Engenho de busca para smartphones

Equipe:
Carlos Henrique Caloete Pena (chcp)
Matheus Branco de Siqueira (mbs8)

O domínio

- O domínio escolhido possui muitos sites pela a internet o que facilita encontrar resultados relevantes
- Milhares de modelos de smartphones
- Interesse em obter informações sobre smartphones

Sites escolhidos

- Amazon
- Avenida
- Cissa Magazine
- Colombo
- Havan
- iByte
- Kabum
- Magazine Luiza
- Ricardo Eletro
- Taqi

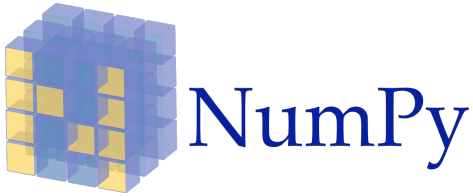
Crawler

Dificuldades

- Tempo de coleta
- Páginas não encontradas
- Instabilidade da conexão
- Permissão negada pelo robots
- Sites instáveis

404
Page not found

Bibliotecas utilizadas



Urllib

Threading

Estratégias utilizadas

1. BFS
2. Palavras-chave nos links
3. Classificador de links

1. BFS

- Prós
  - Coleta “rápida”
  - Simplicidade de implementação

- Contras
  - Coleta muito superficial
  - Links que se afastam do domínio
  - Retorna menos links válidos

url	pages	total_links	loss
havan	966	1027	5,94%
promobit	1016	1017	0,10%
amazon	1019	1027	0,78%
ibyte	1001	1005	0,40%
avenida	1002	1007	0,50%
cissamagazine	1105	1109	0,36%
magazineluiza	999	1015	1,58%
taqi	1011	1014	0,30%
colombo	972	1034	**6,00%**
kabum	1033	1035	0,19%

Resultado do bfs_crawler (bfs_stats)

2. Heurística com palavras-chave

- Prós
  - Retorna mais links relevantes
  - Menos links quebrados
  - Coleta mais profunda

- Contras
  - Coleta mais lenta

url	pages	total_links	loss
colombo	878	1005	**12,64%**
havan	1007	1007	0,00%
amazon	1025	1025	0,00%
cissamagazine	1034	1036	0,19%
taqi	999	1001	0,20%
ibyte	1001	1001	0,00%
magazineluiza	995	1003	0,80%
promobit	972	1004	3,19%
avenida	1018	1018	0,00%
kabum	1001	1002	0,10%

Resultado do heuristic_crawler (heuristic_stats)

3. Classificador de links

- Prós
  - Adaptação e aprendizagem
  - Consegue generalizar mais facilmente

- Contras
  - Requer base de dados para treinar
  - Maior custo em tempo

url	pages	total_links	loss
cissamagazine	936	1000	6,40%
avenida	1000	1000	0,00%
kabum	1000	1000	0,00%
amazon	982	1000	1,80%
ricardoeletro	772	1000	**22,80%**

Resultado do classifier_crawler (classifier_crawler)

Comparando Resultados

Estratégia	Páginas relevantes	Total de páginas	Harvest ratio	Ganho
BFS	685	10124	6,77%	-
Palavras-chave	2209	9930	22,25%	15,48%
Classificador de links	2302	4690	**49,08%**	**26,83%**

Classificador

Dificuldades

- Precisão das anotações
- Dados ruidosos
- Server Busy

store	link	text	label
magazinelu	**https://www.magazineluiza.com.br/celulares-e-telefor**	**Celular e Smartphone Magazi**	**0**
amazon	http://www.amazon.com.br/Celular-Xiaomi-Redmi-Vers%C3	Celular Xiaomi Redmi Note GB R	1
amazon	http://www.amazon.com.br/Celular-Xiaomi-Redmi-Vers%C3	Celular Xiaomi Redmi Note GB R	1
amazon	http://www.amazon.com.br/Celular-Redmi-Note-128GB-Nept	Celular Redmi Note GB GB Nep	1
amazon	https://www.amazon.com.br/Celular-Redmi-Note-128GB-Nep	Celular Redmi Note GB GB Nep	1
amazon	https://www.amazon.com.br/Celular-Redmi-Note-128GB-Nep	Celular Redmi Note GB GB Nep	1
amazon	https://www.amazon.com.br/Celular-Redmi-Note-128GB-Nep	Celular Redmi Note GB GB Nep	1
amazon	https://www.amazon.com.br/Celular-Redmi-Note-128GB-Nep	Celular Redmi Note GB GB Nep	1
amazon	https://www.amazon.com.br/Celular-Redmi-Note-128GB-Nep	Celular Redmi Note GB GB Nep	1
amazon	https://www.amazon.com.br/Celular-Xiaomi-Lite-Global-128G	Celular Xiaomi Mi Lite Global Dual	1
amazon	https://www.amazon.com.br/Celular-Xiaomi-Lite-Global-128G	Celular Xiaomi Mi Lite Global Dual	1
amazon	https://www.amazon.com.br/Celular-Xiaomi-Lite-Global-128G	Celular Xiaomi Mi Lite Global Dual	1
amazon	https://www.amazon.com.br/Celular-Xiaomi-Lite-Global-128G	Celular Xiaomi Mi Lite Global Dual	1
amazon	https://www.amazon.com.br/Celular-Xiaomi-Lite-Global-128G	Celular Xiaomi Mi Lite Global Dual	1
amazon	https://www.amazon.com.br/Celular-Xiaomi-Lite-Global-128G	Celular Xiaomi Mi Lite Global Dual	1
amazon	https://www.amazon.com.br/Celular-Xiaomi-Lite-Global-128G	Celular Xiaomi Mi Lite Global Dual	1
amazon	https://www.amazon.com.br/Celular-Xiaomi-Lite-Global-128G	Celular Xiaomi Mi Lite Global Dual	1
amazon	https://www.amazon.com.br/Celular-Xiaomi-MI-9T-Dual/dp/B	Celular Xiaomi MI T Dual GB Pre	1
amazon	https://www.amazon.com.br/Celular-Xiaomi-MI-9T-Dual/dp/B	Celular Xiaomi MI T Dual GB Pre	1
amazon	https://www.amazon.com.br/Celular-Xiaomi-Vers%C3%A3o-	Celular Xiaomi Mi Play GB GB RA	1
amazon	https://www.amazon.com.br/Celular-Xiaomi-Vers%C3%A3o-	Celular Xiaomi Mi Play GB GB RA	1
amazon	https://www.amazon.com.br/Smartphone-Motorola-Moto-XT1	Smartphone Motorola Moto G XT	1
amazon	https://www.amazon.com.br/Smartphone-Motorola-Moto-XT1	Smartphone Motorola Moto G Plus	1
amazon	https://www.amazon.com.br/Smartphone-Samsung-Galaxy-	Smartphone Samsung Galaxy Note	1
amazon	https://www.amazon.com.br/Smartphone-Samsung-Galaxy-	Smartphone Samsung Galaxy Note	1
amazon	https://www.amazon.com.br/Smartphone-Motorola-Action-XT	Smartphone Motorola One Action XT	1
amazon	https://www.amazon.com.br/Smartphone-Motorola-Action-XT	Smartphone Motorola One Action XT	1
amazon	https://www.amazon.com.br/Smartphone-Motorola-Action-XT	Smartphone Motorola One Action XT	1
amazon	https://www.amazon.com.br/Smartphone-Motorola-Action-XT	Smartphone Motorola One Action XT	1
amazon	https://www.amazon.com.br/Smartphone-Motorola-Action-XT	Smartphone Motorola One Action XT	1

Estratégias utilizadas

Dado uma página fazer a coleta de texto

- Link
- Cabeçalho
- Dados dentro de tabelas

Sistema operacional	android
Baterias ou pilhas	1 Polímero de lítio baterias ou pilhas necessárias (inclusas).
Modelo	REDMI-NT7-128-BLU
Tecnologia sem fio	HSDPA, 4G, 3G, GSM, Wifi, 2G
Tecnologia de conexão	4G, 3G, Wifi, Bluetooth Wireless, 2G
Cor	Neptune Blue
Classificação de potência da bateria ou pilha	4000 milliampere_hour
Tempo de conversa	23 horas
Tamanho da tela	6.3 polegadas
Tamanho da memória RAM instalada	4 GB

Metodologia de Testes

- Foram anotados manualmente mais de 2600 páginas de 10 sites diferentes
- Foi utilizado a divisão Leave One Group Out, para obter as métricas
  - Onde o grupo da página seria o domínio
- Foram testadas seis máquinas de aprendizagem
  - GaussianNb, DecisionTree, SVC, MLP, LogisticRegression, RandomForest
- Com três variações no pré processamento dos dados
  - Usar apenas palavras em caixa baixa
  - Eliminar Stop words
  - Utilizar TF IDF

Variações

				GaussianNb			
Lower	Stop Words	TF IDF		Acuracy	Precision	Recall	Train Time
			Mean	0.8336	0.8097	0.7549	0.1257
0	0	0	Std	0.2000	0.2417	0.2345	0.0123
			Mean	0.7658	0.7442	0.7225	0.1384
0	0	1	Std	0.2383	0.2106	0.2403	0.0307
			Mean	0.8714	0.8284	0.7881	0.1372
0	1	0	Std	0.1735	0.2432	0.2364	0.0129
			Mean	0.8286	0.7883	0.7729	0.0841
0	1	1	Std	0.1950	0.1990	0.2377	0.0193
			Mean	0.7616	0.7903	0.7347	0.0721
1	0	0	Std	0.2448	0.2013	0.2069	0.0096
			Mean	0.7470	0.7249	0.7285	0.3708
1	0	1	Std	0.2211	0.1920	0.2182	0.0326
			Mean	0.7748	0.7870	0.7325	0.0819
1	1	0	Std	0.2155	0.1966	0.2026	0.0113
			Mean	0.7642	0.7160	0.7335	0.0771
1	1	1	Std	0.2094	0.2115	0.2329	0.0103

Variações

				DecisionTree			
Lower	Stop Words	TF IDF		Acuracy	Precision	Recall	Train Time
			Mean	0.8670	0.8103	0.7993	0.3496
0	0	0	Std	0.1585	0.2178	0.2165	0.1135
			Mean	0.8215	0.8219	0.7822	0.3721
0	0	1	Std	0.2045	0.1782	0.1897	0.0936
			Mean	0.8653	0.8183	0.7992	0.2882
0	1	0	Std	0.1580	0.2231	0.2157	0.0857
			Mean	0.8230	0.8163	0.7981	0.2500
0	1	1	Std	0.2100	0.1933	0.1983	0.0697
			Mean	0.8581	0.8133	0.7883	0.1856
1	0	0	Std	0.1546	0.2206	0.2067	0.0582
			Mean	0.7985	0.7954	0.7635	1.0199
1	0	1	Std	0.2015	0.1632	0.1779	0.2964
			Mean	0.8589	0.8140	0.7923	0.2269
1	1	0	Std	0.1558	0.2213	0.2107	0.0758
			Mean	0.8243	0.8158	0.8093	0.2224
1	1	1	Std	0.2144	0.1725	0.1961	0.0654

Variações

				SVC			
Lower	Stop Words	TF IDF		Acuracy	Precision	Recall	Train Time
0	0	0	Mean	0.8270	0.7813	0.6634	2.7348
			Std	0.1354	0.2239	0.1811	0.6833
0	0	1	Mean	0.7501	0.3750	0.5000	4.0000
			Std	0.1490	0.0745	0.0000	0.0000
0	1	0	Mean	0.8181	0.7027	0.6350	3.0566
			Std	0.1307	0.2382	0.1837	0.6328
0	1	1	Mean	0.7501	0.3750	0.5000	3.0000
			Std	0.1490	0.0745	0.0000	0.0000
1	0	0	Mean	0.8046	0.7873	0.6648	6.1887
			Std	0.1740	0.2376	0.2148	4.1193
1	0	1	Mean	0.7501	0.3750	0.5000	14.0000
			Std	0.1490	0.0745	0.0000	2.0000
1	1	0	Mean	0.8057	0.7380	0.6426	1.9738
			Std	0.1531	0.2505	0.2055	0.4315
1	1	1	Mean	0.7501	0.3750	0.5000	2.0000
			Std	0.1490	0.0745	0.0000	0.0000

Variações

				MLP			
Lower	Stop Words	TF IDF		Acuracy	Precision	Recall	Train Time
			Mean	0.8903	0.8665	0.8474	11.6260
0	0	0	Std	0.1536	0.1699	0.1606	3.0704
			Mean	0.8696	0.8304	0.8087	9.8523
0	0	1	Std	0.1760	0.1982	0.2362	1.6026
			Mean	0.8852	0.8599	0.8567	9.0294
0	1	0	Std	0.1575	0.1808	0.1676	2.8338
			Mean	0.8762	0.8218	0.8638	5.0464
0	1	1	Std	0.1689	0.2038	0.1915	0.7650
			Mean	0.8879	0.8227	0.8338	24.8311
1	0	0	Std	0.1517	0.2118	0.1945	3.3435
			Mean	0.8552	0.7737	0.8226	24.0375
1	0	1	Std	0.1487	0.2027	0.1892	7.1442
			Mean	0.9028	0.8746	0.8876	5.1970
1	1	0	Std	0.1513	0.1615	0.1569	0.8987
			Mean	0.8943	0.8434	0.8878	4.9544
1	1	1	Std	0.1525	0.1760	0.1634	0.8416

Variações

				LogisticRegression			
Lower	Stop Words	TF IDF		Acuracy	Precision	Recall	Train Time
			Mean	0.8875	0.8882	0.8129	0.2511
0	0	0	Std	0.1495	0.1546	0.1711	0.0691
			Mean	0.8590	0.8390	0.7843	0.2204
0	0	1	Std	0.1697	0.2013	0.2006	0.0241
			Mean	0.8944	0.8955	0.8381	0.1706
0	1	0	Std	0.1572	0.1590	0.1776	0.0448
			Mean	0.8767	0.8534	0.8314	0.1513
0	1	1	Std	0.1609	0.1873	0.1844	0.0168
			Mean	0.8878	0.8802	0.8396	0.9416
1	0	0	Std	0.1538	0.1519	0.1713	0.1204
			Mean	0.8590	0.8464	0.7930	0.1818
1	0	1	Std	0.1523	0.1809	0.1717	0.0286
			Mean	0.8988	0.8949	0.8654	0.1373
1	1	0	Std	0.1564	0.1542	0.1719	0.0129
			Mean	0.8728	0.8527	0.8307	0.1300
1	1	1	Std	0.1562	0.1822	0.1753	0.0115

Variações

				RandomForest			
Lower	Stop Words	TF IDF		Acuracy	Precision	Recall	Train Time
			Mean	0.8907	0.8977	0.8370	1.0673
0	0	0	Std	0.1607	0.1551	0.1912	0.1536
			Mean	0.8880	0.8819	0.8042	0.9616
0	0	1	Std	0.1511	0.1577	0.1978	0.1084
			Mean	0.8878	0.8463	0.8397	0.6467
0	1	0	Std	0.1691	0.2340	0.2039	0.0727
			Mean	0.9000	0.8994	0.8548	0.7094
0	1	1	Std	0.1510	0.1506	0.1678	0.0914
			Mean	0.8874	0.8487	0.8456	2.8228
1	0	0	Std	0.1659	0.2296	0.2086	0.3105
			Mean	0.8910	0.9007	0.8259	0.8009
1	0	1	Std	0.1497	0.1500	0.1860	0.2264
			Mean	0.8918	0.8492	0.8551	0.6037
1	1	0	Std	0.1695	0.2337	0.2066	0.0791
			Mean	0.8863	0.8438	0.8318	0.6363
1	1	1	Std	0.1652	0.2287	0.2049	0.0864

Melhor Modelo

				MLP			
Lower	Stop Words	TF IDF		Accuracy	Precision	Recall	Train Time
1	1	0	Mean	0.9028	0.8746	0.8876	5.1970
			Std	0.1513	0.1615	0.1569	0.8987

				Logistic Regression			
Lower	Stop Words	TF IDF		Accuracy	Precision	Recall	Train Time
1	1	0	Mean	0.8988	0.8949	0.8654	0.1373
			Std	0.1564	0.1542	0.1719	0.0129

Tunning

- Penalty
- C
- Solver

GridSearchCV Acuracy	GridSearchCV Precision	GridSearchCV Recall	GridSearchCV Train Time
0.532608695652174	0.6022104108664473	0.6840092235891913	16.709161520004272
0.9962264150943396	0.98	0.9979253112033195	14.215339183807373
0.9076923076923077	0.9152661064425771	0.8504728132387707	15.733324766159058
0.7183098591549296	0.6123188405797102	0.5151960784313725	11.940999984741211
1.0	1.0	1.0	14.29494333267212
0.989100817438692	0.969005956813105	0.9850789372352715	11.966851234436035
0.9944444444444445	0.9946808510638299	0.9942528735632183	14.028168439865112
0.9843260188087775	0.9514541715628673	0.9341101694915255	12.386742353439331
0.9653465346534653	0.9686238902340598	0.9618673560703987	8.241345405578613
1.0	1.0	1.0	14.570685148239136

	GridSearchCV Acuracy	GridSearchCV Precision	GridSearchCV Recall	GridSearchCV Train Time
mean	0.9088055092939131	0.8993560227562597	0.8922912762823068	13.408756136894226
std	0.15803331512515031	0.15604523971403092	0.16575994734160307	2.3908620783826193

Wrapper

Exemplo de informações relevantes

Smartphone Xiaomi Mi 8 Lite 64GB 4GB RAM Preto

por Xiaomi

994 avaliações de clientes | 1000+perguntas respondidas

Preço: R$1.019,00

Em até 10x R$ 101,90 sem juros Calculadora de prestações

Cor: **Preto**

- Conectividade sem fio: GSM, 3G, 4G, GPS, WiFi e Bluetooth
- Armazenamento externo até 256 GB
- Memória Interna: 64 GB
- Câmera Frontal: 24.0 MP
- Tela 6.26"

Nome do produto	Celular Smartphone Galaxy Note10+ 6,8" 256GB Samsung - Preto
SKU	309904-2
Marca	Samsung
Código de barras	7892509109901
Garantia de fabrica	12 meses

Informações do produto

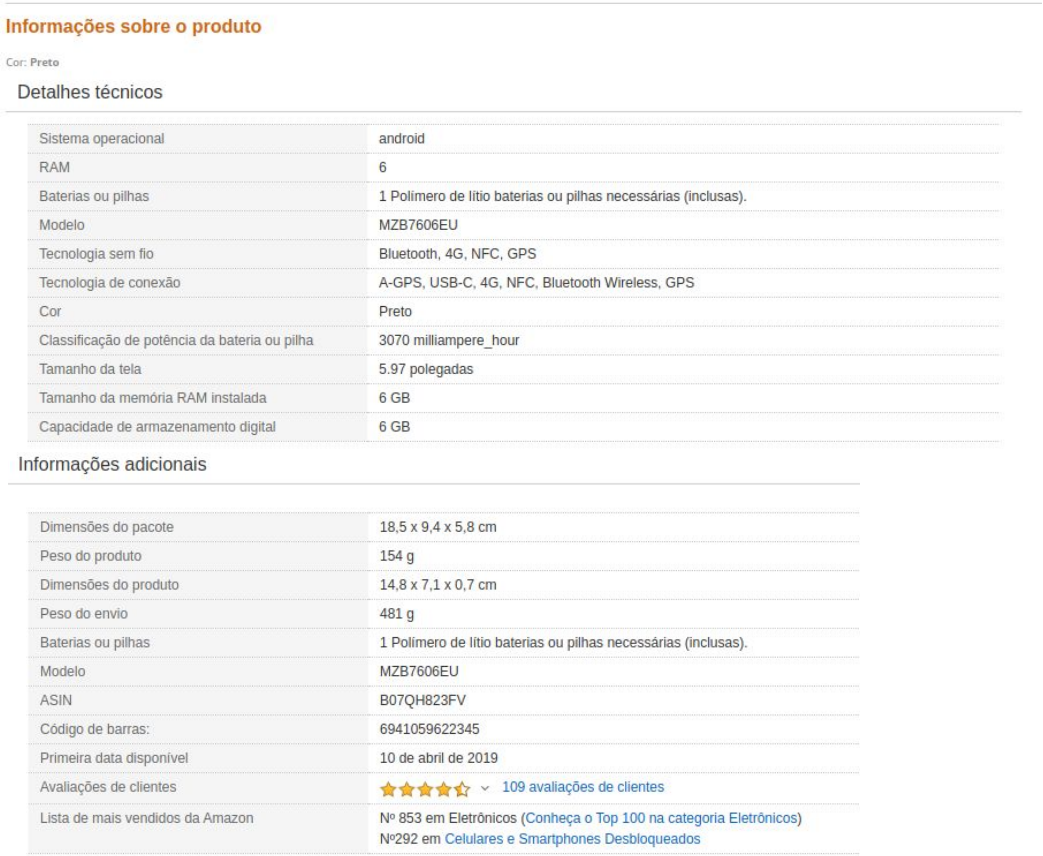
Smartphone Samsung Galaxy A10 32GB Preto 4G - 2GB RAM 6,2" Câm. 13MP + Câm. Selfie 5MP

Tela 6.2" Polegadas
Câmera frontal de 5 MP Câmera traseira 13 MP
Processador Octa-Core 1.6GHz
Memória Ram de 2 GB
Memória 32 GB, expansível até 512 GB

Itens Inclusos:
Aparelho celular
carregador
cabo USB
fone de ouvido
Extrator de Chip
Manual do usuário

Campos buscados

1. Preço
2. Sistema operacional
3. RAM
4. Capacidade de armazenamento digital
5. Modelo
6. Cor
7. Tamanho da tela
8. Avaliações de clientes



Informações sobre o produto

Cor: **Preto**

Detalhes técnicos

Sistema operacional	android
RAM	6
Baterias ou pilhas	1 Polímero de lítio baterias ou pilhas necessárias (inclusas).
Modelo	MZB7606EU
Tecnologia sem fio	Bluetooth, 4G, NFC, GPS
Tecnologia de conexão	A-GPS, USB-C, 4G, NFC, Bluetooth Wireless, GPS
Cor	Preto
Classificação de potência da bateria ou pilha	3070 milliampere_hour
Tamanho da tela	5.97 polegadas
Tamanho da memória RAM instalada	6 GB
Capacidade de armazenamento digital	6 GB

Informações adicionais

Dimensões do pacote	18,5 x 9,4 x 5,8 cm
Peso do produto	154 g
Dimensões do produto	14,8 x 7,1 x 0,7 cm
Peso do envio	481 g
Baterias ou pilhas	1 Polímero de lítio baterias ou pilhas necessárias (inclusas).
Modelo	MZB7606EU
ASIN	B07QH823FV
Código de barras:	6941059622345
Primeira data disponível	10 de abril de 2019
Avaliações de clientes	109 avaliações de clientes
Lista de mais vendidos da Amazon	Nº 853 em Eletrônicos (Conheça o Top 100 na categoria Eletrônicos) Nº292 em Celulares e Smartphones Desbloqueados

Extratores Individuais

- Biblioteca BeautifulSoup em conjunto com regex
- XPath
- Distância de Levenshtein
- Dificuldades:
  - Nomenclatura varia de fabricante
  - Poucas informações
  - Encontrar o valor de venda correto
  - Informações importantes presente em imagens

'Preço' : 'R$1.998,00',
'Sistema operacional': 'android',
'RAM' : '6',
'Capacidade de armazenamento digital': '6 GB',
'Modelo': 'MZB7606EU',
'Cor' : 'Preto',
'Tamanho da tela': '5.97 polegadas',
'Avaliações de clientes': '4.7 de 5 estrelas 109 avaliações de clientes '}


```
XPATH_MODEL = r"""//*[@id="pag-detalhes"]/div/div[6]/div[2]/p/text()"""
```

```
price_roi = s.find_all("div", attrs={"class":"preco_normal"},text=True)
if len(price_roi):
    price = re.findall("R\$ \d{1,5}\,\d{1,3}", str(price_roi[0]))
    if len(price):
        fields_kabum["Preço"] = price[0]
```