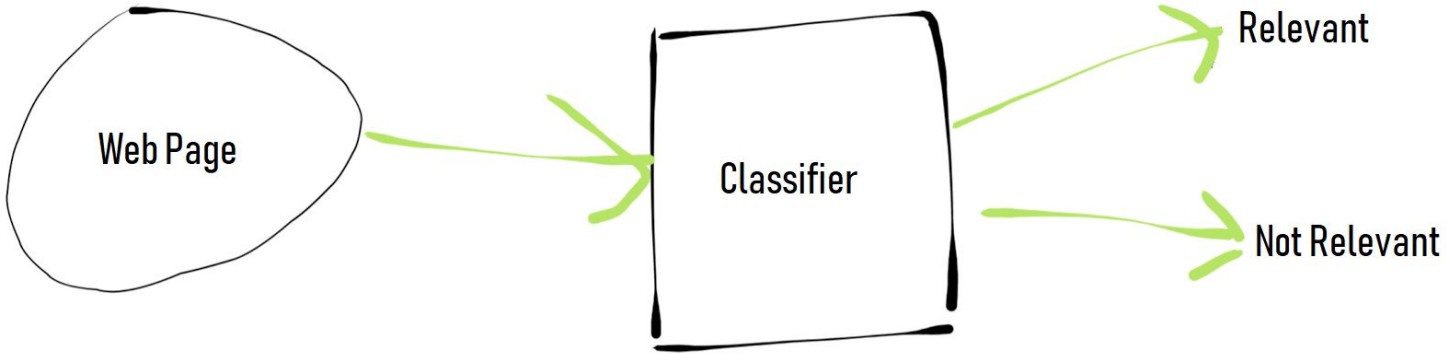
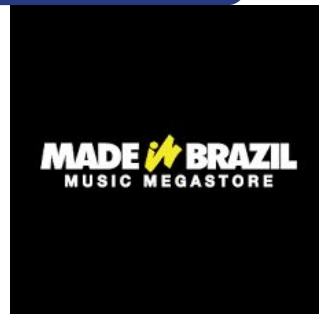
Coleta e Busca de Entidades Estruturadas em um Domínio

Violões
Lucas Melo, Mário Wessen, Mateus Gonçalves

Tarefa 2: Detectar Páginas com Instâncias



2.1 Rotular Exemplos Positivos e Negativos



2.2 Criação de Conjunto de Features

Bag of Words

Criação de Bag of Words: Filtros

BoW_plaintext	18.067 palavras
BoW_lowercase	18.067 palavras
BoW_noStopwords (lowercase + noStopwords)	15.968 palavras
BoW_noStopwords (lowercase + noStopwords + stemming)	15.220 palavras

2.3 Classificadores Utilizados

Gaussian Naive Bayes

Multinomial Naive Bayes

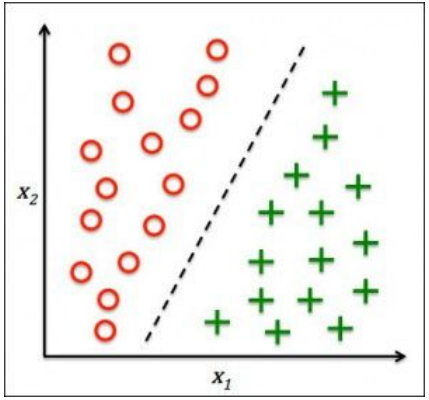
Logistic Regression

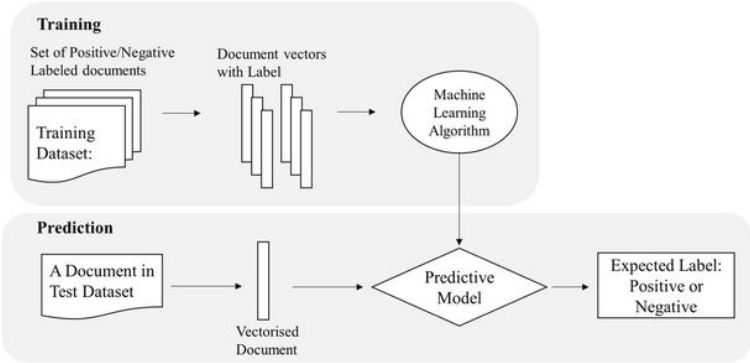
Multilayer Perceptron

Random Forest

Decision Tree

Support Vector Machines





Medições

Gaussian NB

	Accuracy	Precision	Recall	Training Time
bow_plain	0.91	0.92	0.93	0.31249 s
bow_lowercase	0.92	0.92	0.94	0.07812 s
bow_nostopword	0.92	0.92	0.94	0.07812 s
bow_stemming	0.90	0.91	0.92	0.09297 s
bow_vectorizer	0.96	0.96	0.96	0.24414 s

Multinomial NB

	Accuracy	Precision	Recall	Training Time
bow_plain	0.89	0.88	0.94	0.03125 s
bow_lowercase	0.88	0.88	0.92	0.08583 s
bow_nostopword	0.85	0.87	0.88	0.03124 s
bow_stemming	0.82	0.87	0.81	0.10331 s
bow_vectorizer	0.79	0.83	0.79	0.01562 s

Multilayer Perceptron

	Accuracy	Precision	Recall	Training Time
bow_plain	0.88	0.91	0.89	8.00461 s
bow_lowercase	0.91	0.92	0.94	3.01587 s
bow_nostopword	0.92	0.91	0.94	3.58022 s
bow_stemming	0.93	0.93	0.95	1.72491 s
bow_vectorizer	0.86	0.89	0.86	6.81927 s

Logistic Regression

	Accuracy	Precision	Recall	Training Time
bow_plain	0.93	0.92	0.98	0.15590 s
bow_lowercase	0.95	0.93	0.98	0.50960 s
bow_nostopword	0.92	0.90	1.0	0.12539 s
bow_stemming	0.93	0.90	1.0	0.38176 s
bow_vectorizer	0.88	0.89	0.88	0.09374 s

Random Forest

	Accuracy	Precision	Recall	Training Time
bow_plain	0.92	0.95	0.96	1.15658 s
bow_lowercase	0.92	0.94	0.96	0.32847 s
bow_nostopword	0.92	0.94	0.96	0.32812 s
bow_stemming	0.89	0.89	0.96	0.31249 s
bow_vectorizer	0.96	0.96	0.96	0.44086 s

Decision Tree

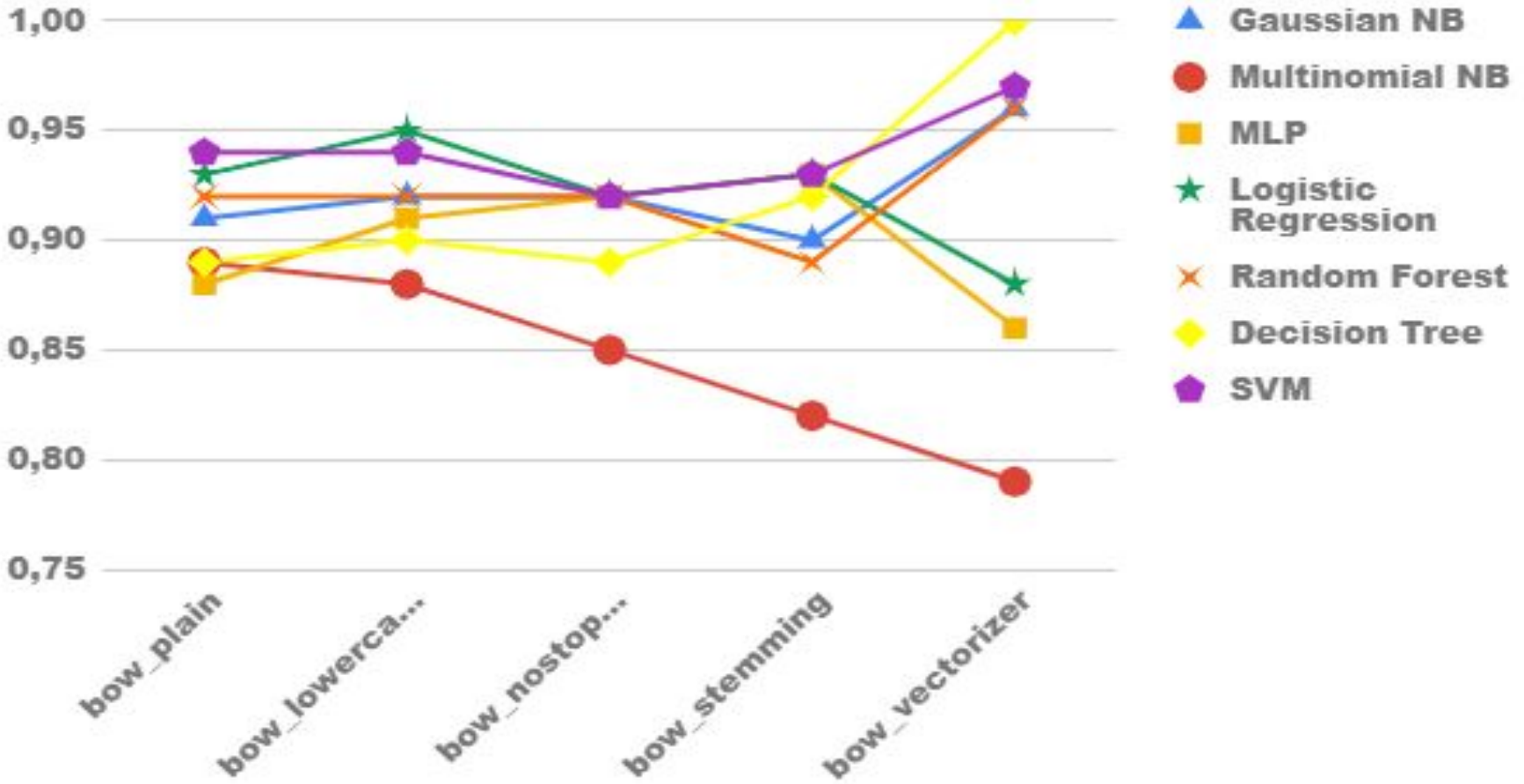
	Accuracy	Precision	Recall	Training Time
bow_plain	0.89	0.86	0.84	0.16350 s
bow_lowercase	0.90	0.91	0.85	0.17511 s
bow_nostopword	0.89	0.88	0.83	0.10937 s
bow_stemming	0.92	0.90	0.86	0.09375 s
bow_vectorizer	1.0	1.0	1.0	0.21874 s

Support Vector Machines

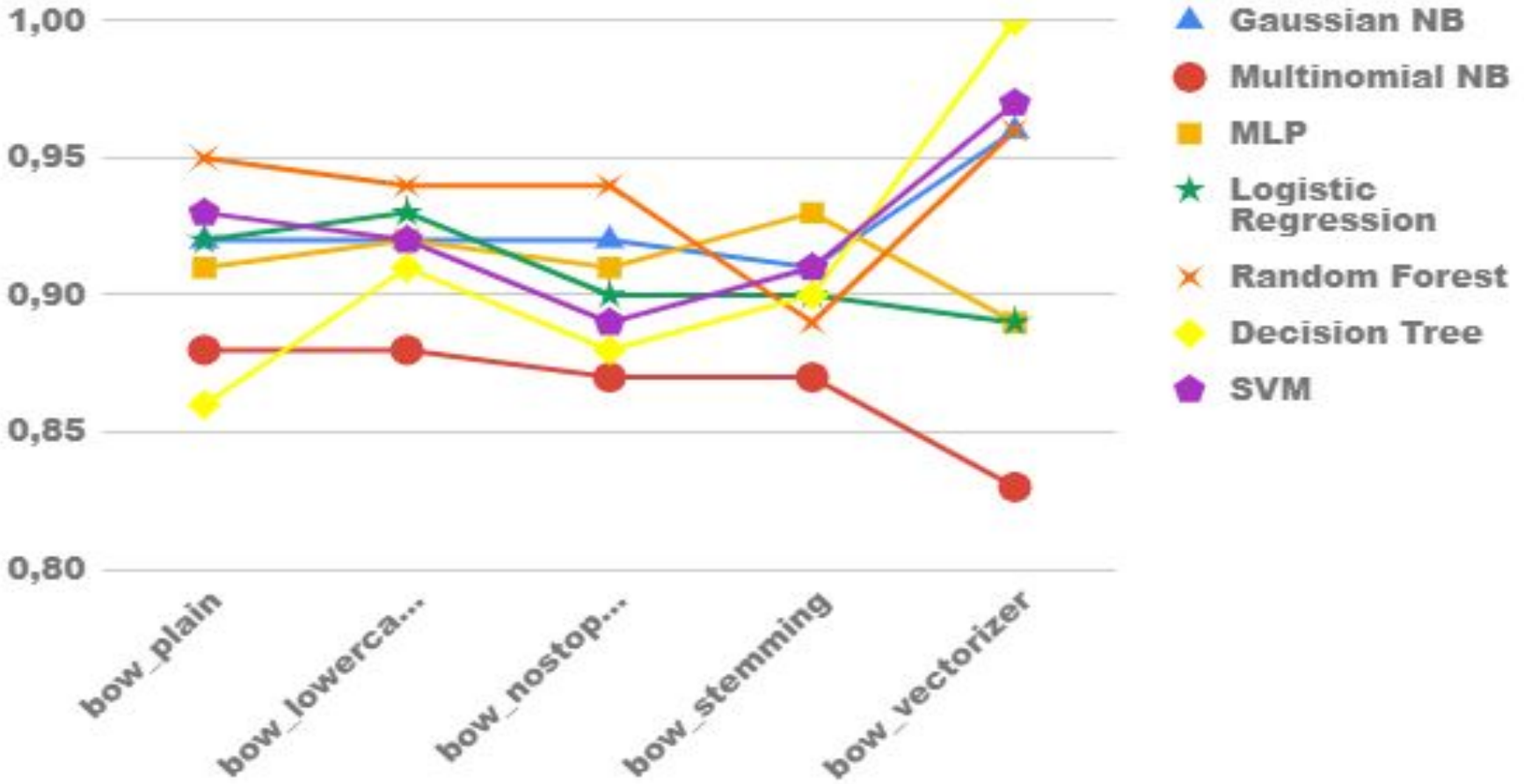
	Accuracy	Precision	Recall	Training Time
bow_plain	0.94	0.93	0.99	0.44316 s
bow_lowercase	0.94	0.92	0.98	0.49111 s
bow_nostopword	0.92	0.89	0.99	0.36505 s
bow_stemming	0.93	0.91	0.99	1.06472 s
bow_vectorizer	0.97	0.97	0.97	0.09375 s

Gráficos

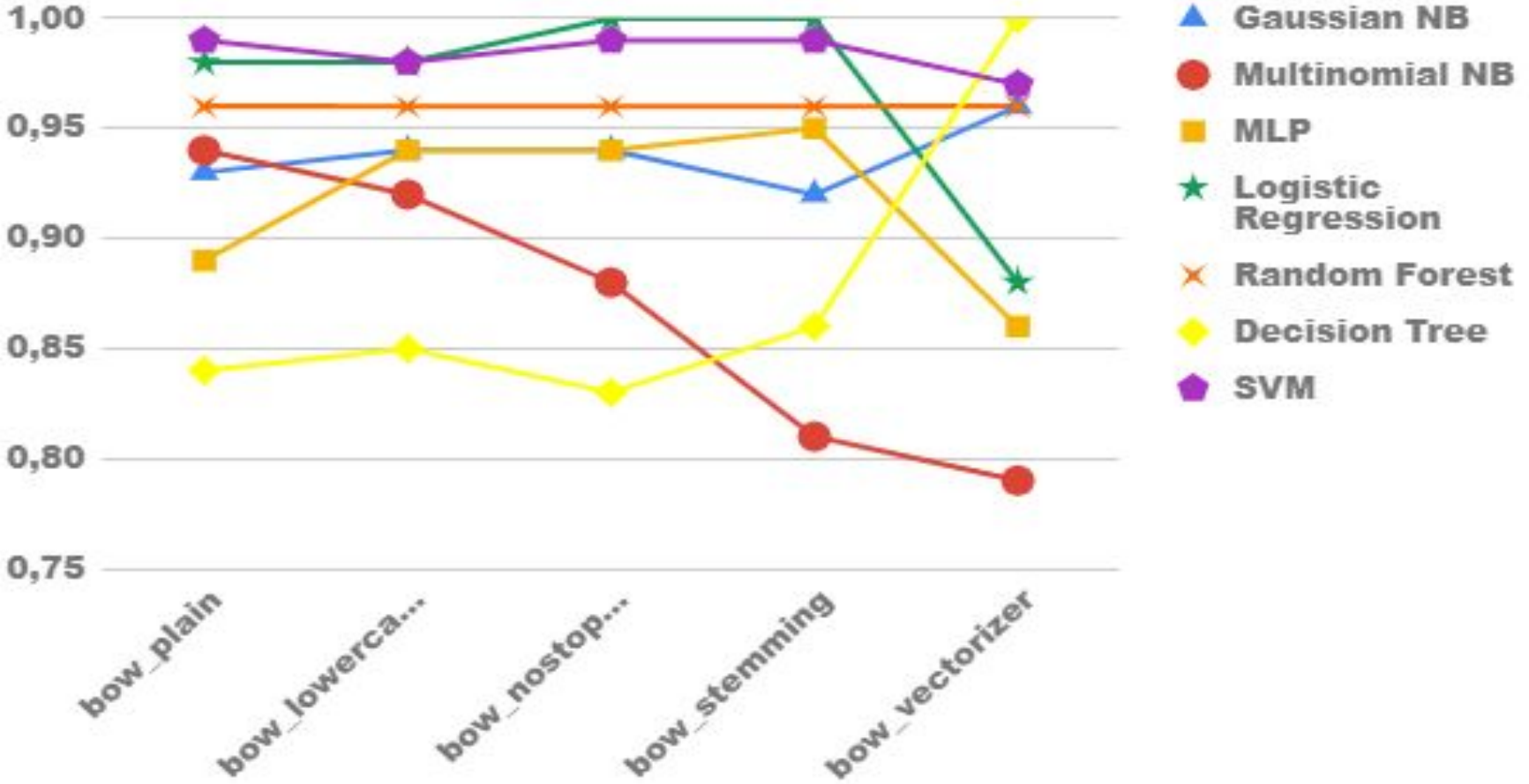
Accuracy


1,00
0,95
0,90
0,85
0,80
0,75
bow_plain
bow_lowerca...
bow_nostop...
bow_stemming
bow_vectorizer
Gaussian NB
Multinomial NB
MLP
Logistic Regression
Random Forest
Decision Tree
SVM

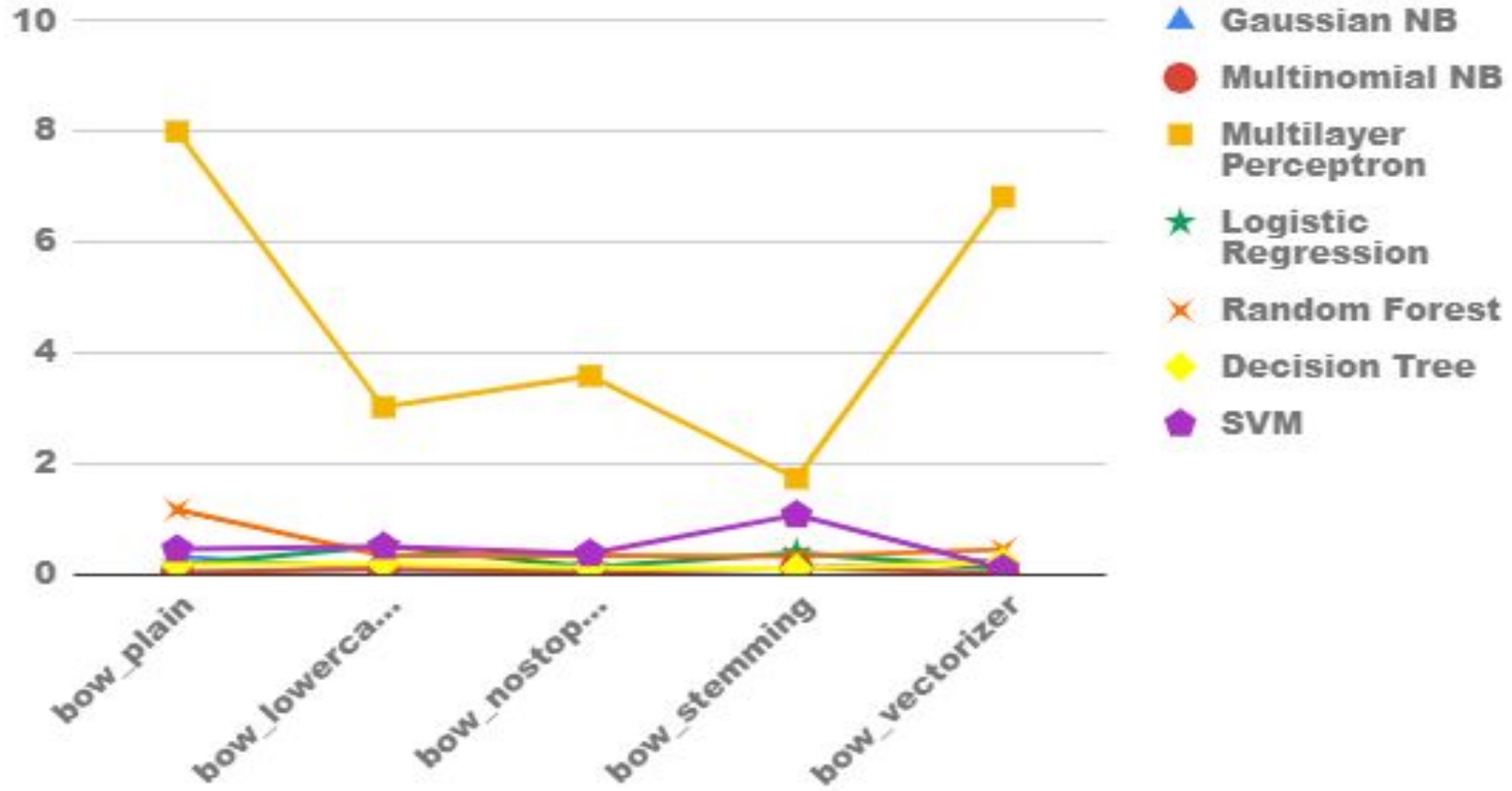
Precision


1,00
0,95
0,90
0,85
0,80
bow_plain
bow_lowerca...
bow_nostop...
bow_stemming
bow_vectorizer
Gaussian NB
Multinomial NB
MLP
Logistic Regression
Random Forest
Decision Tree
SVM

Recall


1,00
0,95
0,90
0,85
0,80
0,75
bow_plain
bow_lowerca...
bow_nostop...
bow_stemming
bow_vectorizer
Gaussian NB
Multinomial NB
MLP
Logistic Regression
Random Forest
Decision Tree
SVM

Training Time



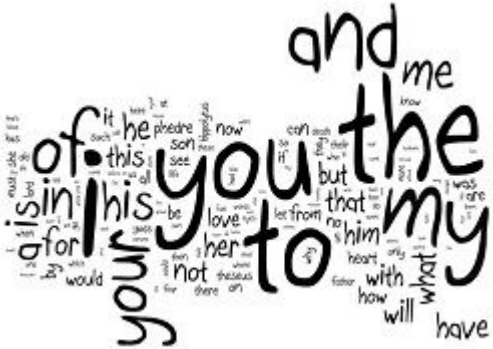
2.4 Avaliação dos Resultados

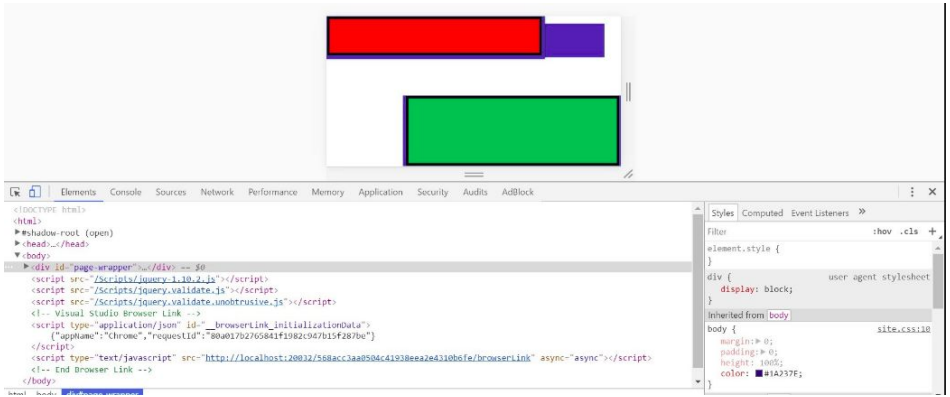
- Métodos utilizados apresentaram resultados relativamente semelhantes para a maioria das Bag of Words.
- Com exceção do Multinomial Naive Bayes que teve desempenho inferior tanto em Acurácia quanto em Precisão e Recall.
- Multilayer Perceptron apresentou, como esperado, tempo de treinamento maior para todos as Bag of Words.

Tarefa 3: Extrair informações

3.1 Técnicas Utilizadas

- Tags diretas
- Redução de espaço de procura
- Stopwords





3.2 Problemas

- Sites que mudam de template
- Sites desestruturados
- Atualização de página

3.3 Resultados

	Recall	Precision	F-Measure
Mil Sons	0.95	0.70	0.80
Nova Music	0.33	0.27	0.30
Mundo Max	0.80	0.86	0.83
Made in Brazil	1.00	0.40	0.57
Playtech	0.00	0.00	0.00
Multisom	0.56	0.58	0.57
Walmart	0.92	0.92	0.92
Casas Bahia	1.00	0.60	0.75

3.4 Avaliação dos Resultados

- A redução de espaço de busca funcionou muito bem, na maioria dos casos
- Técnica de PLN + RegEx se encaixou no domínio que escolhemos
- Técnica de classificação para redução de espaço