Extração de Dados na WEB: Remédios

Iago Raphael (irvm) e João Lucas (jlgm)

Domínio

- Farmácias Online
- Informações sobre remédios
  - Preços, marca
- Páginas do domínio escolhidas manualmente
  - Por relevância nas buscas no Google

Páginas

- Onofre
- Ultra Farma
- Farma Delivery
- Farma22
- Pague Menos
- Farmagora
- Drogaria São Paulo
- Medicamentos Brasil
- Farmacia Cristo Rei
- Sare Drogarias

Crawling

- Primeiro crawler faz uma Busca em Largura a partir do root
  - Ex. a partir de www.onofre.com.br
- Coloca todos os links acessíveis na fila
- Páginas visitadas muitas vezes não tinham instância de remédios
- Lista de URLs visitadas: https://raw.githubusercontent.com/jlgm/DrugCrawling/master/urls?token=ADW4eH4jG_QzGncZJjAwL047vS5JL9S6ks5ZJwpawA%3D%3D

Crawling (BFS)

```
1   https://www.onofre.com.br
2   https://www.onofre.com.br/medicamentos/antidiarreico/44/03
3   https://www.onofre.com.br/medicamentos/antiflatulento/54/03
4   https://www.onofre.com.br/medicamentos/antiacido/95/03
5   https://www.onofre.com.br/medicamentos/laxante/248/03
6   https://www.onofre.com.br/medicamento/55/01
7   https://www.onofre.com.br/medicamentos/hipoglicemiante/234/03
8   https://www.onofre.com.br/medicamentos/antigotoso/57/03
9   https://www.onofre.com.br/medicamentos/tireoidiano/342/03
10  https://www.onofre.com.br/medicamentos/antibiotico/32/03
11  https://www.onofre.com.br/medicamentos/anti-infeccioso/67/03
12  https://www.onofre.com.br/medicamentos/antidiuretico/46/03
```

```
122 http://www.ultrafarma.com.br/
123 http://www.ultrafarma.com.br/minha_cesta.html
124 http://www.ultrafarma.com.br/cliente/duvidas_sugestoes.html
125 http://www.ultrafarma.com.br/atendimento/televendas.html
126 http://www.ultrafarma.com.br/atendimento/chat_offline.html
127 http://www.ultrafarma.com.br/atendimento.html
128 http://www.ultrafarma.com.br/meus_pedidos.html
129 http://www.ultrafarma.com.br/institucional/nossas_lojas.html
130 http://www.ultrafarma.com.br/institucional/ultraempresas.html
131 http://www.ultrafarma.com.br/meu_perfil.html
132 http://www.ultrafarma.com.br/categoria-372/ordem-1/Medicamentos.html
```

Crawling (BFS, bons exemplos)

```
110  https://www.onofre.com.br/dorflex-com-36-comprimidos/63596/05
111  https://www.onofre.com.br/adoless-60mcg15mcg-c-28-comprimidos/791/05
112  https://www.onofre.com.br/alexa-c-24-comprimidos-revestidos/1137/05
113  https://www.onofre.com.br/triquilar-c-21-comprimidos/41566/05
114  https://www.onofre.com.br/cerazette-75mg-com-3-cartelas-de-28-comprimidos-cada/56225/05
115  https://www.onofre.com.br/magnesia-bisurada-com-40-comprimidos/27797/05
116  https://www.onofre.com.br/mylanta-plus-morango-240ml/29555/05
117  https://www.onofre.com.br/sal-de-fruta-eno-com-2-envelopes-de-5g/35483/05
118  https://www.onofre.com.br/engov-com-24-comprimidos/60776/05
119  https://www.onofre.com.br/nexium-40mg-com-28-comprimidos/29889/05
120  https://www.onofre.com.br/tecta-40mg-com-60-comprimidos/39755/05
```

Crawling

- Solução: heurística!
- Usar fila de prioridade (heap) ao invés de fila comum
- Usar como critério de prioridade o número de - (hífens) e _ (underlines) presentes na URL
- Lista de URLs visitadas: https://raw.githubusercontent.com/jlgm/DrugCrawling/master/urls-heuristic?token=ADW4eGdtHFJbMxQ_aG_SNCnmpK1dZWRgks5ZJwpKwA%3D%3D

BFS com heurística

```
1  https://www.onofre.com.br
2  https://www.onofre.com.br/insulina-tresiba-flextouch-com-1-sistema-de-aplicacao-de-3ml/60622/05
3  https://www.onofre.com.br/cerazette-75mg-com-3-cartelas-de-28-comprimidos-cada/56225/05
4  https://www.onofre.com.br/rinosoro-9mg-09-gotas-nasais-pediatrico-adulto-com-30ml/33996/05
5  https://www.onofre.com.br/sal-de-fruta-eno-com-2-envelopes-de-5g/35483/05
6  https://www.onofre.com.br/esomeprazol-magnesio-40mg-com-28-comprimidos-medley-genericos/54460/05
7  https://www.onofre.com.br/victoza-injetavel-6mg-com-2-sistemas-de-aplicacao/42174/05
8  https://www.onofre.com.br/forxiga-10mg-com-30-comprimidos-revestidos/59510/05
9  https://www.onofre.com.br/naridrin-12hs-gotas-nasais-com-30ml/29636/05
10 https://www.onofre.com.br/neosoro-gotas-nasais-adulto-com-30ml/29796/05
11 https://www.onofre.com.br/propilracil-100mg-c-30-comprimidos-biolab/32800/05
12 https://www.onofre.com.br/rosuvastatina-10mg-com-30-comprimidos-medley/59035/05
13 https://www.onofre.com.br/sal-de-fruta-eno-c2-env/35483/07
14 https://www.onofre.com.br/aerolin-5mgml-gotas-para-nebulizacao-com-10ml/817/05
15 https://www.onofre.com.br/aerolin-100mcg-spray-com-200-doses/819/05
16 https://www.onofre.com.br/aerolin-nebules-25mg-com-20-flaconetes/816/05
```

Crawling estatísticas

- Harvest Ratio da BFS sem heuristica:
  - 107 páginas relevantes de 1002 visitadas
  - HR = 0.106786427146
- Harvest Ratio com heurística:
  - 856 páginas relevantes de 1199 visitadas
  - HR = 0.713928273561

Crawling: melhorias

- Adicionar blacklist de palavras
  - ex. categoria, quem-somos
- Dar prioridade a URLs que possuem unidade de medida
  - Ex. 50ml, 30g

Classificação

- Modelo escolhido: Naïve Bayes
  - Features: bag of words
  - Vocabulário restrito às seguintes palavras: “adicionar”, “informações”, “detalhes”, “descrição”, “contraindicações”, “inicial', “home”, “ordenar”, “página”, “abrace”
  - Páginas foram convertidas para somente texto.

Classificação

- Modelo escolhido: Naïve Bayes
  - Um objeto treinado com o conjunto de páginas obtidas por busca em largura simples (1080 documentos), e outro com o conjunto de páginas obtidas com uso da heurística (2000 documentos).
  - Treinamento: todo o conjunto.
  - Validação cruzada: k-fold com 8 partições.

Classificação

- Resultados - conjunto 1 (sem heurística)
  - Accuracy: 0.9611
  - Precision: 0.7744
  - Recall: 0.8956
  - F-measure: 0.8293
  - Confusion matrix:
  - [103 12]
  - [30 935]

Classificação

- Resultados - conjunto 2 (com heurística)
  - Accuracy: 0.8275
  - Precision: 0.8492
  - Recall: 0.9410
  - F-measure: 0.8924
  - Confusion matrix:
  - [862 54]
  - [153 131]

Extração

- Classificador coloca páginas rotuladas como “possui instância” em pasta chamada “classificados”
- Extrator visita essa página e salva os atributos das instâncias válidas em um documento estruturado (formato JSON)
- Dados gerados: https://raw.githubusercontent.com/jlgm/DrugCrawling/master/extraction/data.js?token=ADW4eGWdhVvaZKSCDauuUs811Qv3hHEmks5ZJwopwA%3D%3D

Extração


```
[{'price': u'Por R$ 17,95 cada', 'produto': u' Complexo B 320 mg - Sidney Oliveira Leve 100 Pague 75 Comprimidos ', 'site':
'www.ultrafarma.com.br', 'farmacia': 'Ultra Farma', 'sumario': u'Ultrafarma. T\xe1 no cora\xe7\xe3o da gente! - COMPLEXO B 320 MG - SIDNEY
OLIVEIRA LEVE 100 PAGUE 75 COMPRIMIDOS '}, {'price': u'\r\n                                        Por:\r\n                                        \r\n
R$ 19,32\r\n\r\n                                        \n', 'produto': u'\nSUBSTIANE OLHOS ANTI-AGE 15ML\nC\xd3DIGO DO PRODUTO: 717791 | Marca: LA
ROCHE POSAY\n', 'site': 'www.farmagora.com.br', 'farmacia': 'Farmagora', 'sumario': u'Substiane Olhos Anti-Age 15Ml | Farmagora'}, {'price':
u'\nR$ 19,29', 'produto': u'Acetilciste\xedna 600mg Com 16 Envelopes - EMS - Gen\xe9rico', 'site': 'www.onofre.com.br', 'farmacia':
'Onofre', 'sumario': u'\r\n\t\r\n    Acetilciste\xedna 600mg Com 16 Envelopes - EMS - Gen\xe9rico | Onofre\r\n\r\n'}, {'price': u'\nR$
10,13', 'produto': u'Aturgyl Spray Nasal Com 15ml', 'site': 'www.onofre.com.br', 'farmacia': 'Onofre', 'sumario': u'\r\n\t\r\n    Aturgyl
Spray Nasal Com 15ml | Onofre\r\n\r\n'}, {'price': u'Por: R$ 15,90', 'produto': u'Kit Banho Fiorucci Flor de Cerejeira Sabonete L\xedquido
500ml + Lo\xe7\xe3o Hidratante 500ml', 'site': 'www.drogariasaopaulo.com.br', 'farmacia': 'Drogaria Sao Paulo', 'sumario': u'Kit Banho
Fiorucci Flor de Cerejeira Sabonete L\xedquido 500ml + Lo\xe7\xe3o Hidratante 500ml - Drogaria Sao Paulo'}, {'price': u'R$22,99', 'produto':
u'\nEnfagrow 800G \n', 'site': 'www.farmaciacristorei.com.br', 'farmacia': 'Farmacia Cristo Rei', 'sumario': u'Enfagrow 800G  - Leites e
F\xf3rmulas Especiais - Sa\xfade'}, {'price': u'Por R$ 5,99 cada', 'produto': u' Creme para Preven\xe7\xe3o de Assaduras Rahda Baby com 45
Gramas', 'site': 'www.ultrafarma.com.br', 'farmacia': 'Ultra Farma', 'sumario': u'Ultrafarma. T\xe1 no cora\xe7\xe3o da gente! - Creme para
preven\xe7\xe3o de assaduras rahda baby com 45 gramas'}, {'price': u'R$ 19,46', 'produto': u'Bico Mamadeira Naturform Lillo - farma22',
'site': 'www.farma22.com.br', 'farmacia': 'Farma 22', 'sumario': u'Bico Mamadeira Naturform Lillo - farma22'}, {'price': u'R$128,90',
'produto': u'\nEscova Modeladora Gama Italy Turbo Plus 2200 Rotating Styler 127V 1000/1100W - Branca\n', 'site': 'www.farmadelivery.com.br',
'farmacia': 'Farma Delivery', 'sumario': u'Escova Modeladora Gama Italy Turbo Plus 2200 Rotating Styler 127V 1000/1100W - Branca - Farma
Delivery'}, {'price': u'\nR$ 448,52\nno pagamento \xe0 vista.\n', 'produto': u'Champix - Parar de fumar - Pre\xe7o e Bula', 'site':
'www.saredrogarias.com.br', 'farmacia': 'Sare Drogarias Online', 'sumario': u'Champix - Parar de fumar - Pre\xe7o e Bula'}, {'price':
u'R$138,06', 'produto': u'\nAparelho de Press\xe3o Arterial Digital Auto Bra\xe7o Omron HEM-7113 (Detector de Arritmia e Hipertens\xe3o)\n',
'site': 'www.farmadelivery.com.br', 'farmacia': 'Farma Delivery', 'sumario': u'Aparelho de Press\xe3o Arterial Digital Auto Bra\xe7o Omron
HEM-7113 (Detector de Arritmia e Hipertens\xe3o) - Farma Delivery'}, {'price': u'R$306,89', 'produto': u'\nCadeira Para Autom\xf3vel Graco
Junior Maxi de 3 \xe0 12 Anos Estampa Orbit\n', 'site': 'www.farmadelivery.com.br', 'farmacia': 'Farma Delivery', 'sumario': u'Cadeira Para
Autom\xf3vel Graco Junior Maxi de 3 \xe0 12 Anos Estampa Orbit - Farma Delivery'}, {'price': u'Por R$ 11,99 cada', 'produto': u' Escova
Dental Dr Axell Profissional Leve 4 Pague 3 Gr\xe1tis 4 Protetores de Cerdas', 'site': 'www.ultrafarma.com.br', 'farmacia': 'Ultra Farma',
'sumario': u'Ultrafarma. T\xe1 no cora\xe7\xe3o da gente! - ESCOVA DENTAL DR AXELL PROFISSIONAL LEVE 4 PAGUE 3 GR\xc1TIS 4 PROTETORES DE
```

Extração - Acurácia

- Total extrações possíveis (N):
  - 963
- Total extrações realizadas (E):
  - 739
- Total extrações corretas (C):
  - 739
- Recall:
  - R = 0.767393561786
- Precision:
  - P = 1.0
- F-Measure:
  - F = 0.86839012926

Dúvidas?

Obrigado!