Relatório : Desafio Seazone - Estágio Dev Python

Francisco Silveira Burigo

1.Introdução

O projeto consiste em adquirir dados, por meio de web scraping, do site olx.com.br, sobre terrenos à venda em Florianópolis e região. Foi utilizado Python e suas bibliotecas, para conseguir adquirir os dados. Dividindo o código em duas principais partes, que serão melhor detalhadas à frente.

A segunda parte do projeto, é com os dados obtidos fazer uma análise, conseguindo assim encontrar tamanho médio dos terrenos e valores, por cada região encontrada.

2. Código Aquisição

Para conseguir adquirir os dados foi utilizado a linguagem Python, e dentro do código dele, que pode ser visto acessando https://github.com/FranciscoBurigo/OLX_scrapper, dividimos o código em dois.

A primeira parte do código, é utilizada apenas para acessar a parte de buscas do site olx, e assim salvar os links dos URLs do terrenos desejados em uma lista para ser utilizado na segunda parte.

Nesta primeira parte utilizamos a ferramenta **requests** junto com **Beautifulsoup**, utilizando como parser do **Beautifulsoup** o **LXML**, pois foi verificado que com este parser, foi muito mais rápido o processo, sem perda significativa de aquisição de dados. É interessante utilizar o comando **try** durante a aquisição, pois caso algum link o sistema apresente erro, ele consegue continuar adquirindo dados do próximo link, sem fazer com que o arquivo pare de rodar. Esta mesma ideia foi utilizada na segunda parte, evitando assim que caso surgisse algum problema na aquisição, não parasse o código.

Na segunda parte foi utilizado o **Selenium** e o **Beautifulsoup,** o que deixou esta parte muito mais demorada que a primeira, se fizermos um comparativo, a primeira parte roda em segundos, sendo no máximo alguns minutos, já a segunda leva algumas horas.

Utilizando as duas ferramentas, fizemos a varredura e adquirimos os dados necessários e que foram requisitados para a análise de dados. Todos os dados são salvos em um dicionário criado e depois exportados como .csv.

A ideia inicial desta parte era fazer a aquisição como na primeira, para ser mais rápido e eficiente, mas devido a alguns dados, que por causa da página ser renderizada, acabavam dando erro na hora de fazer a aquisição, assim a solução encontrada foi utilizar o **Selenium**. O principal dado a apresentar problema para fazer a aquisição foi o nome do vendedor.

3. Análise de Dados

Como o código em python exportava um arquivo .csv com os dados, utilizamos o **Google Sheets**, para fazer uma análise. Podemos ver o resultado da análise no final deste documento, ou acessando o Git e vendo as possibilidades para acessar.

Podemos ver que o número de anúncios analisados foi de 360, mas podemos ver que na tabela de aquisição, temos quase 420 dados. Isto ocorre, pois muitos dados vieram com o seu tamanho zerado, impossibilitando de fazer a análise.

No resultado apresentado no final deste documento, podemos ver as regiões como municípios, mas na planilha original ainda pode considerar os bairros, podendo ser mais preciso na análise.

Outro ponto interessante, é que quando fizemos a aquisição, havia 3 tipos de “terrenos”, terrenos e lotes, sítios e chácaras, e fazendas. Podendo assim, ser considerado somente o de maior interesse. Mas mesmo considerando todos, algumas cidades possuem apenas um ou dois dados, o que torna a análise muito fraca.

4. Melhorias

O código escrito ele considera um número x de páginas para varrer e não o número total, a melhor maneira, seria fazer uma varredura para descobrir quantas páginas de anúncios existem e assim fazer a varredura, cobrindo todos os dados o que não foi realizado neste projeto.

A segunda melhoria seria buscar uma outra solução para a aquisição de dados, pois quando utilizamos o selenium ficou extremamente lento, como foi a primeira vez que fiz **web scraping**, não tinha total conhecimento das ferramentas, com mais tempo,poderia ter tentado buscar soluções melhores e mais rápidas. Inclusive este foi um dos motivos de não varrer todas as páginas, pois para varrer metade levava em torno de 2-3h.

Uma terceira melhoria, seguindo ainda a linha de raciocínio, é fazer uma maior aquisição de dados, inclusive buscando em outros sites, para ter uma maior cobertura e podendo fazer análises mais completas, pois mesmo se dobrarmos os dados coletados, supondo que fizemos a varredura completa, teríamos cidades com 4-5 dados, o que é muito pouco para uma boa análise.

5.Conclusão

O projeto conseguiu entregar tudo que foi solicitado, mas pode ser melhorado e ampliado. A falta de prática com web scraping, fez com que eu perdesse bastante tempo no começo para entender e conseguir utilizar as ferramentas com confiança. Sobre o desafio fiquei muito animado em realizar, pois na universidade acabamos

ficando muita na teoria e pouca prática, já este desafio foi o contrário. Consegui aprender muito sobre esta área, que tenho bastante interesse, e não ficar como nas disciplinas apenas fazendo projetos que estão longe de projetos reais.

Município	*Bairro*	Nº de Anúncios	Tamanho médio [m²]	Tamanho Max [m²]	Tamanho Min [m²]	Preço médio por m²	Preço Max por m²	Preço Min por m²
Aguas Mornas		2	1440	1780	1100	R$ 189,79	R$ 297,75	R$ 81,82
Aguas Mornas Total		2	1440	1780	1100	R$ 189,79	R$ 297,75	R$ 81,82
Alfredo Wagner Total		2	8197	13394	3000	R$ 33,12	R$ 39,57	R$ 26,67
Antonio Carlos Total		3	369,3333333	379	360	R$ 663,53	R$ 750,00	R$ 501,82
Araranguа Total		6	834,1666667	3260	300	R$ 319,18	R$ 613,50	R$ 162,04
Balneario Arroio do Silva Total		5	342	510	300	R$ 185,49	R$ 266,67	R$ 127,45
Balneario Gaivota Total		25	337,52	524	264	R$ 439,64	R$ 2.133,33	R$ 93,33
Biguacu Total		20	423,55	1813	150	R$ 840,77	R$ 1.775,90	R$ 120,00
Botuvera Total		1	59000	59000	59000	R$ 3,39	R$ 3,39	R$ 3,39
Braco do Norte Total		1	360	360	360	R$ 333,33	R$ 333,33	R$ 333,33
Canelinha Total		3	37253,33333	70000	5760	R$ 16,55	R$ 27,78	R$ 9,72
Criciuma Total		11	686	1888	6	R$ 5.743,12	R$ 58.000,00	R$ 115,29
Florianopolis Total		115	3773,947826	177000	52	R$ 2.056,62	R$ 51.442,31	R$ 28,30
Garopaba Total		22	37156,63636	750000	335	R$ 850,06	R$ 3.086,73	R$ 0,43
Governador Celso Ramos Total		9	1299,888889	3650	252	R$ 636,55	R$ 1.714,29	R$ 191,78
Icara Total		1	1005	1005	1005	R$ 348,26	R$ 348,26	R$ 348,26
Imarui Total		2	12750	25000	500	R$ 79,80	R$ 136,00	R$ 23,60
Imbituba Total		19	16576,63158	291730	180	R$ 542,06	R$ 1.500,00	R$ 3,77
Jaguaruna Total		1	240	240	240	R$ 354,17	R$ 354,17	R$ 354,17
Laguna Total		1	450	450	450	R$ 400,00	R$ 400,00	R$ 400,00
Major Gercino Total		3	2713,333333	4000	1140	R$ 129,20	R$ 201,75	R$ 42,50
Palhoca Total		38	1517,631579	18514	12	R$ 1.757,68	R$ 36.666,67	R$ 13,89
Passo de Torres Total		10	308,7	420	210	R$ 332,65	R$ 450,00	R$ 216,67
Paulo Lopes Total		2	1525,5	2547	504	R$ 335,83	R$ 396,83	R$ 274,83
Pescaria Brava Total		2	453	456	450	R$ 245,00	R$ 245,00	R$ 245,00
Rancho Queimado Total		3	12204	30000	300	R$ 754,31	R$ 2.166,67	R$ 46,26
Santo Amaro da Imperatriz Total		2	2049	4000	98	R$ 2.330,92	R$ 4.591,84	R$ 70,00
Sao Joao do Sul Total		1	4000	4000	4000	R$ 75,00	R$ 75,00	R$ 75,00
Sao Jose Total		41	9151,146341	238627	200	R$ 682,55	R$ 3.452,38	R$ 11,65
Sao Ludgero Total		1	378	378	378	R$ 370,37	R$ 370,37	R$ 370,37
Sao Pedro de Alcantara Total		1	40000	40000	40000	R$ 15,00	R$ 15,00	R$ 15,00
Tijucas Total		2	7156	14000	312	R$ 141,42	R$ 256,41	R$ 26,43
Tubarao Total		4	370	400	300	R$ 348,24	R$ 375,00	R$ 300,00
Urussanga Total		1	360	360	360	R$ 413,89	R$ 413,89	R$ 413,89
Total geral		**360**	**6603,508333**	**750000**	**6**	**R$ 1.328,54**	**R$ 58.000,00**	**R$ 0,43**