

Remédios

RAFAEL JOSÉ (RJS4)
ALEXANDRE FERREIRA (AFC2))

Visão geral do projeto

- Crawling
 - 2 estratégias
 - Busca em largura
 - Heurística
 - Estatística
 - Harvest ratio
- Classificador
 - Rotular exemplos positivos e negativos
 - criação de features
 - comparar estratégia

Extrator



Motivação

satisfazer a necessidade do usuário comparando quais técnicas de utilização darão um melhor desempenho (número de páginas relevante), para as necessidades do cliente





-
- Incompatibilidade do python3 com https
 - ambiente isolado
 - Dificuldade na coleta de tantas páginas na web(1000 paginas por site)
 - Dificuldade na instalação de algumas ferramentas (ex BeautifulSoup)
 - Como distinguir páginas relevantes de não relevantes
 - Como estruturar tantas páginas web em tags.

Domínio

- Farmácias Online
- Informações sobre remédios
 - Valor, tipo(marca).
- Páginas do domínio escolhidas manualmente (Por relevância nas buscas do google e padronização do site - estrutura física)

Páginas visitadas

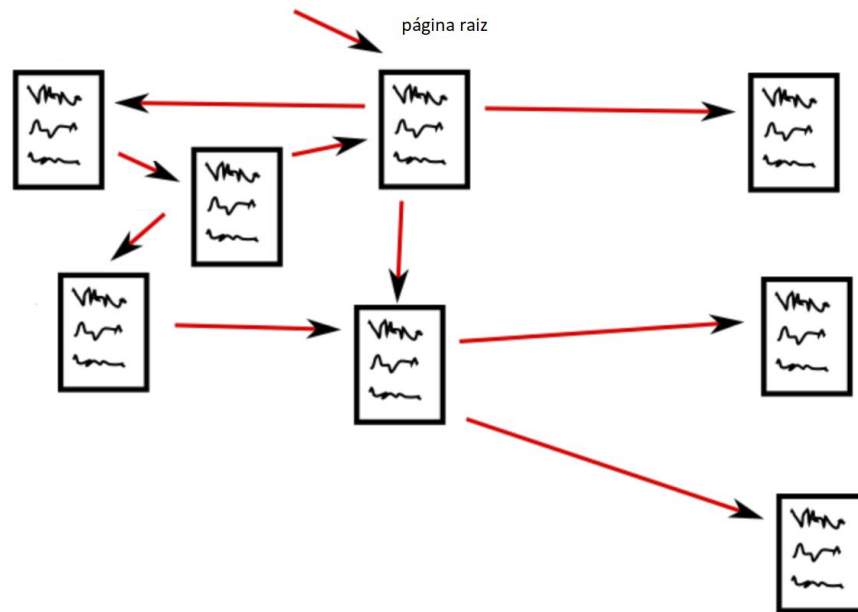


drogaria brasil



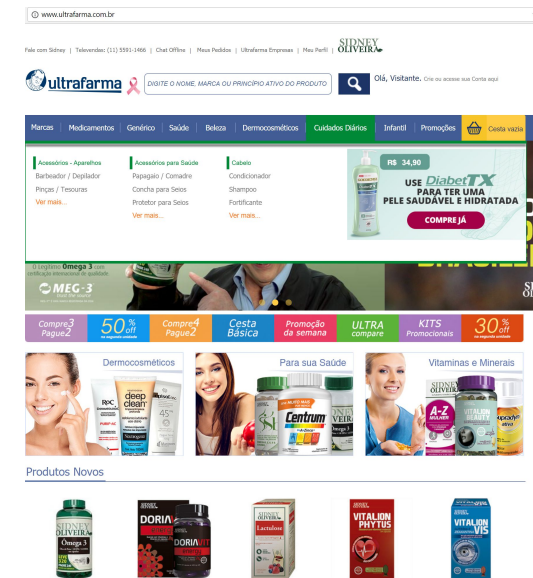
Crawling

O crawler deve fazer uma busca partindo da primeira página(página raiz)
abaixo temos alguns exemplos

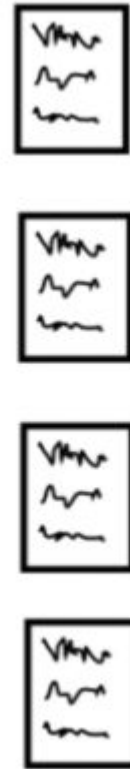
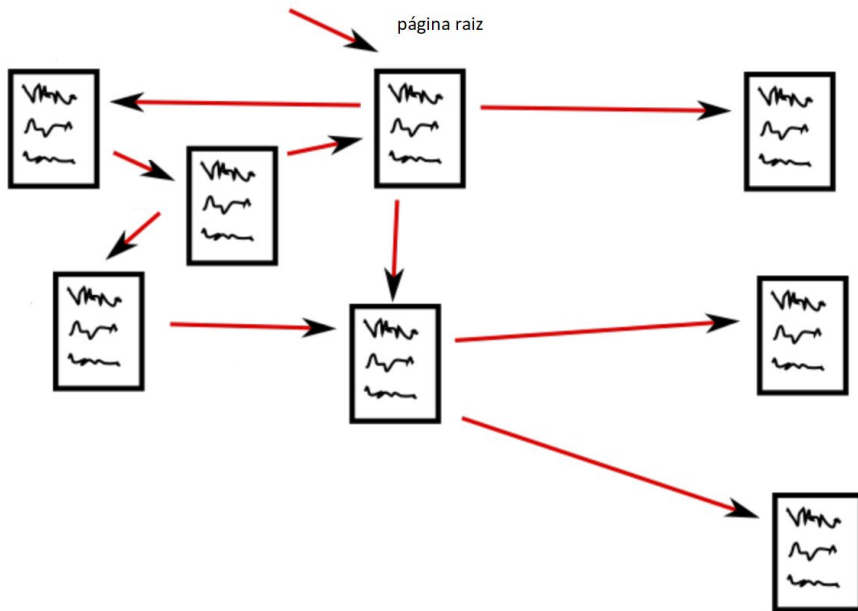


Crawling

Exemplos de páginas raízes



Crawling (busca em largura)



- Se inicia pela raiz
- Coloca todos os links acessíveis a partir da raiz em uma fila
- páginas visitadas muitas vezes não tinham instâncias de remédios
- link github para url visitadas

<https://pastebin.com/5dFsxAEA>

Crawling(B. em largura) bons exemplos!

Drogaria Onofre [BR] | <https://www.onofre.com.br/dorflex-com-36-comprimidos/63596/05>

OUTUBRO ROSA Prevenir é o melhor remédio. SAIBA MAIS

À VISITANTE, [FAÇA LOGIN](#) | [CRIAR CONTA](#) | [ESQUECI MINHA SENHA](#) TELEFONADAS: 4007-2526

Drogaria Onofre CVS Health A drogaria onofre agora faz parte da rede de farmácias CVS Health. [Clique aqui e saiba mais.](#)

Busca: Digite o nome do produto, marca ou princípio ativo... R\$ 0,00

MEDICAMENTOS SAÚDE BELEZA E BEM ESTAR MAMÃE E BEBÊ OFERTAS E LANÇAMENTOS COMO COMPRAR

ogaria Onofre > Medicamentos > Analgésico > Analgésico e Relaxante Muscular



Cód: 631948 MS: 1130001830132
Dorflex Com 36 Comprimidos

Sanofi Aventis
Princípio Ativo: Citrato de Orfenadrina-Dipirona Monodratada-Cafeína

Salvar no Facebook

Preço válido para compras feitas pela internet. Imagem meramente ilustrativa.

Insira o CRM do Médico: UF

NÃO USE ESTE MEDICAMENTO DURANTE A GRAVIDEZ E EM CRIANÇAS MENORES DE TRÊS MESES DE IDADE. SE PERSISTIREM OS SINTOMAS O MÉDICO DEVERÁ SER PROCURADO.

Selecione a Quantidade:

QUANTIDADE

R\$ 18,98

R\$ 12,90

+ 1 COMPRAR

DORFLEX COM 36 COMPRIMIDOS É UM MEDICAMENTO, SEU USO PODE TRAZER RISCOS. PROCURE O MÉDICO E O FARMACÊUTICO. LEIA A BULA. NÃO USE ESTE MEDICAMENTO DURANTE A GRAVIDEZ E EM CRIANÇAS MENORES DE TRÊS MESES DE IDADE. SE PERSISTIREM OS SINTOMAS O MÉDICO DEVERÁ SER PROCURADO.

Drogaria Onofre [BR] | <https://www.onofre.com.br/adoless-60mcg15mcg-c-28-comprimidos/791/05>

OUTUBRO ROSA Prevenir é o melhor remédio. SAIBA MAIS

OLÁ VISITANTE, [FAÇA LOGIN](#) | [CRIAR CONTA](#) | [ESQUECI MINHA SENHA](#) TELEFONADAS: 4007-2526

Drogaria Onofre CVS Health A drogaria onofre agora faz parte da rede de farmácias CVS Health. [Clique aqui e saiba mais.](#)

Busca: Digite o nome do produto, marca ou princípio ativo... R\$ 0,00

MEDICAMENTOS SAÚDE BELEZA E BEM ESTAR MAMÃE E BEBÊ OFERTAS E LANÇAMENTOS COMO COMPRAR

Drogaria Onofre > Medicamentos > Anticoncepcionais > Anticoncepcional



Cód: 168955 MS: 1039001400011
Adoless 60mcg/15mcg C/ 28 Comprimidos

Farmosquima
Princípio Ativo: Gestodeno+Etinilestradiol

Salvar no Facebook

Preço válido para compras feitas pela internet. Imagem meramente ilustrativa.

Insira o CRM do Médico: UF

VENDA SOB PRESCRIÇÃO MÉDICA.

Selecione a Quantidade:

QUANTIDADE

R\$ 35,90


R\$ 27,64

+ 1 COMPRAR

"10% de desconto na terceira unidade. Para participar insira 3 unidades no carrinho de compras. "Oferta válida na compra de medicamento de uso contínuo sujeito à prescrição médica."

Crawling (B. largura) maus exemplos

Drogaria Onofre [BR] | https://www.onofre.com.br/medicamentos/antiacido/95/03

Onofre Digite o nome do produto, marca ou princípio ativo...  R\$ 0,00

SEÇÕES

ANTIÁCIDO

AZIA E MÁ DIGESTÃO

POR FORMA FARMACÊUTICA:

- ☐ COMPRIMIDOS (2)
- ☐ SUSPENSÃO (13)
- ☐ PÓ (16)
- ☐ PASTILHAS (2)
- ☐ ENVELOPE (7)
- ☐ COMPRIMIDOS (3)
- ☐ COMPRIMIDOS MASTIGÁVEIS (2)
- ☐ LÍQUIDO (1)

POR SABOR:

- ☐ SACHÊ (1)







POR QUANTIDADE:

- ☐ 10 COMPRIMIDOS (3)
- ☐ 2 COMPRIMIDOS (2)
- ☐ 20 COMPRIMIDOS (1)
- ☐ 30 COMPRIMIDOS (3)
- ☐ 40 COMPRIMIDOS (1)
- ☐ 50 COMPRIMIDOS (1)
- ☐ 6 COMPRIMIDOS (1)
- ☐ 240 ML (3)
- ☐ 100 GRAMAS (2)

POR FABRICANTE/ MARCA:

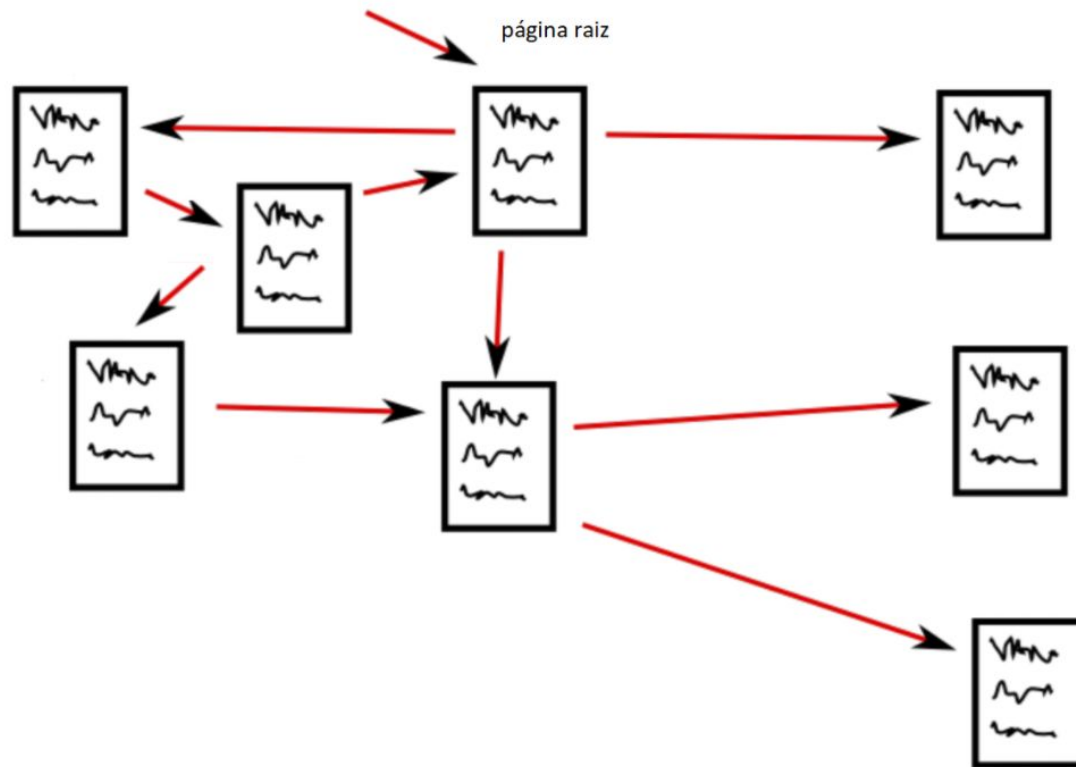
- ☐ A.D.V. (1)
- ☐ Cosmed Indústria S.A. (3)
- ☐ D.J.M. Dorsay (4)
- ☐ EMS (6)
- ☐ Eurofarma (1)
- ☐ Glaxosmithkline Brasil (17)

Exibir Todos os Fabricantes/ Marcas

 22% DE DESCONTO ENGOV COM 24 COMPRIMIDOS	 ESTOMAZIL PÓ LARANJA ENVELOPE DE 5G	 GASTROL PÓ LIMÃO 1 ENVELOPE 5G
NEPIRAMINA-HIDRÓXIDO DE ALUMÍNIO-ÁCIDO MS: 1781700930074 D.J.M. Dorsay R\$ 21,13 R\$ 16,48	ÁCIDO CÍTRICO-CARBONATO DE SÓDIO-BICARBONATO MS: 1781700390576 Cosmed Indústria S.A. R\$ 2,52	HIDRÓXIDO ALUMÍNIO-HIDRÓXIDO MS: 1558403960028 Neo Química R\$ 2,74
COMPRAR	COMPRAR	COMPRAR
	 22% DE DESCONTO	

Não tem uma relevância !!!

Crawling (heurística)



- Usar o número de “-” e “_” como prioridade
- URL visitadas
<https://pastebin.com/KwjgAH2m>

Crawling (heurística)

```
1  https://www.onofre.com.br
2  https://www.onofre.com.br/insulina-tresiba-flexitouch-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
```

A parte da estatística



Busca em largura

páginas visitadas	páginas relevantes	HR
1002	107	0.1067

Heurística

páginas visitadas	páginas relevantes	HR
1199	856	0.7139

Fazer um up grade

Dar prioridade a URLs que possuem unidades de medidas. por exemplo 50 ML

<https://www.onofre.com.br>

<https://www.onofre.com.br/insulina-tresiba-flexitouch-com-1-sistema-de-aplicacao-de-3ml/60622/05>

<https://www.onofre.com.br/cerazette-75mg-com-3-cartelas-de-28-comprimidos-cada/56225/05>

<https://www.onofre.com.br/rinosoro-9mg-09-gotas-nasais-pediatrico-adulto-com-30ml/33996/05>

<https://www.onofre.com.br/sal-de-fruta-eno-com-2-envelopes-de-5g/35483/05>

UPGRADE

Classificação

- Modelo escolhido: Naive 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 de heurística (2000 documentos).
 - Treinamento: todo o conjunto.
 - Validação cruzada: K-fold com 8 partições.

Classificação

- FEATURES
- VOCABULARY = ['adicionar', 'informacoes', 'informações',
- 'detalhes', 'descricao', 'descrição',
- 'contraindicações', 'contraindicações',
- 'inicial', 'home', 'ordenar', 'pagina', 'página',
- 'abrace']

Classificação

- Resultados - conjunto 1(sem heurística)
 - Documentos classificados: 1080
 - Accuracy: 0.961111111111
 - Precision: 0.895652173913
 - Recall: 0.895652173913
 - F1-Score: 0.818840705431
 - Confusion matrix:
 - $\begin{bmatrix} 103 & 12 \\ 30 & 935 \end{bmatrix}$

Classificação

- Resultados - conjunto 2 (com heurística)
 - Documentos classificados: 1200
 - Accuracy: 0.826666666667
 - Precision: 0.941048034934
 - Recall: 0.941048034934
 - F1-Score: 0.89226917669
 - Confusion matrix:
 - $\begin{bmatrix} 862 & 54 \\ 154 & 130 \end{bmatrix}$

Extração

O extrator visitará essas páginas e extrairá os atributos dessa páginas para documentos no formatos json

exemplos :

```
[{'price': '\nR$ 9,50', 'farmacia': 'Onofre', 'sumario': '\n\t\n    Acido Mefenâmico 500mg com 24 Comprimidos - Medley Generico | Onofre\n\n', 'produto': 'Acido Mefenâmico 500mg com 24 Comprimidos - Medley Generico', 'site': 'www.onofre.com.br'}, {'price': 'R$63,90', 'farmacia': 'Farma Delivery', 'sumario': 'Escova Dental Adulto Curaprox CS 5460 Ultra Soft Sensitive Trio Colorida c/ 3 Unidades - Farma Delivery', 'produto': '\nEscova Dental Adulto Curaprox CS 5460 Ultra Soft Sensitive Trio Colorida c/ 3 Unidades\n', 'site': 'www.farmadelivery.com.br'}, {'price': 'R$12,04', 'farmacia': 'Farma Delivery', 'sumario': 'Ranitidina 150mg c/ 20 Comprimidos Genérico EMS - Farma Delivery', 'produto': '\nRanitidina 150mg c/ 20 Comprimidos Genérico EMS\n', 'site': 'www.farmadelivery.com.br'}, {'price': 'R$1.134,00', 'farmacia': 'Farma Delivery', 'sumario': 'Accutrend Plus Roche Aparelho Monitor p/ Determinação de Glicose +Colesterol +Triglicérides +Lactato - Farma Delivery', 'produto': '\nAccutrend Plus Roche Aparelho Monitor p/ Determinação de Glicose +Colesterol +Triglicérides +Lactato\n', 'site': 'www.farmadelivery.com.br'}, {'price': 'R $68,00', 'farmacia': 'Farma Delivery', 'sumario': 'Mnوتاuro Focus & Energy Pó Para Preparo de Composto Líquido Sabor Laranja Pote 300g - Farma Delivery', 'produto': '\nMnوتاuro Focus & Energy Pó Para Preparo de Composto Líquido Sabor Laranja Pote 300g\n', 'site': 'www.farmadelivery.com.br'}, {'price': 'R$ 10,63', 'farmacia': 'Farma 22', 'sumario': 'Absorvente Carefree Todo Dia Flexi Sem Perfume 40un - farma22', 'produto': 'Absorvente Carefree Todo Dia Flexi Sem Perfume 40un - farma22', 'site': 'www.farma22.com.br'}, {'price': 'Por: R$ 3,99', 'farmacia': 'Drogaria Sao Paulo', 'sumario': 'Absorvente Always Básico Malha Suave Com Abas com 8 Unidades - Drogaria Sao Paulo', 'produto': 'Absorvente Always Básico Malha Suave Com Abas com 8 Unidades', 'site': 'www.drogariasao paulo.com.br'}, {'price': '\nR$ 46,19', 'farmacia': 'Onofre', 'sumario': '\n\t\n    Amoxil 500mg Com 15 Cápsulas | Onofre\n\n', 'produto': 'Amoxil 500mg Com 15 Cápsulas', 'site': 'www.onofre.com.br'}, {'price': 'R$ 24,29', 'farmacia': 'Farma 22', 'sumario': 'Fralda Huggies Turma da Mônica Tripla Proteção P - farma22', 'produto': 'Fralda Huggies Turma da Mônica Tripla Proteção P - farma22', 'site': 'www.farma22.com.br'}, {'price': 'Por R$ 11,50 cada', 'farmacia': 'Ultra Farma', 'sumario': 'Ultra Farma. Tá no coração da gente! - Escova dental aleg kids fun macia leve 4 pague 3', 'produto': 'Escova Dental Aleg Kids Fun Macia Leve 4 Pague 3', 'site': 'www.ultrafarma.com.br'}, {'price': 'Por R$ 7,15 cada', 'farmacia': 'Ultra Farma', 'sumario': 'Ultra Farma. Tá no coração da gente! - Talco desodorante para os pés rahda canforado com 100 gramas', 'produto': 'Talco Desodorante para os Pés Rahda Canforado com 100 Gramas', 'site': 'www.ultrafarma.com.br'}, {'price': 'Por: R$ 7,30', 'farmacia': 'Drogaria Sao Paulo', 'sumario': 'Absorvente Intimus Gel Normal Cobertura Suave sem Abas 16 Unidades - Drogaria Sao Paulo', 'produto': 'Absorvente Intimus Gel Normal Cobertura Suave sem Abas 16 Unidades', 'site': 'www.drogariasao paulo.com.br'}, {'price': '\nR$ 33,79\n\nno pagamento à vista.\n', 'farmacia': 'Sare Drogarias Online', 'sumario': 'Bepantol - Solução tratar estrias - Preço Bepantol', 'produto': 'Bepantol - Solução tratar estrias - Preço Bepantol', 'site': 'www.saredrogarias.com.br'}, {'price': 'R$13,99', 'farmacia': 'Farma Delivery', 'sumario': 'Escova Interdental Reutilizável Orvital 0.8mm c/ 5 Unidades - Farma Delivery', 'produto': '\nEscova Interdental Reutilizável Orvital 0.8mm c/ 5 Unidades\n', 'site': 'www.farmadelivery.com.br'}, {'price': 'Por: R$ 250,77', 'farmacia': 'Drogaria Sao Paulo', 'sumario': 'Kit Roc C-Supérieur Serum 15g + Gel Creme Anti-Oxidante 15ml + Gel Creme Anti-Oxidante Olhos 15g - Drogaria Sao Paulo', 'produto': 'Kit Roc C-Supérieur Serum 15g + Gel Creme Anti-Oxidante 15ml + Gel Creme Anti-Oxidante Olhos 15g', 'site': 'www.drogariasao paulo.com.br'}, {'price': 'Por R$ 69,90 cada', 'farmacia': 'Ultra Farma', 'sumario': 'Ultra Farma. Tá no coração da gente! - Ômega 3-6-9 mulher - sidney oliveira kamila sikora leve 240 pague 180 cápsulas', 'produto': 'Ômega 3-6-9 Mulher - Sidney Oliveira Kamila Sikora Leve 240 Pague 180 Cápsulas', 'site': 'www.ultrafarma.com.br'}, {'price': '\nR$ 243,40\n\nno pagamento à vista.\n', 'farmacia': 'Sare Drogarias Online', 'sumario': 'DDAVP Spray nasal 2,5ml - DDAVP Preço e bula', 'produto': 'DDAVP Spray nasal 2,5ml - DDAVP Preço e bula', 'site': 'www.saredrogarias.com.br'}, {'price': 'Por: R$ 23,16', 'farmacia': 'Drogaria Sao Paulo', 'sumario': 'Fralda Descartável Huggies Turma Da Mônica Veste Fácil Jumbo Xg 16 Unidades - Drogaria Sao Paulo', 'produto': 'Fralda Descartável Huggies Turma Da Mônica Veste Fácil Jumbo Xg 16 Unidades', 'site': 'www.drogariasao paulo.com.br'}, {'price': '\n\n\nR$ 19,32\n\n', 'farmacia': 'Farmagora', 'sumario': 'Imecap Hair Queda Intensa Kit Shampoo + Loção + Cápsulas | Farmagora', 'produto': '\nIMECAP HAIR QUEDA INTENSA KIT SHAMPOO + LOÇÃO + CÁPSULAS\n\nCÓDIGO DO PRODUTO: 728243 | Marca: DIVCOM PHARMA\n', 'site': 'www.farmagora.com.br'}, {'price': '\n\n\nR$ 19,32\n\n', 'farmacia': 'Farmagora', 'sumario': 'Alcool Lbs 70% Solução 50ML | Farmagora', 'produto': '\nALCOOL LBS 70% SOLUÇÃO 50ML\n\nCÓDIGO DO PRODUTO: 702805 | Marca: LBS\n', 'site': 'www.farmagora.com.br'}, {'price': 'Por R$ 13,90 cada', 'farmacia': 'Ultra Farma', 'sumario': 'Ultra Farma. Tá no coração da gente! - Lactulose 667 mg/ml xarope sabor ameixa sidney oliveira com 120 ml', 'produto': 'Lactulose 667 mg/ml Xarope Sabor Ameixa Sidney Oliveira com 120 mL', 'site': 'www.ultrafarma.com.br'}, {'price': '\nR$ 17,83', 'farmacia': 'Onofre', 'sumario': '\n\t\n    Propilracil 100mg c/ 30 Comprimidos - Biolab | Onofre\n\n', 'produto': 'Propilracil 100mg c/ 30 Comprimidos - Biolab', 'site': 'www.onofre.com.br'}, {'price': 'Por R$ 11,99 cada', 'farmacia': 'Ultra Farma', 'sumario': 'Ultra Farma. Tá no coração da gente! - ÔMEGA 3 CONCENTRADO 33 EPA/22 DHA 1000 MG - SIDNEY OLIVEIRA 20 CÁPSULAS', 'produto': 'Ômega 3 Concentrado 33 Epa/22 Dha 1000 mg - Sidney Oliveira 20 Cápsulas', 'site': 'www.ultrafarma.com.br'}, {'price': 'Por R$ 69,90 cada', 'farmacia': 'Ultra Farma', 'sumario': 'Ultra Farma. Tá no coração da gente! - ÓLEO DE COCO 1000 MG - SIDNEY OLIVEIRA LEVE 240 PAGUE 180 CÁPSULAS', 'produto': 'Óleo de Coco 1000 mg - Sidney Oliveira Leve 240 Pague 180 Cápsulas', 'site': 'www.ultrafarma.com.br'}
```

Extração : Estatística



- 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

