Recuperando informação de Ecommerces

Focando em **lojas de moda** na categoria **blusas.**

Por Mateus Nunes, Ramom Pereira

Prepi

Motivação

Migrar clientes de concorrêntes para o Prepi.

Tarefa 1 - O Crawler

Prepi

Seleção dos domínios.

Prepi

Tarefa 1 - Crawler / Domínio

- **Mercado de ecommerce muito pulverizado**
- **Existem Big players que competem regionalmente no segmento de criação de ecommers moda com startup, esse foi o foco.**
- **Para atender demanda das cadeiras foi necessário focar em uma categoria, a escolhida foi a de “blusas”.**
- **Foco resumido: buscar sites que foram criados pelos nossos concorrêntes**

Prepi

Tarefa 1 - Crawler / Domínio

- **Foi previamente selecionado 10 sites, mas fomos melhorando a lista conforme buscávamos links e entediamos a “dimensão” do site.**
- **Finalizando com a lista:**

"https://www.damyller.com.br",

"https://www.boutiquedassi.com.br",

"https://www.cuticutibaby.com.br",

"https://www.bluk.com.br",

"https://www.lojavillevie.com.br",

"https://www.achados96.com.br",

"https://www.bellaseda.online",

"https://www.gregory.com.br",

"https://www.lojasexclusiva.com.br",

“https://www.dwz.com.br",

Desenvolvendo *Crawler*

Tarefa 1 - Crawler / Desenvolvendo o Crawler

- Foi estudo especialmente a biblioteca de BeautifulSoup para criação do crawler.
- Bibliotecas auxiliares: sys, time, pandas, requests, validators, csv.

Prepi

- Estrutura básica:

Criado uma classe de crawler contendo as principais funções do crawler (respeito a rôbos, get e setters, e o crawler em sí)

Desenvolvendo *Crawler*

Dores de cabeça

- Principalmente com a captura de links inválidos, o crawler retorna formatos que são não úteis para o projeto : “tel”, “mailto”, “javascript”.

➡ Resolução: trabalhar com validators e excluindo links com formatos não úteis.

- Principalmente com a captura de links inválidos, o crawler retorna formatos que são não úteis para o projeto : “tel”, “mailto”, “javascript”.

➡ Resolução: trabalhar com validators e excluindo links com formatos não úteis.

- Conexões demasiadas e oscilações da internet

➡ Resolução: Respeitar overload do site com sleeps, e usar internet mais estáveis.

- Permissões do robots não aceitas.

  ➡ Resolução: Respeitar os robots, não tem pra onde correr.

- Erros de https nas páginas.

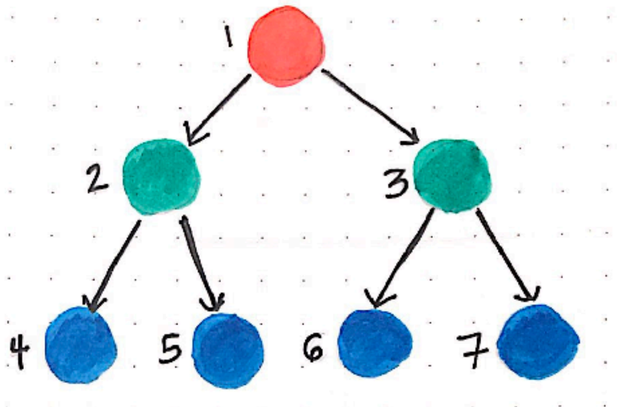
➡ Resolução: tratar excessões na execução do código

- Qualidade de alguns domínios

➡ Resolução: substituição de domínios a posteriori.

Desenvolvendo *Crawler*

Estratégias Utilizadas



BFS

👍🏽 Menor custo de execução

👍🏽 Rápida implementação

— — — — — —

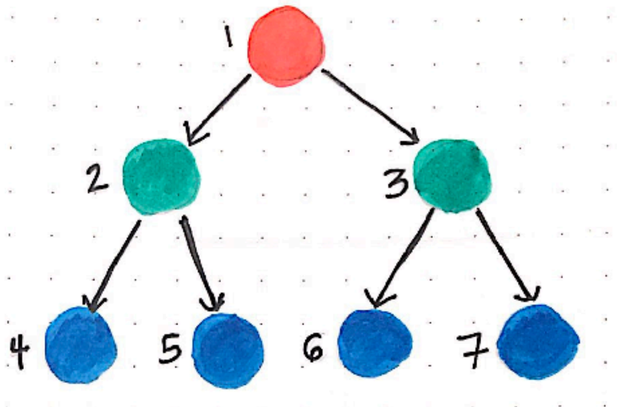
👎🏽 Coleta “aleatória”

👎🏽 Navegação em área completamente fora do interesse

👎🏽 Piores resultados no Haverst Ratio

Prepi

BFS - Resultados (overview)

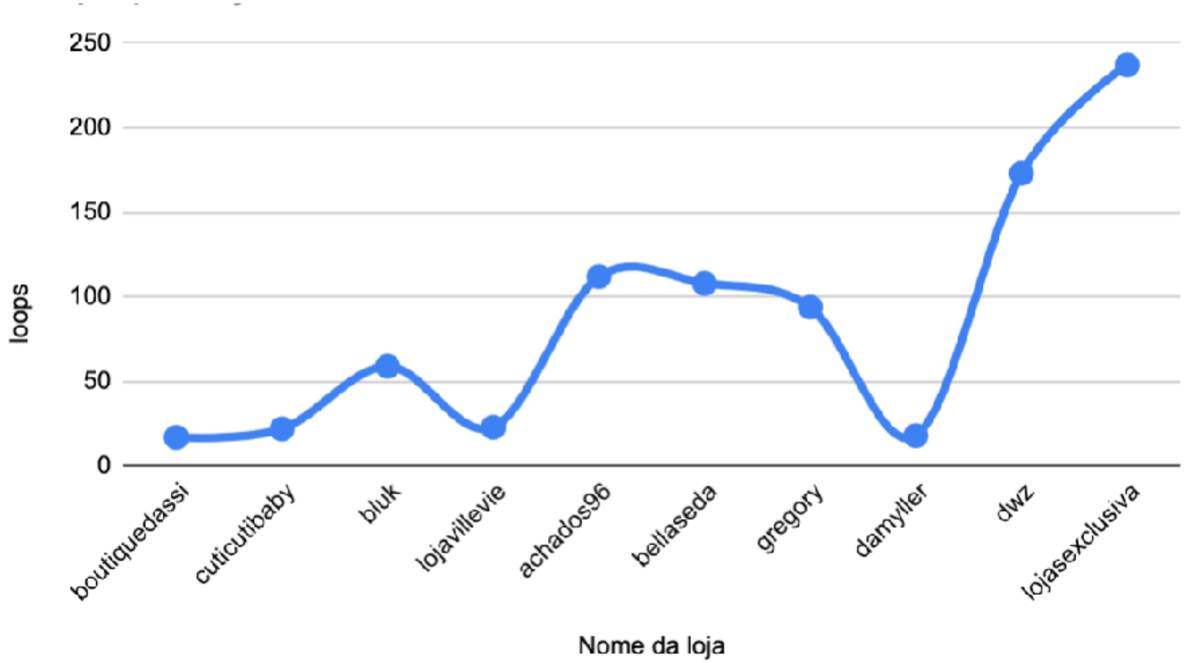


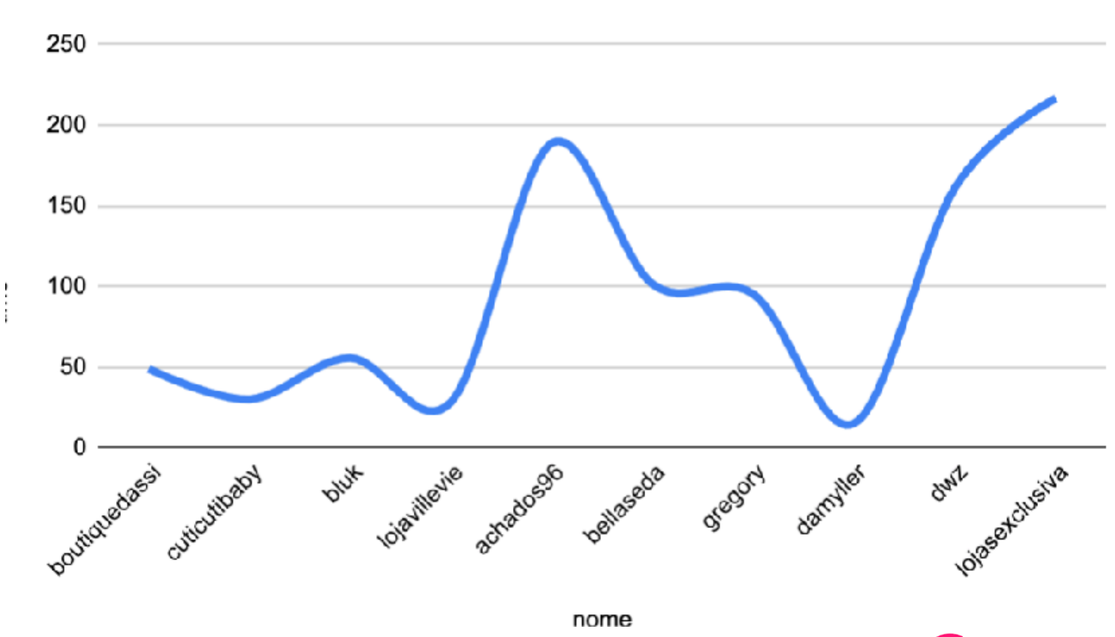
	Páginas Buscadas	Paginas Relevantes
Soma das lojas	10319	2330

Loops (de busca em profundidade) e Tempo

863 Total de loops

15,7 Minutos

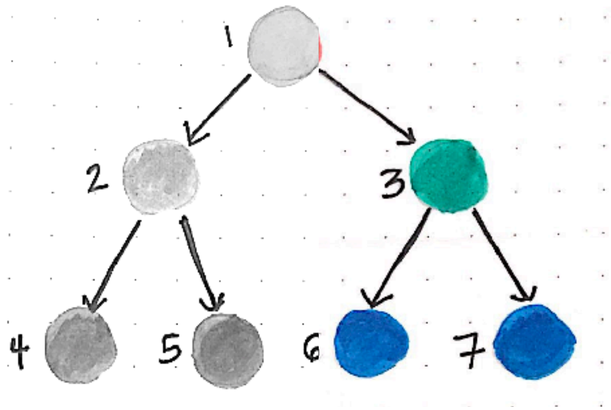




Prepi

Loops (de busca em profundidade) e Tempo

URL	Loops	Tempo
boutiquedassi	17	49.04772782
cuticutibaby	22	30.22799373
bluk	59	55.83135104
lojavillevie	23	28.46018314
achados96	112	188.9455893
bellaseda	108	101.4449992
gregory	94	95.37044597
damyller	18	15.260566
dwz	173	161.3772371
lojasexclusiva	237	216.5528548
media	**76.5**	**75.6008985**
total	**863**	**942.5189481**



Heurística

Utilizando checagem do subdomínio e keyworks positivas e negativas com um score para cada link

👍🏽 Busca mais inteligente e asertiva.

👍🏽 Busca com aprofundamento maior.

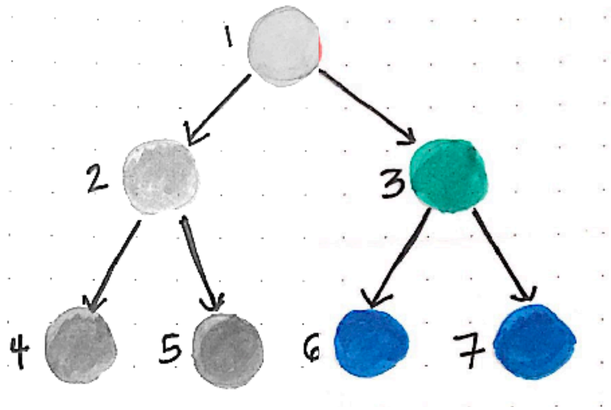
👍🏽 Melhores resultados no Haverst Ratio

— — — — — — —

👎🏽 Maior custo de execução (muito considerável)

👎🏽 Maior tempo para implementação

Prepi



Heurística

Utilizando checagem do subdomínio.

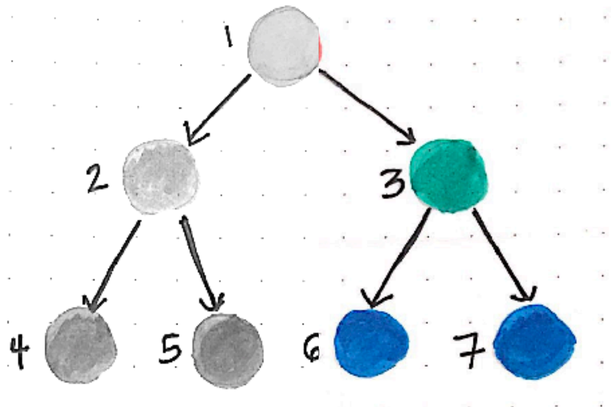
Keywords positivas e negativas

```
positives=["blusa","blusinhas","blus","t-shirt","tshirt"]
negatives=["calça","short","jeans","saia","vestido","biquine"]
```

Criando-se um score que se incrementa com a presença de palavra positiva na url e decrementa com presença de palavras negativas.

Prepi

Heurística - Resultados (overview)

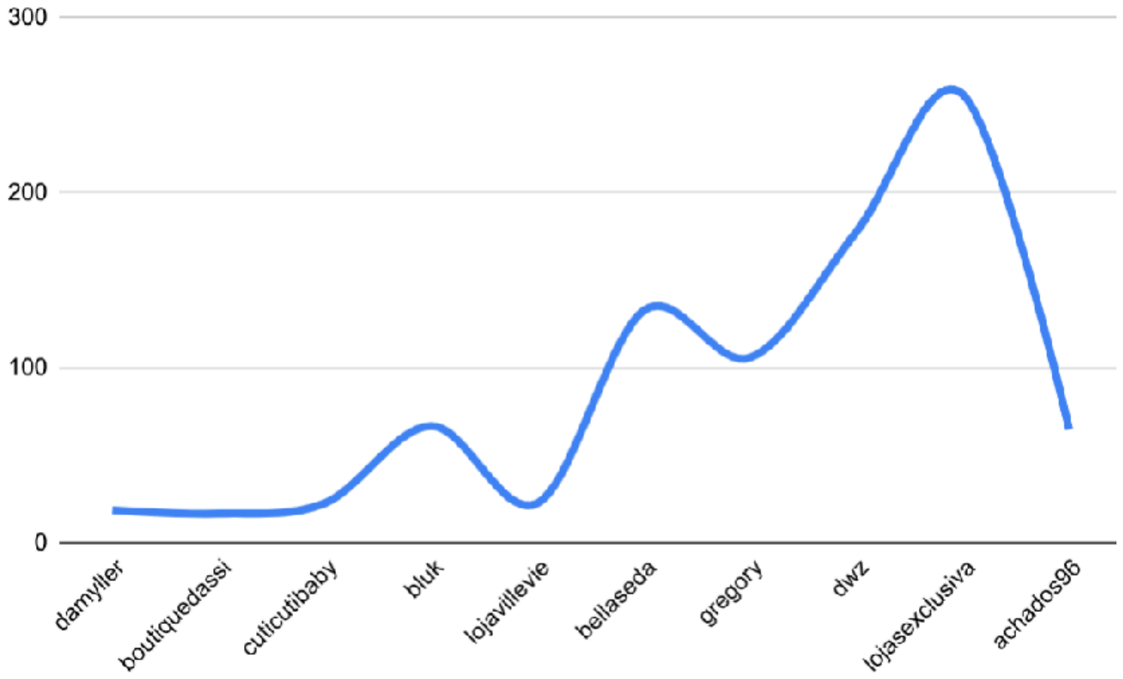


	Páginas Buscadas	Páginas Relevantes
Soma das lojas	9509	2871

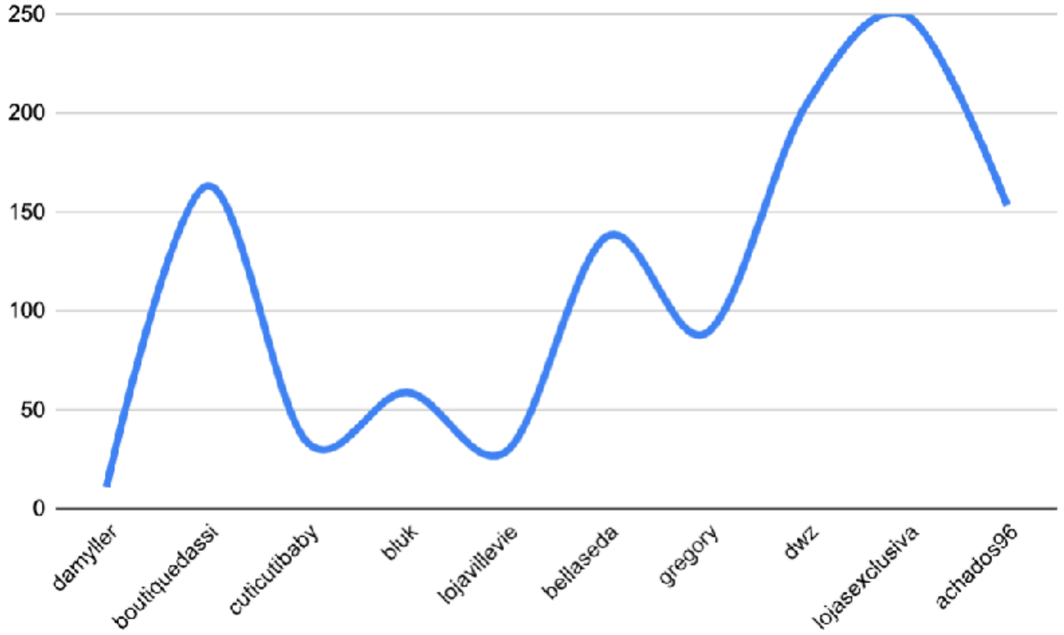
Prepi

Loops (de busca em profundidade) e Tempo

887 Total de loops



18,8 Minutos



Prepi

Loops (de busca em profundidade) e Tempo

URL	Loops	Tempo
damyller	19	11.32347512
boutiquedassi	17	163.3065042
cuticutibaby	23	34.13603091
bluk	67	59.08766317
lojavillevie	23	29.3817811
bellaseda	133	137.8630621
gregory	106	89.08289623
dwz	178	205.0935972
lojasexclusiva	256	249.5186079
achados96	65	153.5818949
media	**66**	**113.4729792**
total	**887**	**1132.375513**

Heurística vs BSF **Haverst Ratio**

	Páginas Buscadas	Páginas Relevantes	Haverst Ratio	Tempo	Ganho de Haverst	Perca de Tempo
Heuristica	9509	2871	30,1%	18,8 min	33,78%	19,75%
BSF	10319	2330	22,5%	15,7 min		

Resultados detalhados no proximo slide

Prepi

Heurística vs BSF **Dados Completos**

BSF

URL	Paginas Buscadas	Páginas Relevantes	HR
achados96	1012	177	0.1749011858
bellaseda	1001	293	0.2927072927
bluk	1043	189	0.1812080537
boutiquedassi	1070	276	0.2579439252
cuticutibaby	1001	223	0.2227772228
damyller	1016	175	0.1722440945
dwz	1149	238	0.2071366406
gregory	1004	290	0.2888446215
lojasexclusiva	1000	230	0.23
lojavillevie	1023	239	0.2336265885
total	10319	2330	0.2257970734

Heurística

URL	Paginas Buscadas	Páginas Relevantes	HR
achados96	332	97	0.2921686747
bellaseda	1001	342	0.3416583417
bluk	1020	316	0.3098039216
boutiquedassi	1058	294	0.2778827977
cuticutibaby	1027	263	0.2560856865
damyller	1034	203	0.1963249516
dwz	1015	342	0.3369458128
gregory	1000	390	0.39
lojasexclusiva	1003	301	0.3000997009
lojavillevie	1019	323	0.3169774289
total	9509	2871	0.3019244926

* HR= Harvest Ratio

Prepi

Pontos de melhorias

Prepi

Tarefa 1 - Crawler / Melhorias

- SOON Aprimoramento de heurísticas
- SOON Utilização de threads
- SOON Utilização de mais 1 estratégia.

Prepi

Tarefa 2 - O Classificador

Prepi

Extração e processando conteúdo

Prepi

Coleta

- **Coleta manual de 10 positivos (contendo blusas) em média.**
- **Coleta manual de 10 negativos em média.**

Baixar paginas e Processando

- **Dentro dos links do domínio, foi utilizado curl para baixar a página no links**
- **Processamento de conteúdos eliminando stop words e regex.**

Criação de database

- **Com os dados processados foi criado um database e dividiu-se ele para treinar e testar o classificador.**

Desenvolvendo o classificador

Seleção e execução de classificadores

- **Foram escolhidos 8 classificadores que são:**

  Naive Bases

  Logistic Regresson

  Random Forest

  KNN

  SVM

  Ada Boost

  MLP

  Decision Tree

Estratégias e Resultados

- **Foram utilizadas as estratégias de bag of words e TF-IDF**

Bag of Words

	Model	Time	Accuracy	Precision	Recall	F1-Measure
0	Naive Bayes	0.018991	0.753086	0.711111	0.820513	0.761905
1	Linear Regression	0.112701	0.753086	NaN	NaN	NaN
2	Random Forest	1.729755	0.814815	0.800000	0.820513	0.810127
3	KNN	0.006499	0.506173	0.487179	0.487179	0.487179
4	SVM	0.164042	0.481481	0.481481	1.000000	0.650000
5	Ada Boost	8.992370	0.703704	0.653061	0.820513	0.727273
6	MLP	2.615867	0.777778	0.744186	0.820513	0.780488
7	Decision Tree	0.027460	0.802469	0.767442	0.846154	0.804878

```
Testando Bag of Words
              precision    recall  f1-score   support

           0       0.81      0.69      0.74        42
           1       0.71      0.82      0.76        39

    accuracy                           0.75        81
   macro avg       0.76      0.76      0.75        81
weighted avg       0.76      0.75      0.75        81
```

Prepi

Estratégias e Resultados

- **Foram utilizadas as estratégias de bag of words e TF-IDF**

TF-IDF

	Model	Time	Accuracy	Precision	Recall	F1-Measure
0	Naive Bayes	0.011468	0.777778	0.744186	0.820513	0.780488
1	Linear Regression	0.016986	0.604938	NaN	NaN	NaN
2	Random Forest	1.995227	0.617284	0.590909	0.666667	0.626506
3	KNN	0.011142	0.506173	0.487805	0.512821	0.500000
4	SVM	0.180463	0.481481	0.481481	1.000000	0.650000
5	Ada Boost	15.612709	0.703704	0.682927	0.717949	0.700000
6	MLP	6.759238	0.777778	0.744186	0.820513	0.780488
7	Decision Tree	0.077739	0.740741	0.714286	0.769231	0.740741

```
Testando TF-IDF
              precision    recall  f1-score   support

           0       0.67      0.48      0.56        42
           1       0.57      0.74      0.64        39

    accuracy                           0.60        81
   macro avg       0.62      0.61      0.60        81
weighted avg       0.62      0.60      0.60        81
```

Prepi

Escolhendo os melhores

- Dentre as estratégias o Bag of Words se destacou
- Dentre todas métricas foi escolhido para o classificador o Decision Tree e Random Forest.

  Dentre esses Decision Tree foi escolhido pela o tempo de execução.

Desenvolvendo *Classificação*

Dores de cabeça

- Tempo para classificação
- Links removidos após a escolha.
- Overfit no classificador.

Obrigado!

Duvidas, comentários?

Por Mateus Nunes, Ramom Pereira