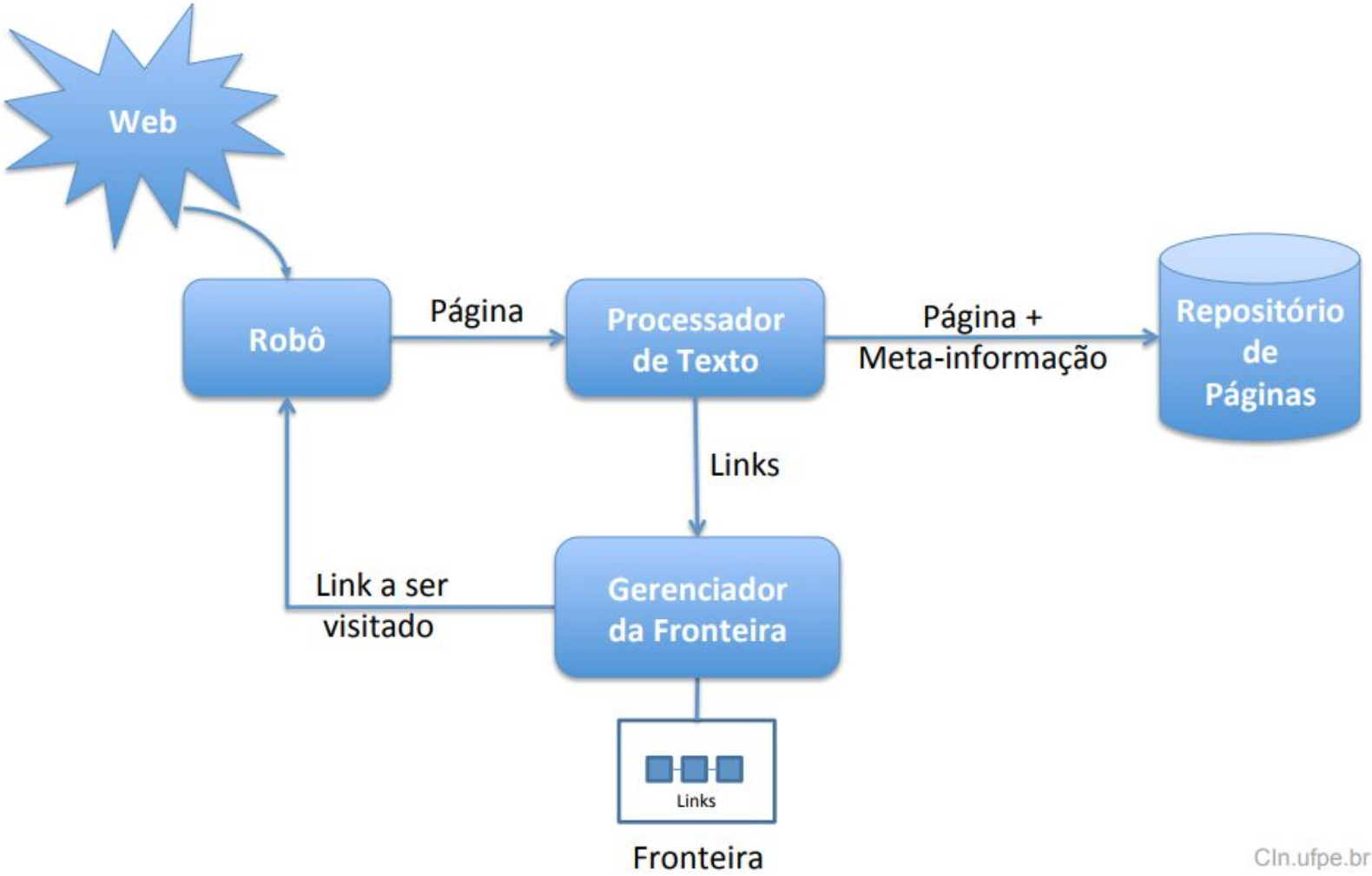
Recuperação de Informação

Projeto 1:
Crawler *(Larícia Mota - lmmc2)*
Classificador *(Gabriel Vinicius - gvmgs)*

Sites no domínio: Séries

- Rotten Tomatoes
- IMDB
- The Movie DB
- Trakt
- TV Guide
- Metacritic
- The TV DB
- TV.com

Crawler

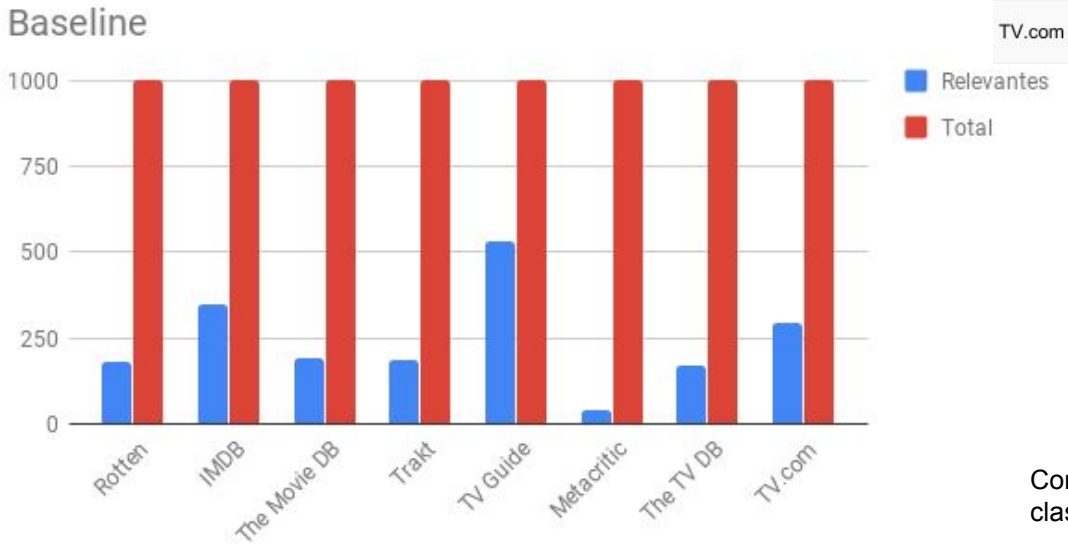

Web
Robô
Página
Processador de Texto
Página + Meta-informação
Repositório de Páginas
Links
Link a ser visitado
Gerenciador da Fronteira
Links
Fronteira

Estratégias

- Baseline
  - BFS - Busca em largura
- Heurísticas
  - Heurística 1 - Palavras positivas
  - Heurística 2 - Palavras positivas e negativas

Resultados Baseline

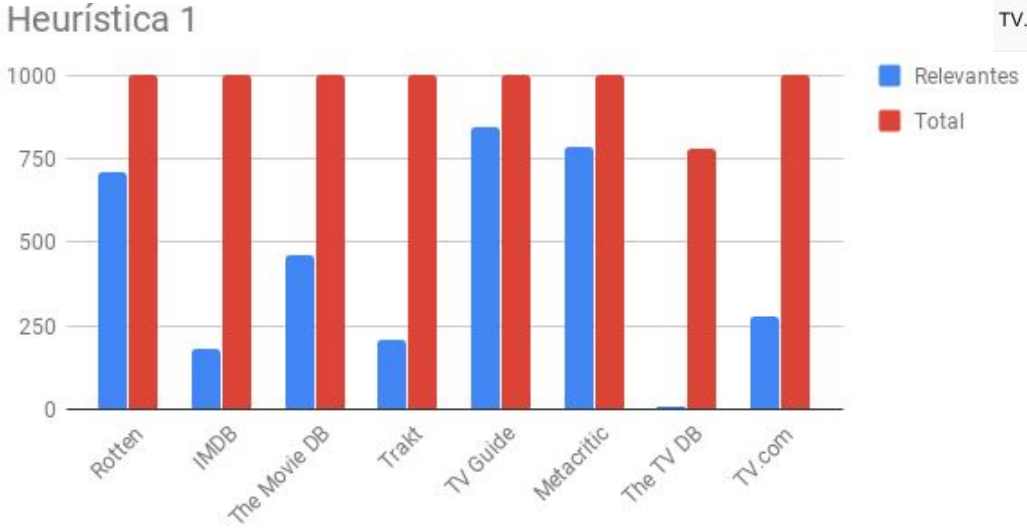
Site	Relevantes	Total	Harvest Ratio
Rotten Tomatoes	181	1000	0,181
IMDB	347	1000	0,347
The Movie DB	191	1000	0,191
Trakt	187	1000	0,187
TV Guide	531	1000	0,531
Metacritic	37	1000	0,037
The TV DB	168	1000	0,168
TV.com	294	1000	0,294



Comparação realizada através de algoritmo simples de classificação de páginas, analisando-se palavras presentes.

Resultados Heurística 1

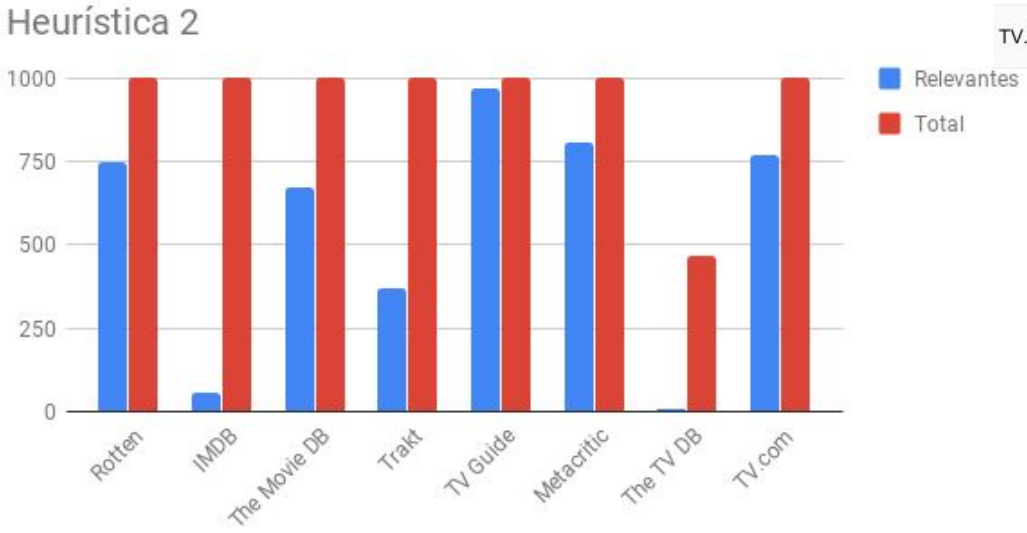
Site	Relevantes	Total	Harvest Ratio
Rotten Tomatoes	710	1000	0,71
IMDB	182	1000	0,182
The Movie DB	460	1000	0,46
Trakt	205	1000	0,205
TV Guide	845	1000	0,845
Metacritic	786	1000	0,786
The TV DB	8	781	0,01024327784891
TV.com	276	1000	0,276



Comparação realizada através de algoritmo simples de classificação de páginas, analisando-se palavras presentes.

Resultados Heurística 2

Site	Relevantes	Total	Harvest Ratio
Rotten Tomatoes	746	1000	0,746
IMDB	56	1000	0,056
The Movie DB	672	1000	0,672
Trakt	371	1000	0,371
TV Guide	971	1000	0,971
Metacritic	805	1000	0,805
The TV DB	5	465	0,01075268817
TV.com	767	1000	0,767



Comparação realizada através de algoritmo simples de classificação de páginas, analisando-se palavras presentes.

Classificador

Abordagem

- 220 amostras
- 91088 features (incluindo estrutura HTML)
- TF-IDF
  - TF-Normalizado x IDF
  - Palavras da estrutura HTML mantidas
  - Redução para 51051 features
- Classificadores
  - 70% Teste - 10% Validação - 20% Teste
  - MLP (**com variações**), Decision Tree, Naive Bayes, SVM, Logistic Regression, **Random Forest**, **Gradient Boosting**.

Comparação de Métricas - Teste

	MLP - 20 neuronios	MLP - 40 neuronios	Decision Tree	Random Forest	Naive Bayes	SVM	Logistic Regression	MLP - ADAMAX	MLP - SGD	Gradient Boosting
Precision	91.304	95.455	88.000	95.652	78.571	86.364	95.455	98.431	99.802	98.976
Recall	95.455	95.455	100.000	100.000	100.000	86.364	95.455	90.909	95.455	100.000
Accuracy	93.478	95.652	93.478	97.826	86.957	86.957	95.652	95.652	97.826	97.826
Curva Roc	99.432	99.621	93.750	99.432	87.500	90.152	99.432	98.295	99.811	99.432

Matriz de Confusão

MLP - 20 N

	T	F
T	22	2
F	1	21

MLP - 40 N

	T	F
T	23	1
F	1	21

Decision Tree

	T	F
T	21	3
F	0	22

Random Forest

	T	F
T	23	1
F	0	22

Naive Bayes

	T	F
T	18	6
F	0	22

SVM

	T	F
T	21	3
F	3	19

Logistic Regression

	T	F
T	23	1
F	1	21

MLP - ADAMAX

	T	F
T	24	0
F	2	20

MLP - SGD

	T	F
T	24	0
F	1	21

Gradient Boosting

	T	F
T	23	1
F	0	22

Tempo de Treinamento

	MLP - 20 neuronios	MLP - 40 neuronios	Decision Tree	Random Forest	Naive Bayes	SVM	Logistic Regression	MLP - ADAMAX	MLP - SGD	Gradient Boosting
Segundos	9.87	15.06	0.2	0.07	0.27	7.47	0.37	9.95	22.45	14.40