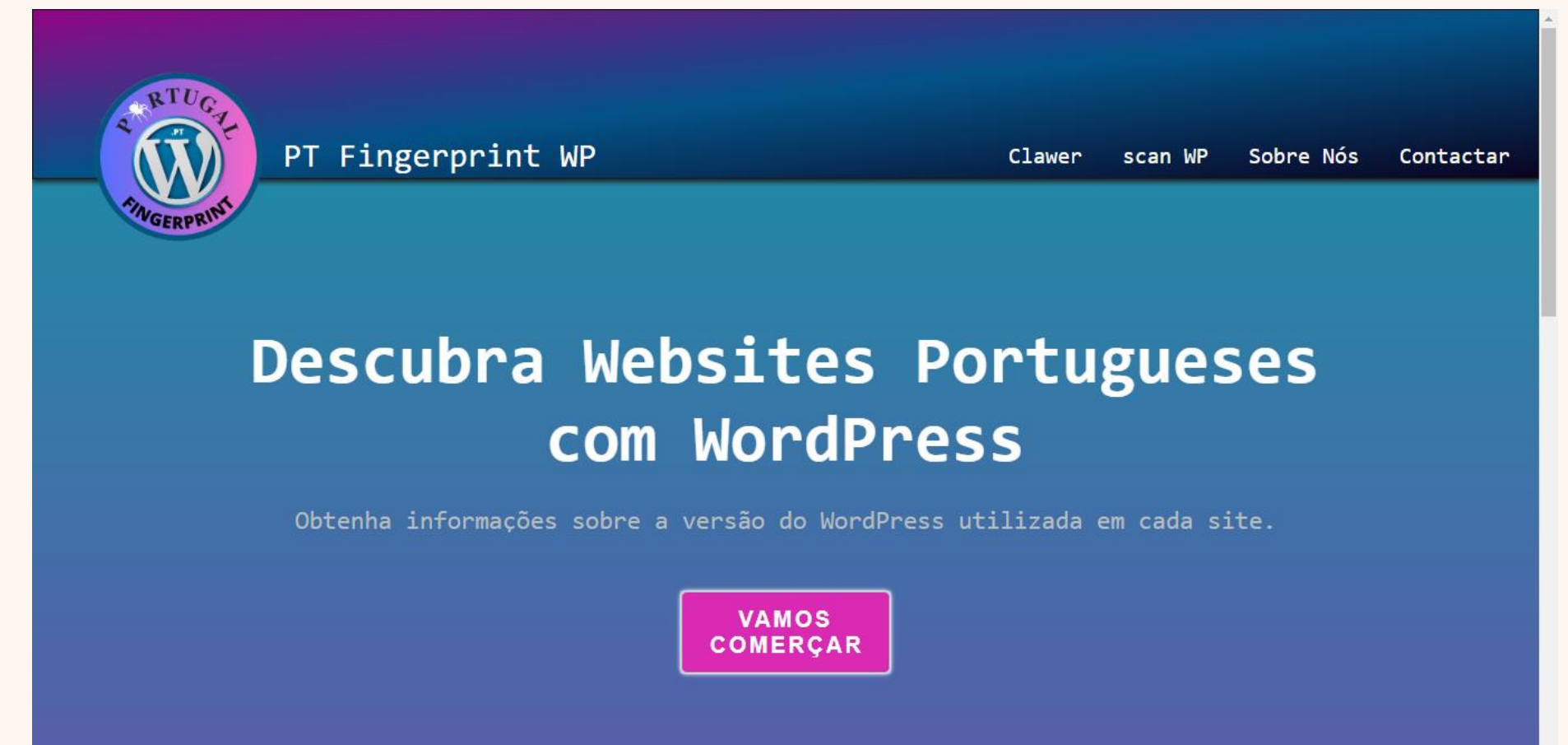
O crawler desenvolvido segue as seguintes etapas:

- ❑ Envio de solicitação: O spider envia uma solicitação ao domínio fornecido e analisa o código fonte do site.
- ❑ Verificação do uso do WordPress: O crawler verifica se o site utiliza o WordPress, levando em consideração vários aspectos do domínio.
- ❑ Extração da versão do WordPress: Caso seja identificado o uso do WordPress, o crawler extrai a versão específica utilizada pelo site.
- ❑ Coleta de informações relevantes: O crawler retorna informações como endereço IP, servidores de nomes, administradores e outros detalhes do domínio.

Essas etapas permitem ao crawler identificar e extrair informações sobre sites que utilizam o WordPress.

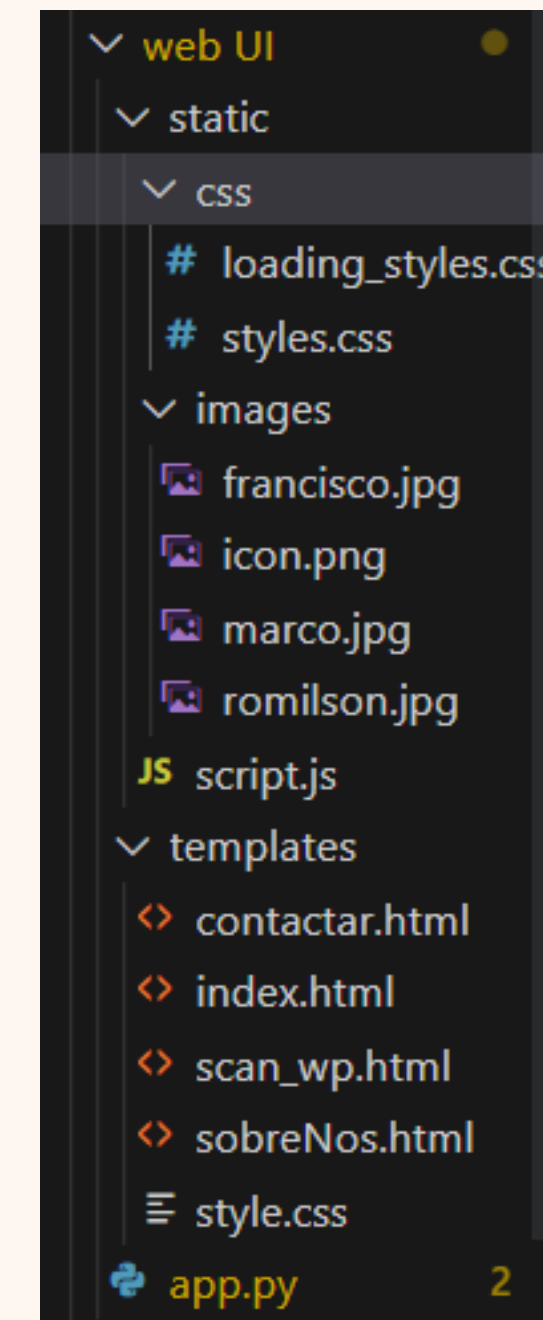
Web interface

- ❑ Além dos objetivos inicialmente propostos pelo docente, decidimos criar uma web interface para tornar o uso dos nossos crawlers mais conveniente.
- ❑ Desenvolvemos a web interface utilizando tecnologias como HTML, CSS, JavaScript e o servidor Python Flask. Com isso, oferecemos aos utilizadores uma interface amigável e intuitiva para a utilização dos nossos crawlers.



Server "app"

- ❑ Utilizamos a biblioteca Flask para desenvolver o servidor, a fim de conectar a interface web aos crawlers. Para isso, criamos rotas e endpoints específicos que permitem executar as funcionalidades da interface.
- ❑ Quando um utilizador faz uma solicitação através da interface web, o servidor responde executando o crawler correspondente. No caso específico do crawler que verifica o domínio de um site, o servidor passa o domínio inserido pelo utilizador como parâmetro durante a execução.
- ❑ Após a conclusão da execução, o servidor recupera o resultado armazenado no banco de dados e o exibe na interface web



```
1 from flask import Flask, render_template, request
2 import subprocess
3 import json
4 import os
5 import sys
6 import mysql.connector
7
8 app = Flask(__name__)
9
10 # Função para ler as informações dos domínios do banco de dados
11 > def read_domain_info_from_database(): ...
12
13
14
15
16
17
18 @app.route('/', methods=['GET', 'POST'])
19 > def index(): ...
20
21
22
23
24
25
26
27 @app.route('/sobrenos')
28 def sobrenos():
29     return render_template('sobreNos.html')
30
31
32
33 @app.route('/contato')
34 def contato():
35     return render_template('contactar.html')
36
37
38
39 @app.route('/scan_wp', methods=['GET', 'POST'])
40 > def scan_wp(): ...
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95 if __name__ == '__main__':
96     app.run()
97
```


Resultado do crawler PT WordPress Fingerprint

- ❑ Neste exemplo o crawler “PT WordPress Fingerprint” procura os sites portugueses desenvolvidos com o wordpress, guarda os domínios dos sites, o endereço ip, a versão do wordpress na qual foram desenvolvidas e o name server, guarda os resultados na base de dados e posteriormente apresenta os resultados na web interface.


	id	domain	ip	version	host	name_servers
<input type="checkbox"/> Editar Copiar Apagar	1	www.farmaciaarade.pt	89.26.249.46	N/A	a.cp.cloudlink.pt	elle.ns.cloudflare.com, elmo.ns.cloudflare.com
<input type="checkbox"/> Editar Copiar Apagar	2	gumportugal.pt	109.71.42.35	5.9.7	secundus.motioncreator.net	ns5.motioncreator.pt, ns6.motioncreator.pt
<input type="checkbox"/> Editar Copiar Apagar	3	www.urgencias.pt	94.46.14.35	N/A	cp52.webserver.pt	ns2.mydnspt.net, ns1.mydnspt.net, ns8.mydnspt.net,...
<input type="checkbox"/> Editar Copiar Apagar	4	b-lizzard.pt	195.22.8.67	6.2.2	hati.dnshati.com	ns1.dnshati.com, ns2.dnshati.com
<input type="checkbox"/> Editar Copiar Apagar	5	dupladiabolika.pt	88.198.36.249	5.4.13	249.purpleprofile.pt	ns1.purpleprofile.com, ns2.purpleprofile.com
<input type="checkbox"/> Editar Copiar Apagar	6	adegadepetiscos.pt	193.70.24.82	6.2.2	cluster027.hosting.ovh.net	dns11.ovh.net, ns11.ovh.net
<input type="checkbox"/> Editar Copiar Apagar	7	revistadois.pt	94.46.13.220	N/A	sv01.sulinformacao.pt	ns1.dnscpanel.com, ns2.dnscpanel.com, ns3.dnscpane...
<input type="checkbox"/> Editar Copiar Apagar	8	oeirasdigital.pt	130.185.84.150	5.4.13	cp12.webserver.pt	ns1.amenworld.com, ns2.amenworld.com
<input type="checkbox"/> Editar Copiar Apagar	9	mappingout.iscte-iul.pt	193.136.189.103	6.2	afdevlives.iscte-iul.pt	dns1.iscte.pt, dns3.iscte.pt, ns02.fccn.pt
<input type="checkbox"/> Editar Copiar Apagar	10	international.uac.pt		N/A		ns-1572.awsdns-04.co.uk, ns-732.awsdns-27.net, ns-...
<input type="checkbox"/> Editar Copiar Apagar	11	magmastudio.pt		6.2.2		ns2.wp-ns.com, ns1.wp-ns.com
<input type="checkbox"/> Editar Copiar Apagar	12	lvm.pt	188.93.230.91	6.1.3	lvm.ibername.com	dns1.ibername.com, dns2.ibername.com

 PT Fingerprint WP				
Foi encontrado 224 sites Portugueses que usan WordPress				
Resultados do Clawer				
Domínio	IP	Version	Host	Name Servers
www.farmaciaarade.pt	89.26.249.46	N/A	a.cp.cloudlink.pt	• elle.ns.cloudflare.com • elmo.ns.cloudflare.com
gumportugal.pt	109.71.42.35	5.9.7	secundus.motioncreator.net	• ns5.motioncreator.pt • ns6.motioncreator.pt
www.urgencias.pt	94.46.14.35	N/A	cp52.webserver.pt	• ns2.mydnspt.net • ns1.mydnspt.net • ns8.mydnspt.net • ns7.mydnspt.net
b-lizzard.pt	195.22.8.67	6.2.2	hati.dnshati.com	• ns1.dnshati.com • ns2.dnshati.com
dupladiabolika.pt	88.198.36.249	5.4.13	249.purpleprofile.pt	• ns1.purpleprofile.com • ns2.purpleprofile.com
adegadepetiscos.pt	193.70.24.82	6.2.2	cluster027.hosting.ovh.net	• dns11.ovh.net • ns11.ovh.net
revistadois.pt	94.46.13.220	N/A	sv01.sulinformacao.pt	• ns1.dnscpanel.com • ns2.dnscpanel.com • ns3.dnscpanel.com
oeirasdigital.pt	130.185.84.150	5.4.13	cp12.webserver.pt	• ns1.amenworld.com • ns2.amenworld.com

Resultados do crawler Find wordpress in web

❑ No primeiro exemplo o crawler “find
wordpress” conseguiu identificar que o site
passado como parâmetro não é feito com o
wordpress

❑ No segundo exemplo o crawler “find
wordpress” conseguiu identificar que o site
passado como parâmetro é feito com o
wordpress



PT Fingerprint WP

Resultado da Verificação

pinterest.pt

❌ O website não é baseado em WordPress, ou já está escondido

Detalhes do domínio

Domínio: pinterest.pt

Data de criação: 2013-02-05 08:15:20

Data de expiração: 2024-02-04 23:59:20

Registrante

Nome: DNStination, Inc.

Endereço: 3450 Sacramento Street, Suite 405, CA, 94118

E-mail: admin@dnstinations.com, ccops@markmonitor.com

Administrador

Nome: MarkMonitor Inc.

Endereço: 2150 S Bonito Way Suite 150, Meridian, 83642

E-mail: ccops@markmonitor.com

Servidores de Nomes

- ns5.pinterest.com
- ns6.pinterest.com
- ns9.pinterest.com
- ns10.pinterest.com



PT Fingerprint WP

Resultado da Verificação

b-lizzard.pt

✅ O website é baseado em WordPress WordPress 6.2.2

IP: 195.22.8.67

Domínio: b-lizzard.pt

Data de criação: 2006-02-16 00:00:00

Data de expiração: 2023-07-01 23:59:00

Registrante

Nome: B Lizzard - Criatividade, Comunicacao e Servicos Lda

Endereço: Rua Joaquim Rocha Cabral Quinta dos Barros 14-A, Lisboa, 1600-086

E-mail: geral@b-lizzard.pt, multimedia@b-lizzard.pt

Administrador

Nome: PORTAL PME, LDA

Endereço: Rua Aristides de Sousa Mendes, 4C - Escrit. 4, Lisboa, 1600-413

E-mail: dns@hostinet.com, dns@hostingportugal.pt, dns@pme.pt

Servidores de Nomes

- ns1.dnshati.com
- ns2.dnshati.com

Conclusão

O desenvolvimento do crawler e da interface web foi desafiador, porém gratificante. Aprendemos Python, enfrentamos obstáculos e adquirimos habilidades sólidas ao longo do processo. O crawler foi capaz de verificar se os sites utilizavam o WordPress, extrair suas versões e coletar informações relevantes. A interface web permitiu aos utilizadores explorar os resultados.

No geral, esse projeto nos proporcionou aprendizado, trabalho em equipe e superação de desafios.



FIM


ipvcestg

ERSC

ENGENHARIA DE REDES E
SISTEMAS DE COMPUTADORES