IPBeja

INSTITUTO POLITÉCNICO DE BEJA

Sistema de apoio à promoção do turismo rural
Fase de Webscraping

Gonçalo Amaro – 17440,
Pedro Tomás – 18962,
Vítor Abreu – 18966

15 de Dezembro, 2021

Conteúdo

1 Introdução

O objetivo desta fase de trabalho é a recolha da informação que será utilizada no decorrer do projeto, isto é, o "web-scrapping". Simplificando, o processo de "web-scrapping"é simplesmente a recolha de dados de uma forma automatizada, no nosso caso foi utilizado a linguagem python juntamente com algumas bibliotecas como a "Beautifullsoup4". Todas as informações recolhidas foram apenas dentro da localidade de Beja e entre todos os resultados alguns até podem ser comuns, umas vez que todos nós recolhemos por exemplo as "reviews"dos hotéis, atrações e restaurantes.

2 Planeamento

2.1 Divisão de Tarefas

Para a realização desta fase de trabalho o grupo decidiu dividir as tarefas e organizá-las a partir da plataforma "Trello"(https://trello.com/b/PIApdmTA/pi-2021-22) uma vez que o grupo já se encontrava a usar a mesma e já temos mais á vontade. Assim sendo, o site "TripAdvisor"foi realizado pelo aluno Gonçalo Amaro, o "Booking"pelo Pedro Tomás e o "Zomato"pelo Vítor Abreu e todos conseguiram aceder aos seus devidos "websites"e adquirir as informações possíveis.

2.2 Tecnologias Usadas

Na realização do web-scrapping foi desenvolvido um ambiente virtual de python3 para realizar os scripts que iriam recolher as informações.

Como forma de organizar todos os pacotes e possíveis atualizações de bibliotecas dentro do código também foi gerado um ficheiro .txt denominado "requirements"que atualizavamos e usavamos sempre que um dos elementos do grupo iria realizar o seu trabalho.

A linguagem optata para a construção dos scripts foi o python já que é uma das mais acessíveis linguagens de programação disponíveis devido à sua simples síntaxe e não ser complicada e também pela vasta quantidade de bibliotecas disponibilizadas, que mais tarde foram bastante úteis na realização do projeto.

Para finalizar, todos os ficheiros foram guardados em formato .csv uma vez que a quantidade de informação era grande e seria fácil de a organizar no formato indicado.

2.2.1 Ambientes Virtuais de Python

2.2.2 Bibliotecas de Python usadas

3 TripAdvisor

3.1 Estratégia

3.2 Desenvolvimento

3.2.1 Hotéis

	Hotel	Estrelas	Preço
0	Pousada Convento Beja	"4,5 de 5 bolhas"	100
1	Vila Galé Clube de Campo	"4,5 de 5 bolhas"	99
2	Herdade dos Grous	"4,5 de 5 bolhas"	130
3	Hotel Bejense	4 de 5 bolhas	63
4	Herdade do Vau	"4,5 de 5 bolhas"	85
5	Herdade Da Diabroria	4 de 5 bolhas	76
6	Hotel Melius	4 de 5 bolhas	75
7	BejaParque Hotel	"3,5 de 5 bolhas"	80
8	Hotel São Domingos	4 de 5 bolhas	55
9	Maria‘s Guesthouse	5 de 5 bolhas	85 81
10	Hotel Santa Bárbara	4 de 5 bolhas	59
11	Beja Hostel	"3,5 de 5 bolhas"	50
12	Império romano guest house	"4,5 de 5 bolhas"	67
13	Guest House Stories	5 de 5 bolhas	50
14	Monte Das Beatas - Alojamento Local	5 de 5 bolhas	50
15	Hospedaria Santa Maria	3 de 5 bolhas	36
16	Aljana Guest House	"4,5 de 5 bolhas"	99
17	Sesmarias Turismo Rural & SPA	5 de 5 bolhas	90
18	Monte da Floresta B&B	"3,5 de 5 bolhas"	80 76
19	Casa de Pedrogao	4 de 5 bolhas	54 51
20	Hotel Santa Clara	5 de 5 bolhas	58
21	Herdade das Barradas da Serra	5 de 5 bolhas	125
22	Paradise In Portugal	5 de 5 bolhas	79
23	Villa Extramuros	5 de 5 bolhas	
24	Albergaria Do Calvario	5 de 5 bolhas	

3.2.2 Atrações

	Attraction
0	Castelo de Beja
1	Museu Regional de Beja (Museu Rainha D. Leonor)
2	Nucleo Museologico
3	Casa de Santa Vitória
4	Igreja de Nossa Senhora Dos Prazeres E Museu Episcopal
5	Ruínas Romanas de Pisões
6	Museu Visigotico–Igreja de Santo Amaro
7	Jardim Gago Coutinho e Sacadura Cabral
8	Sé Catedral de Beja / Igreja de São Tiago
9	Museu Jorge Vieira/Casa Das Artes
10	Porta de Évora - Arco romano de Beja
11	Pelourinho de Beja
12	Igreja de Santa Maria da Feira
13	Igreja do Salvador
14	Igreja da Misericórdia
15	Estátua da Rainha Dona Leonor
16	Igreja do Carmo
17	Ermida de Santo André
18	Igreja de Nossa Senhora do Pé da Cruz
19	Ermida de Santo Estêvão
20	Bairro da Mouraria
21	Janela Manuelina
22	Arcadas da Praça da República
23	Arco das portas de Avis
24	Monumento ao Prisioneiro Político Desconhecido
25	Palácio dos Maldonados
26	Convento de Santo António em Beja
27	Colégio dos Jesuítas de Beja
28	Piscina Descoberta Municipal de Beja
29	Passo da Rua da Ancha

3.2.3 Restaurantes

	Restaurant
0	Íntimo restaurante
1	Restaurante Dom Dinis
2	Herdade dos Grous Restaurante
3	Adega Tipica Restaurante
4	Bifanas do Márinho
5	Pulo Do Lobo
6	Toi Farois
7	Restaurante Sabores Do Monte
8	Pizzaria Milano
9	Pizaria e Restaurante Mediterrâneo Dona Maria
10	Frango à Guia
11	Casa de Pasto - Tem Avondo
12	Restaurante Espelho D'Água
13	Hamburgueria da Avenida
14	O Arbitro
15	Adega do Castelo - Museu do Vinho
16	Pinguinhas - Tapas e Petiscos
17	Taberna A Pipa
18	Luiz Da Rocha
19	Restaurante Pousada São Francisco
20	Malhadinha Restaurant Wine & Gourmet
21	A Ilha Do Peixe
22	Sabores do Campo
23	Art Deco
24	Os Bolos da Marisa
25	Restaurante Típico O Arcada
26	A Merenda Snack Bar Restaurante
27	A Pracinha
28	Restaurante Alcoforado
29	Restaurante A Lareira

3.3 Resultados

4 Booking

4.1 Estratégia

Para a realização do "web-scrapping"no "website"da Booking.com, inicialmente foi necessário a filtragem pelos hóteis apenas na localidade de Beja, uma vez ser o local que o grupo em conjunto decidiu optar para realizar todas as pesquisas num sítio em comum. Após ter o Booking a apresentar todos os resultados para os hóteis de Beja, foi recolhido o link que redireciona especificamente para esses resultados. Para aceder ás informações específicas de cada elemento da página e mais tarde aceder aos mesmos para retirar a informação pretendida, foi usado a ferramenta de "inspecionar a página"e assim descobrir os nomes das classes e todos os outros elementos que continham conteúdo importante para o projeto, como o nome dos hóteis, preço, classificação, número de comentários e alguns outros detalhes que podessem ser úteis.

Em seguida foi necessário realizar o "web-scrapping"das reviews de cada hotel, a realização desta parte foi um pouco mais difícil uma vez que para as reviews serem bem recolhidas era fulcral que o "web-scrapping"fosse realizado usando outro link, ou seja, foi retirado do site o prefixo de um novo link que seria o "https://www.booking.com/reviews/pt/hotel/"e baseando nos hóteis já retirados foi colocado o nome de cada um á frente do mesmo, criando assim um novo link que seria usado na realização do "web-scrapping"após a criação de um novo link para cada hotel, os processos foram semelhantes aos anteriormente feitos.

Para finalizar, os resultados foram todos guardados em ficheiros .csv para uma mais fácil visualização.

4.2 Desenvolvimento

Inicialmente foi feita a filtragem de apenas os hóteis de Beja.

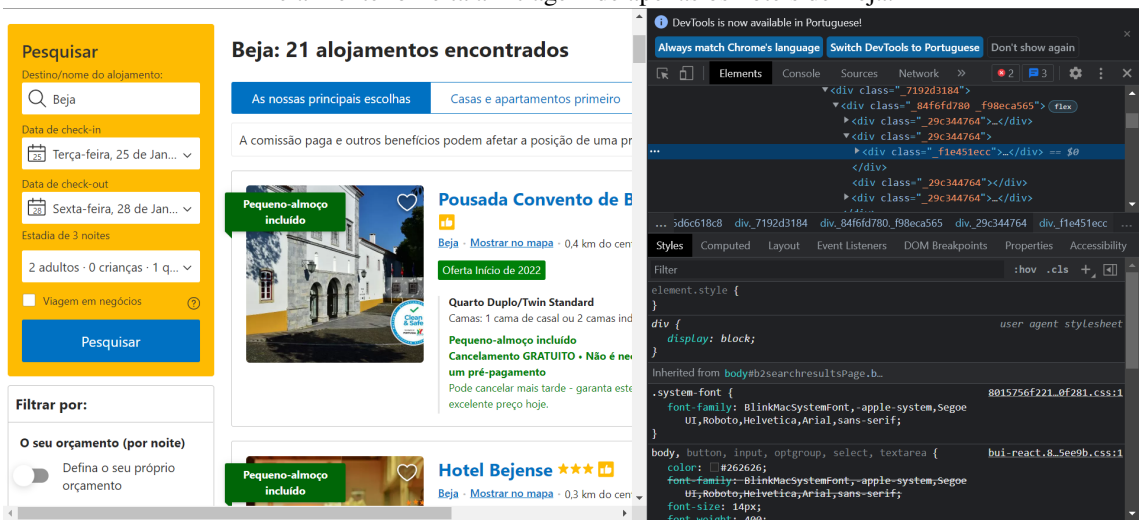

Figura 1: Site Booking.com ao usar a ferramenta inspecionar

No código foi implementado as bibliotecas BeautifulSoup para facilitar a tarefa de realizar o "web-scraping".

Figura 2: "Import"de algumas das bibliotecas necessárias

A partir do "website"ao inspecionar a página era possível retirar os headers que eram valores necessários na realização do "web-scrapping". Também é realizado o pedido HTTP e juntou-se a informação com a biblioteca "BeautifulSoup".

Figura 3: "Import"de algumas das bibliotecas necessárias

Foram criados diferentes arrays para receber as informações e posteriormente colocada a respetiva informação em cada um deles.

Figura 4: "Import"de algumas das bibliotecas necessárias

Devido a alguns "arrays"conterem mais informação, possivelmente devido a algum tipo de informação adicional que possa estar em algum hótel especificamente, para preveneir erros, foi feito um pequeno código para que todos os "arrays"contenham as mesmas dimensões.

```
# bad code
length = len(price)
if length > len(reviews):
    length = len(reviews)
if length > len(hotel):
    length = len(hotel)
if length > len(badge):
    length = len(badge)
if length > len(title):
    length = len(title)
```

Figura 5: "Import"de algumas das bibliotecas necessárias

Por fim todos os resultados contidos nos "arrays"foram guardados num ficheiro .csv denominado "listtable.csv".

```
d1 = {'Hotel': hotel[:length], 'Classificação': badge[:length],
      'Suma': title[:length], 'Avaliações': reviews[:length], 'Preço': price[:length]}
df = pd.DataFrame.from_dict(d1)
print(df)
df.to_csv('listtable.csv')
```

Figura 6: "Import"de algumas das bibliotecas necessárias

Alguns dos resultados dos hóteis em Beja.

	Hotel	Classificação
0	Hotel Bejense	8,4
1	Aljana Guest House Beja	9,3
2	BejaParque Hotel	8,1
3	Pousada Convento de Beja	8,7
4	Guest House Stories	8,7
5	Hotel Melius	8,1
6	Casa do Arco - Beja	9,4
7	Barrote Beja- Alojamento Local	9,7
8	Maria`s Guesthouse	9,4
9	Beja Garden	8,0
10	HI Beja - Pousada de Juventude	9,6
11	Casa do Sembrano	9,5
12	Casa Idalina Villa in Beja's beautiful country...	9,4
13	Império Romano Guest House	9,0
14	Monte das Beatas - Alojamento Local	8,7
15	Casa do Jardim	8,8

Figura 7: "Import"de algumas das bibliotecas necessárias

Construção dos links para realizar o "web-scraping"das "reviews"de cada hotel.

```
reviews_links = []
for link in soup.findAll('a', {'class': 'fb01724e5b'}):

    a = link['href']
    hotel = a.split('/')[5].split('?')[0]
    a = 'https://www.booking.com/reviews/pt/hotel/' + hotel
    reviews_links.append(a)
```

Figura 8: "Import"de algumas das bibliotecas necessárias

Foi realizado o pedido "HTTP"e juntado á biblioteca "BeautifulSoup"para aceder ás "reviews"de cada site e todos os valores foram salvos no formato .csv.

```
count = 0
allreviews = []

for link in reviews_links:
    try:
        response2 = requests.get(link, headers=headers)
        soup2 = BeautifulSoup(response2.content, 'lxml')
        for r in soup2.findAll('span', {'itemprop': 'reviewBody'}):
            try:
                rev = r.text
                allreviews.append(rev + '\n')
            except:
                pass
    except:
        pass
    count += 1
    if allreviews != []:
        seen = set()
        allreviews = [item for item in allreviews if not(
            tuple(item) in seen or seen.add(tuple(item)))]
        dfr = pd.DataFrame.from_dict({'Avaliações': allreviews})
        print(dfr)
        dfr.to_csv('hotel' + str(count) + '.csv')
        allreviews = []
```

Figura 9: "Import"de algumas das bibliotecas necessárias

Algumas das "reviews"de um dos sites disponíveis.

	Avaliações
0	de tudo , o edifício é lindíssimo \n
1	Gosto que as camas tenham lençol, além do edre...
2	Travesseiros um bom o outro pela hora da morte...
3	Amei a suntuosidade do local.\nQuarto silencio...
4	gostei de tudo\n
5	localização e espaço da Pousada\n
6	Temperatura quarto\n
7	Globalmente bem\n
8	Acessibilidade para a recepção e quartos\n
9	Localização da pousada, vista do quarto para o...
10	O quarto era arrumado demasiado tarde, por vol...
11	O quarto era extremamente acolhedor e limpo. A...
12	A rede wi-fi é fraca, a vista do quarto , a il...
13	Da amabilidade e simpatia de todo o pessoal es...
14	do estacionamento\n
15	De todo o espaço\n
16	A piscina devia ser limpa com maior regularida...
17	Da simpatia e amabilidade dos funcionários. As...
18	de nada\n
19	Da amabilidade dos funcionários, da tranquilid...
20	Nada a referir.\n
21	A pousada aproveita de forma exemplar as insta...
22	poucas espreguiçadeiras na piscina, mas são or...
23	O pequeno almoço, empregados e estadia no geral\n

Figura 10: "Import"de algumas das bibliotecas necessárias

4.2.1 Hotéis

	Hotel	Classificação	Preço
0	Hotel Bejense	"8,4"	189
1	Aljana Guest House Beja	"9,3"	330
2	BejaParque Hotel	"8,1"	255
3	Pousada Convento de Beja	"8,7"	270
4	Guest House Stories	"8,7"	135
5	Hotel Melius	"8,1"	242
6	Casa do Arco - Beja	"9,4"	210
7	Barrote Beja- Alojamento Local	"9,7"	255
8	Maria‘s Guesthouse	"9,4"	226
9	Beja Garden	"8,0"	90
10	HI Beja - Pousada de Juventude	"9,6"	327
11	Casa do Sembrano	"9,5"	330
12	Casa Idalina Villa in Beja’s beautiful countryside	"9,4"	210
13	Império Romano Guest House	"9,0"	150
14	Monte das Beatas - Alojamento Local	"8,7"	306
15	Casa do Jardim	"8,8"	270
16	Casa das Histórias	"8,9"	195
17	Casa Centro Histórico Beja - Castelo	"7,9"	180
18	Quinta do Castelo	"9,0"	214
19	Casa do Avô Zé	"8,1"	174
20	Suite na Praça da República	"8,7"	228
21	Herdade da Diabroria - Agroturismo	"8,8"	285
22	Herdade do Vau	"7,8"	270
23	Monte da Corte Ligeira	"8,3"	150

Tabela 1: Tabela com todos os hotéis de Beja retirados do Booking.com

4.3 Resultados

Nº Opinião	Opiniões
0	A localização é excelente assim como as condições do espaço. Local muito bem cuidado e apelativo. Fomos muito bem recebidos. A casa estava muito bem equipada. Muito obrigada, Catarina.
1	Nada digno de registo.
2	A simpatia da Sra Catarina foi fantástica. A casa e as acomodações corresponderam ás expetivas e relação qualidade preço foi perfeita. A localização é óptima, apesar de algum barulho das viaturas que passam junto ás janelas dos quartos. Beja é uma cidade fantástica e voltaremos concerteza. Recomendo.
3	Localização excelente, apartamento espaçoso, e totalmente equipado. O facto de ser uma construção antiga, cria um ambiente muito peculiar, além de que a expessura das paredes ajuda na questão da temperatura (estivémos no verÃ£o, portanto a casa não era quente apesar dos 30 e tal graus na rua).
4	A localização é impecável mesmo no centro histórico de Beja. Casa Limpa e organizada. Dona super prestável e e simpática.
5	De noite ouve-se o barulho da rua com muita facilidade.
6	A casa está muito perto do castelo e está muito bem decorada e limpa. Muito agradável!
7	O dono não tem culpa mas a zona não é muito bem frequentada á noite
8	A casa é muito gira e funcional!
9	Casa muito agradável e bem situada. Os anfitriães são simpáticos e disponíveis. Cozinha muito bem equipada e casa de banho excelente.
10	Da localização, da relação preço/qualidade que entendo adequada.
11	Localização no centro histórico tem vantagens (centralidade) e desvantagens (dificuldades de estacionamento).
12	Localização, decoração, conforto, organização do espaço e o gosto cuidado na decoração.
13	Não tinha microondas.
14	Localização, estacionamento perto, a decoração do alojamento, ter máquina de lavar roupa deu muito geito.
15	Virado para duas ruas públicas isso condiciona a entrada de luz e de ventilação em casa pela noção de segurança; Existem algumas infiltrações junto ao pavimento nalgumas divisões, fruto da data de construção e ser uma casa térrea; Haver imenso equipamento de cozinha, de limpeza, de menage, mas não haver uma pastilha para a máquina de lavar roupa (mero detalhe).
16	A casa em si é antiga e possui algumas divisões de formatos, alturas, níveis de pavimento e arcadas distintas, o que lhe dão um ar muito original; Está bem conservada/renovada, com uma cozinha bastante equipada (incluindo lavagem e limpeza), onde é possível de confeccionar, conservar e lavar. A localização na zona histórica de Beja, no meio do casario.
17	Poderia ter um microondas.
18	Muito bem localizado, anfitriões muito simpáticos. Muito interessante a manutenção de casa típica alentejana. Gostamos todos muito
19	Gostei de tudo. NÃ£o tenho do q reclamar.
20	O apartamento é uma graça. Super completo, amplo, decorado com muito bom gosto e jovial! Adorei!!!!!
21	Serviço conforto e localização.
22	A limpeza poderia ser mais cuidada. Não dar para ligar a chaleira eléctrica porque o quadro não aguentava. O fogão estar rachado não ofereceu muita confiança para cozinhar.
23	Da receção, da decoração, da temperatura da casa, o ser uma casa acolhedora situada no coração do centro histórico.
...	...
28	De tudo. Um lugar ótimo para uma família passar uns dias, a casa está bem situada e as pessoas muito simpáticas. Um lugar a repetir.

Tabela 2: Tabela de resultados de algumas "Reviews"para um dos hotéis (hotel18.csv)

5 Zomato

5.1 Estratégia

5.2 Desenvolvimento

5.2.1 Restaurantes

	Restaurante	Tipo	Preço
0	Adega Típica 25 de Abril	"Alentejana, Portuguesa"	25
1	Dom Dinis	"Bifes, Portuguesa"	30
2	Sushi Alentejano	"Sushi, Japonesa"	35
3	Herdade dos Grous	"Contemporânea, Portuguesa"	60
4	Pulo do Lobo	Portuguesa	25
5	Bar Parque da Vila Beringel	"Snacks, Bebidas"	12
6	Figa's	"Pizza, Italiana"	25
7	Cervejaria Portugalo	"Portuguesa, Bebidas, Petiscos"	25
8	Entre Arcos	"Portuguesa, Grelhados"	25
9	Aperitivo	"Bebidas, Portuguesa"	20
10	O Arbitro	Portuguesa	25
11	Café Central	"Snacks, Bebidas, Portuguesa"	15
12	Gatus Cervejaria Alentejana	"Alentejana, Portuguesa, Marisqueira"	25
13	Hamburgueria da Avenida	Hamburgueria	25
14	Casa de Pasto O Forno	"Snacks, Bebidas, Portuguesa"	12
15	Tennis Courts Club	Portuguesa	25
16	Cervejaria Mira Serra	"Marisqueira, Portuguesa"	40
17	Espelho d'Água	Portuguesa	25
18	TEM Avondo	"Portuguesa, Alentejana"	25
19	Menau	Portuguesa	25
20	Toy Faróis	"Portuguesa, Grelhados"	25
21	Taberna A Pipa	"Alentejana, Portuguesa, Petiscos"	25
22	Pizzeria Milano	"Pizza, Italiana, Portuguesa"	25
23	Dona Maria Deck	"Portuguesa, Petiscos, Snacks"	25
24	O Alemão	"Portuguesa, Petiscos, Alentejana"	25
25	Deliciosa Alvorada	"Portuguesa, Alentejana"	25
26	Moments IPDJ	Portuguesa	20
27	Bar Regional	"Bebidas, Snacks"	6
28	Sushizzaria	"Sushi, Hamburgueria, Pizza"	30
...	...	...	...
155	Vegetariano	Vegetariana	25

5.3 Resultados

6 Próximos Passos

Na seguinte fase de trabalho o grupo irá ter que começar a desenvolver os processos ETL (extract, transform, load) do trabalho já realizado, uma das tarefas será a adaptação de todo o conteúdo textual já armazenado.

7 Conclusão

Concluindo, este trabalho serviu para aprender o que realmente é o "web-scraping"e todos os usos que ele pode ter para a recolha de dados estruturados da web. Para além de aprender o que ele realmente é, também tivemos a oportunidade de trabalhar com ele e usá-lo num caso prático. Houveram algumas dúvidas principalmente para retirar algumas informações de alguns dos "websites"trabalhados e também na criação do ambiente virtual, porem todas as duvidas foram superadas graças ao trabalho em equipa e pesquisas online, para alem da ajuda da docente responsável pelo projeto. Por fim, achamos que o balanço desta parte do trabalho tenha sido bastante positiva.

8 Webgrafia

https://www.youtube.com/watch?v=PRkOFgNAkio
Booking.com