Recuperação de Informação - Projeto 1

Smartphone Web Crawler

Anderson César (accs2) e Mikael Vidal (mvs2)

Centro de Informática UFPE
Universidade Federal de Pernambuco

Abordagem

- Apresentação do domínio escolhido
- Divisão do Trabalho
- Classificador de Páginas
- Web Crawler
  - Heurísticas Utilizadas
  - Comparando Dados
  - Utilização do modelo de classificação de página
- Problemas encontrados

Domínio Escolhido

- Americanas
- Banggod
- Kabum
- Amazon
- Alibaba
- Submarino
- Casas Bahia
- Fast Shop
- Ebay
- Magazine Luiza

Xiaomi Redmi 10 Global Version 6,5 polegadas 90Hz 50MP Quad Câmera 6GB Octa core 4G Smartphone - Versão de outra área Mar azul

Marca: XIAOMI 4 perguntas respondidas ID: 1895259

11.11 SHOPPING FESTIVAL

SALE R$1.097,76 R$1.736,67 -37%

Preço mais baixo em 54 dias

Venda promocional de **Nov 10 a Nov 13**

R$11,62 permissão para novos usuários

6 x R$182,96 sem juros ›

Versão: Versão de outra área

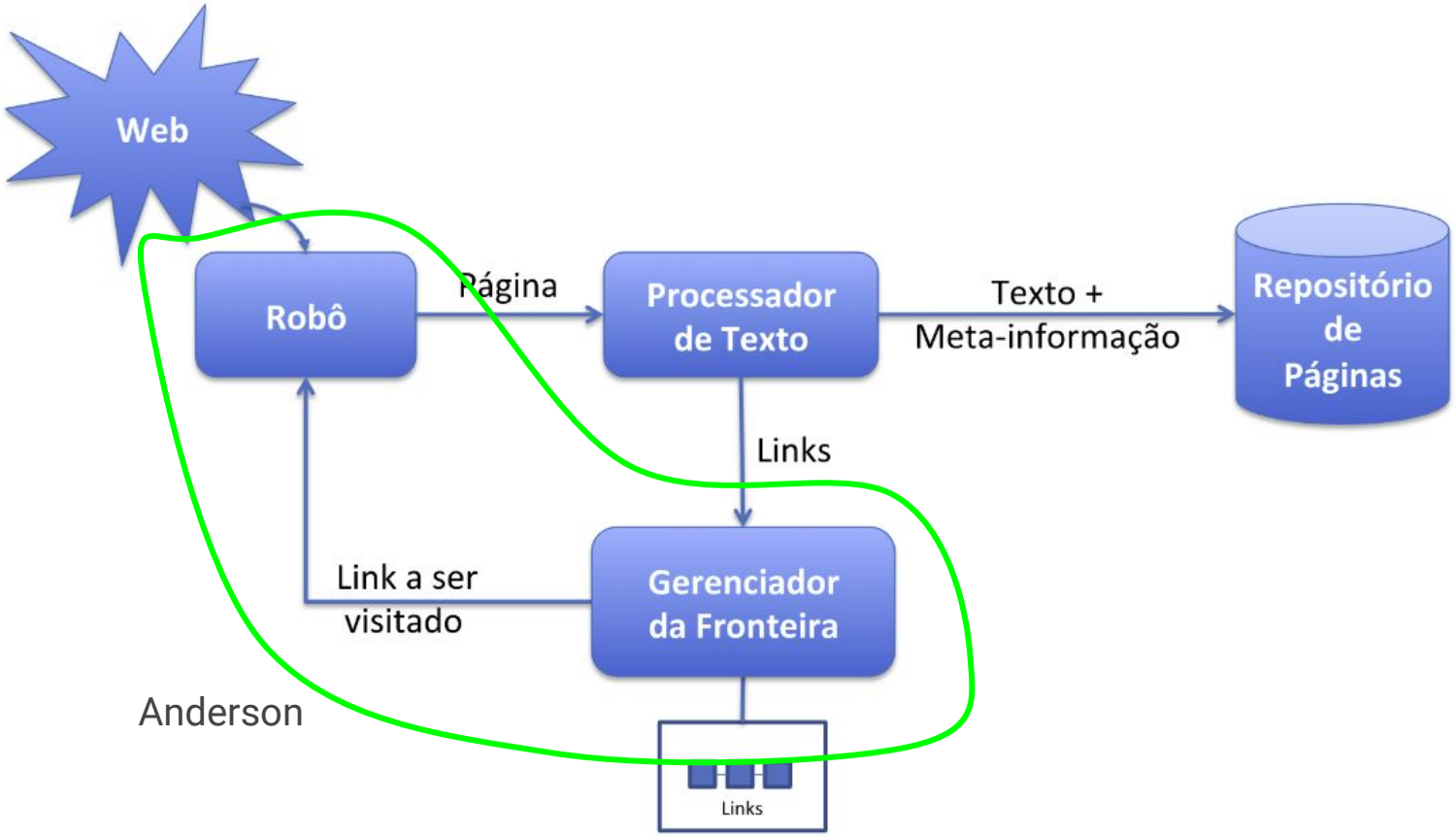
Versão de outra área

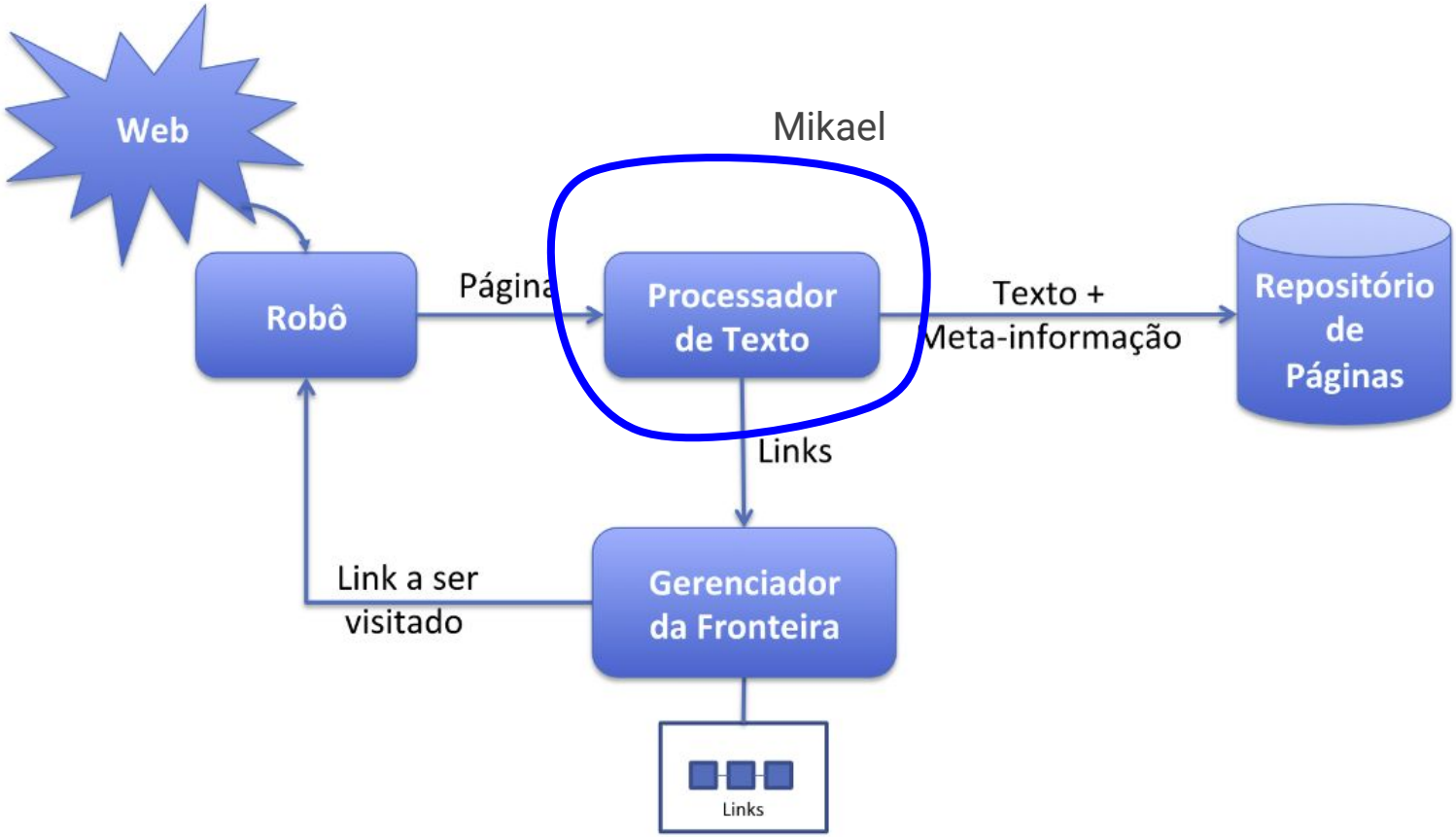
cor.: Mar azul

Enviar De: HK

HK

QTY: 1


Web
Robô
Página
Processador de Texto
Texto + Meta-informação
Repositório de Páginas
Links
Link a ser visitado
Gerenciador da Fronteira
Anderson
Links


Web
Mikael
Robô
Página
Processador de Texto
Texto + Meta-informação
Repositório de Páginas
Links
Link a ser visitado
Gerenciador da Fronteira
Links

Classificador de Páginas

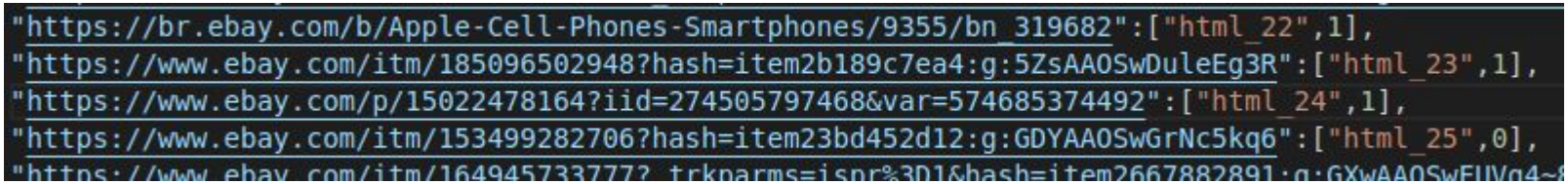
1. Rotular exemplos positivos e negativos (10 positivos e 10 negativos por site)
2. Criar o conjunto de features (ex.: bag of words) usando feature selection (ex. frequência ou information gain)
3. Treinar o classificador com uma ferramenta de ML (ex.: scikit-learn, weka etc)
   - Métodos: Naïve bayes, Decision tree (J48), SVM (SMO), Logistic regression (logistic), Multilayer perceptron
   - Extra: otimizar hiper-parâmetros e diagnosticar modelos
4. Comparar estratégias:
   - Accuracy, precision e recall
   - Tempo de treinamento
   - **Mostrar tabela com os resultados**

Rotulagem de sites

Inicialmente tinham disso ***apenas 40 páginas rotuladas***

Depois foram cerca de ***85 páginas rotuladas***

Estrutura utilizada para ajudar no backup da ***informação que foi rotulada,*** o ***nome de arquivo*** e a sua ***classe*** (**0** - não relevante a smartphone , **1** - relevante).



Conjunto de Features - Bag Of Words

- Escolhido para testes iniciais
- Processamento de texto associados
  - Remoção de caracteres especiais e acentuação
  - Remoção de números e palavras com (***Ex: 1028x ou 122331231***)
  - Stopwords (Portugues e Inglês)
- CountVectorizer()

	aaa	aaaa	aaaaaa	aac	aacute	aad	aaqme	aas	aazl	abaixaram	abaixo	abaixochina	abajures	abas	abd	abdomen
0	6	0	0	0	0	1	0	1	2	0	0	0	0	0	0	0
1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	1	0	0	0	0	14	0	0	1	0	0
3	0	0	0	0	0	1	0	0	0	0	14	0	0	0	0	0
4	0	0	0	0	0	1	0	0	0	0	14	0	0	0	0	0
5	0	0	0	0	0	1	0	0	0	0	14	0	0	0	0	0
6	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	1	0	0	0	0	14	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0
11	0	0	0	15	0	0	0	0	0	0	7	0	0	0	0	0
12	0	0	0	0	3	0	0	0	0	0	7	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0
17	0	0	0	1	0	0	0	0	0	0	7	0	0	0	0	1
18	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
19	1	0	0	0	0	0	0	0	0	0	0	0	0	0	12	0

Conjunto de Features - Information Gain

- Tabela com cerca de **16904 features**
- Utilização da ***RandomForestClassifier***
    - FEATURE IMPORTANCE
- Redução de Features maiores que **0.005**
- Redução para **65** features

Treinamento o classificador

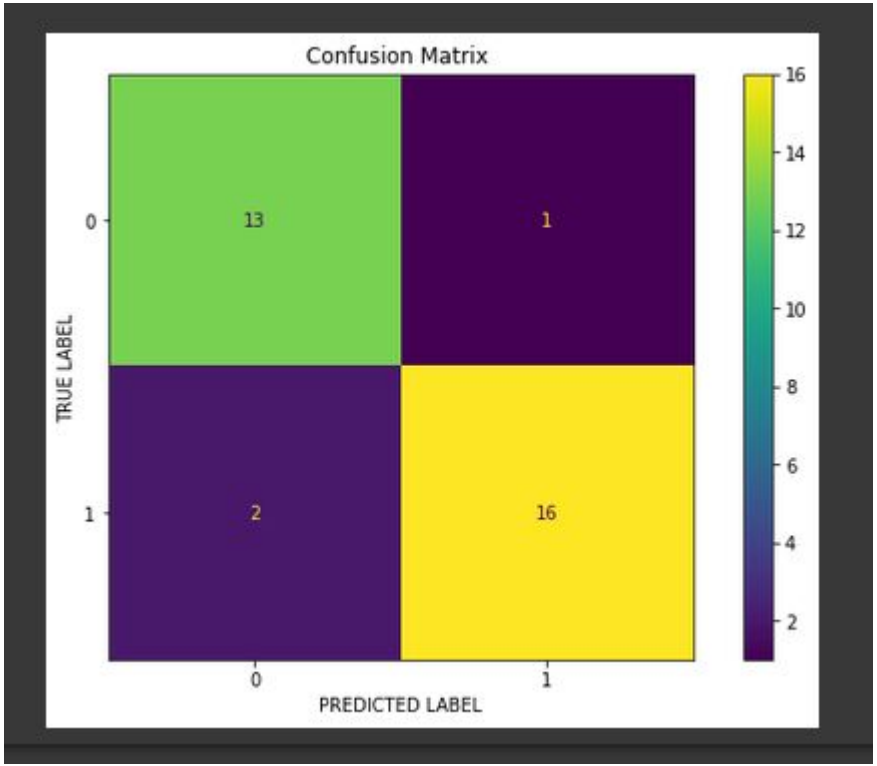
- Naïve bayes,
- Decision tree (J48),
- RandomForestClassifier
- SVM (SMO),
- Logistic regression (logistic),
- Multilayer perceptron

Resultados - Antes de Feature Selection

Métodos	0 ou 1	ACCURACY	PRECISION	RECALL	F1-SCORE	Time
Naive Bayes	0		0.67	0.67	0.67	
	1	0.6	0.6	0.5	0.5	18.9 ms
Decision tree	0		0.62	0.83	0.71	
	1	0.6	0.5	0.25	0.33	5.04 ms
SVM	0		0.57	0.67	0.62	
	1	0.5	0.33	0.25	0.29	24.8 ms
Logistic Regression	0		0.57	0.67	0.62	
	1	0.5	0.33	0.25	0.29	310 ms
Multilayer perceptron	0		0.67	0.67	0.67	
	1	0.6	0.5	0.5	0.5	248 ms

Resultados - Após Feature Selection

Métodos	0 ou 1	ACCURACY	PRECISION	RECALL	F1-SCORE	Time
Naive Bayes	0		0.87	0.93	0.9	
	1	0.90625	0.94	0.89	0.91	1.82 ms
Decision tree	0		0.8	0.86	0.83	
	1	0.84375	0.88	0.83	0.86	5.04 ms
RandomForestClassifier	0		0.87	0.93	0.9	
	1	0.90625	0.94	0.89	0.91	594 ms
SVM	0		0.87	0.93	0.9	
	1	0.90625	0.94	0.89	0.91	3.06 ms
Logistic Regression	0		0.86	0.86	0.86	
	1	0.875	0.89	0.89	0.89	19.1 ms
Multilayer perceptron	0		0.88	0.5	0.64	
	1	0.75	0.71	0.94	0.81	248 ms


Confusion Matrix
13
1
2
16
0
1
TRUE LABEL
0
1
PREDICTED LABEL
16
14
12
10
8
6
4
2

Resultados - Após Feature Selection e alterando alguns hiperparâmetros gerais

Métodos	0 ou 1	ACCURACY	PRECISION	RECALL	F1-SCORE	Time
Naive Bayes	0		0.87	0.93	0.9	6.14 ms
	1	0.90625	0.94	0.89	0.91	
Decision tree	0		0.75	0.86	0.8	1.57 ms
	1	0.8125	0.88	0.78	0.82	
SVM	0		0.8	0.57	0.67	2.73 ms
	1	0.75	0.73	0.89	0.8	
Logistic Regression	0		0.89	0.57	0.7	
	1	0.78125	0.74	0.94	0.83	2.37 ms
Multilayer perceptron	0		0.7	0.5	0.58	
	1	0.6875	0.68	0.83	0.75	390 ms

Web Crawler

1. Encontrar manualmente 10 sites no domínio
2. Implementar 2 estratégias (1000 páginas visitadas por site):
   - Baseline: busca em largura
   - Heurística (usar âncora)
   - **Extra**: implementar um classifcador de links
3. Comparar estatégias:
   - Harvest ratio: (número de páginas relevantes coletadas)/(total de páginas visitadas)
   - **Mostrar tabela com resultados**

- Importante:
  - Evitar sobrecarregar o site
  - Respeitar o robots.txt
  - Detectar o conteúdo da página com o campo Content-Type

Web Crawler

1. **download_url:** Responsável pelo Download das Páginas

2. **get_links:** Responsável por Recuperar os links presentes no html

3. **ranker:** Função que realiza o Rankeamento das Páginas

```
domain_words = {"smartphone","celular", "samsung", "apple","motorola","xiaomi", "android", "ios",
"galaxy", "moto", "lenovo", "tela", "zenfone", "lg",
"telefone","asus","camera","core","mobile","phone","cell", "bateria", "memory", "pixel"}
```

Web Crawler

	Busca em Largura	Busca c/ Heurística
Banggod	0.015	0.488
Kabum	0.002	0.268
Ebay	0.048	0.965
Magazine Luiza	0.0	0.019
Alibaba	0.0	0.066
Amazon	0.0	0.233

Web Crawler

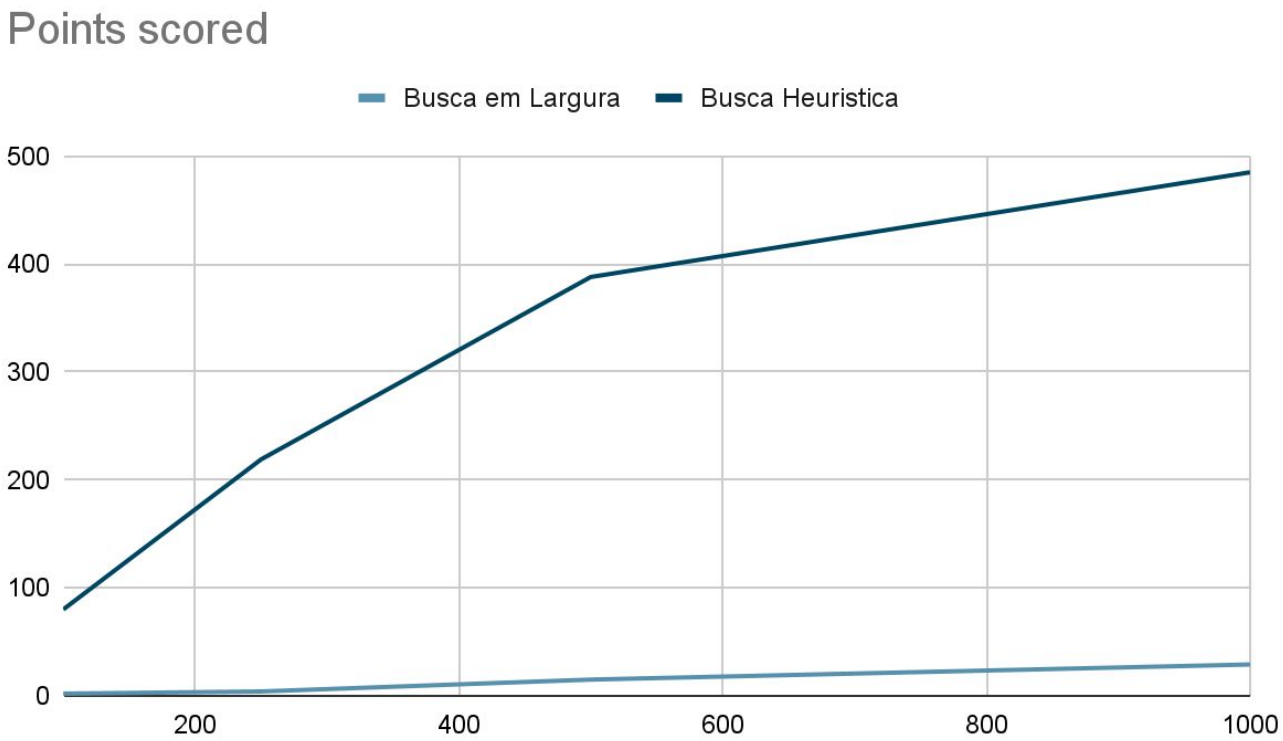
Busca em Largura:

Harvest Ratio Médio: 0.011

Busca c/ Heurística:

Harvest Ratio Médio: 0.339

Web Crawler


Points scored
Busca em Largura
Busca Heuristica
500
400
300
200
100
0
200
400
600
800
1000

Web Crawler

Bakeey High Speed 16GB 32GB 64GB 128GB Class 10 TF/SD Memory Card Flash Drive With Card Adapter For iPhone 12 For Samsung Galaxy S21 Smartphone Tablet Switch Speaker Drone Car DVR GPS Camera

R$52,22

~~R$110,24~~

Buy now

WANSENDA 8GB 16GB 32GB 64GB 128GB 256GB High Speed TF/ SD Memory Card With Card Adapter For Mobile Phone Tablet GPS Camera

R$89,97

~~R$95,72~~

Buy now

Problemas Encontrados