Coleta e Busca de Entidades Estruturadas em um Domínio

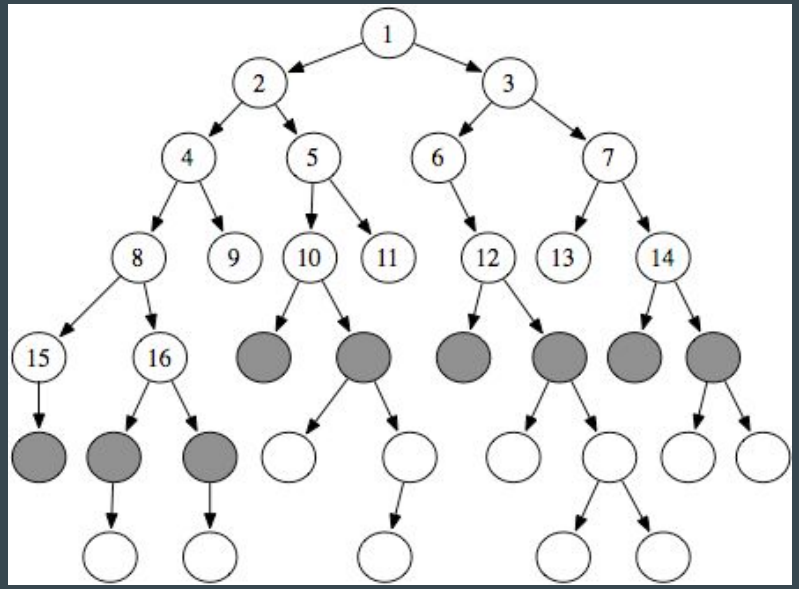
...

Imóveis

Cristiano Oliveira, Renato Sousa e Vinícius Bezerra

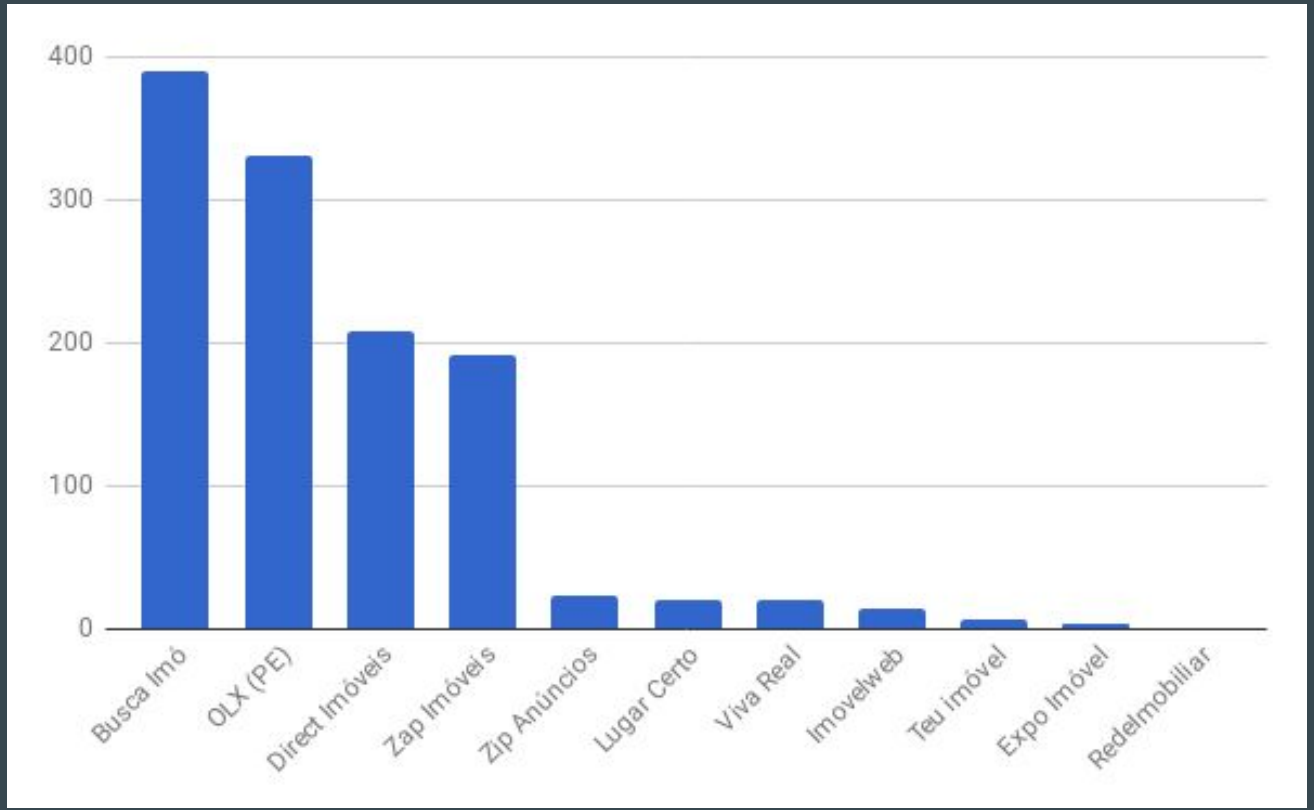
Tarefa 1 : Localizar Páginas Relevantes (crawler)

Tarefa 1.1 : Busca em Largura



- Filtragem do robot.txt realizada;
- Não sobrecarga de acesso assegurada (tempo de espera 1s);
- Sucesso de busca para todas os domínios (1000 links).

Tarefa 1.1 : Busca em Largura - resultados

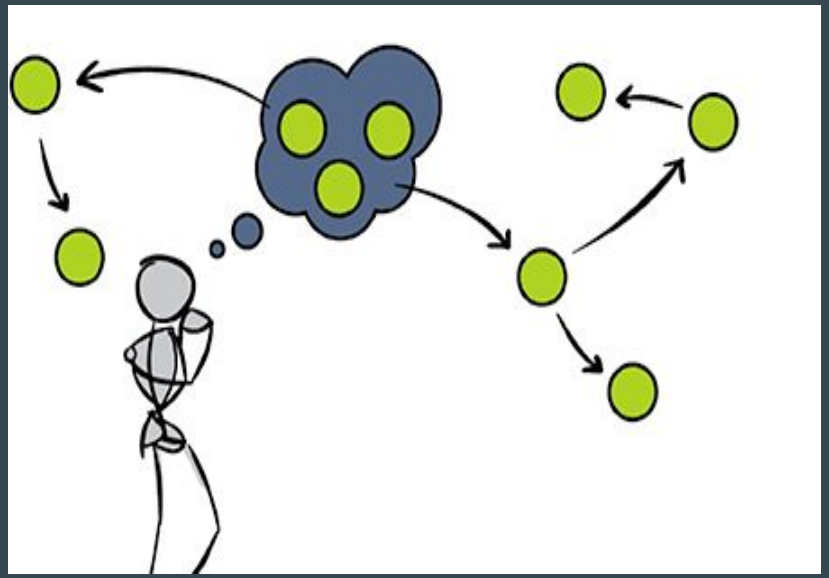


Tarefa 1.1 : Busca em Largura - resultados

Domínio	Harvest Ratio
Busca Imóveis	0.390
OLX (PE)	0.331
Direct Imóveis	0.209
Zap Imóveis	0.192
Zip Anúncios	0.024
Lugar Certo	0.021

Domínio	Harvest Ratio
Viva Real	0.020
Imovelweb	0.014
Teu imóvel	0.007
Expo Imóvel	0.003
RedeImobiliariasecovi	0.0

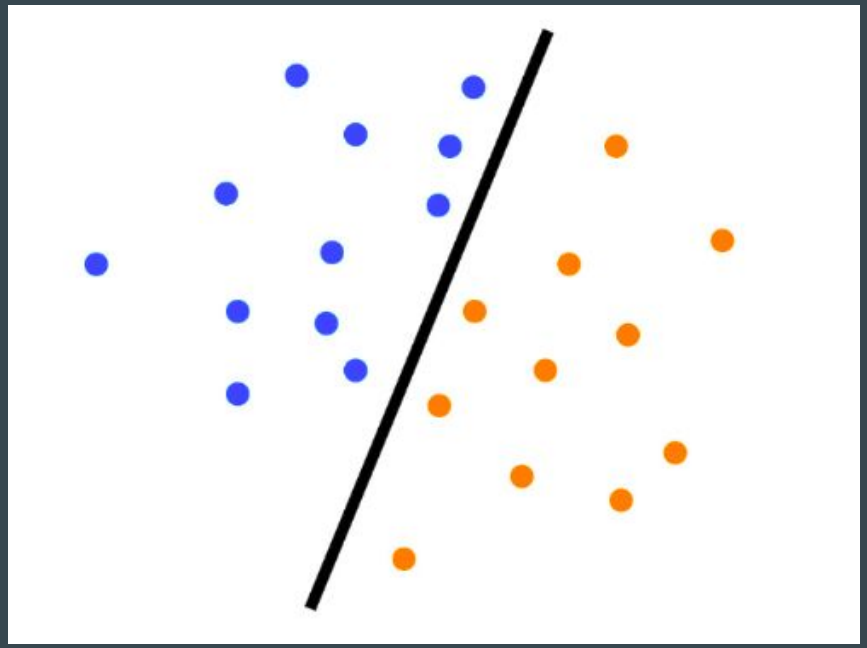
Tarefa 1.2 : Busca Heurística



- Filtragem do robot.txt realizada;
- Não sobrecarga de acesso assegurada (tempo de cálculo ponderado com sleep de 1s);

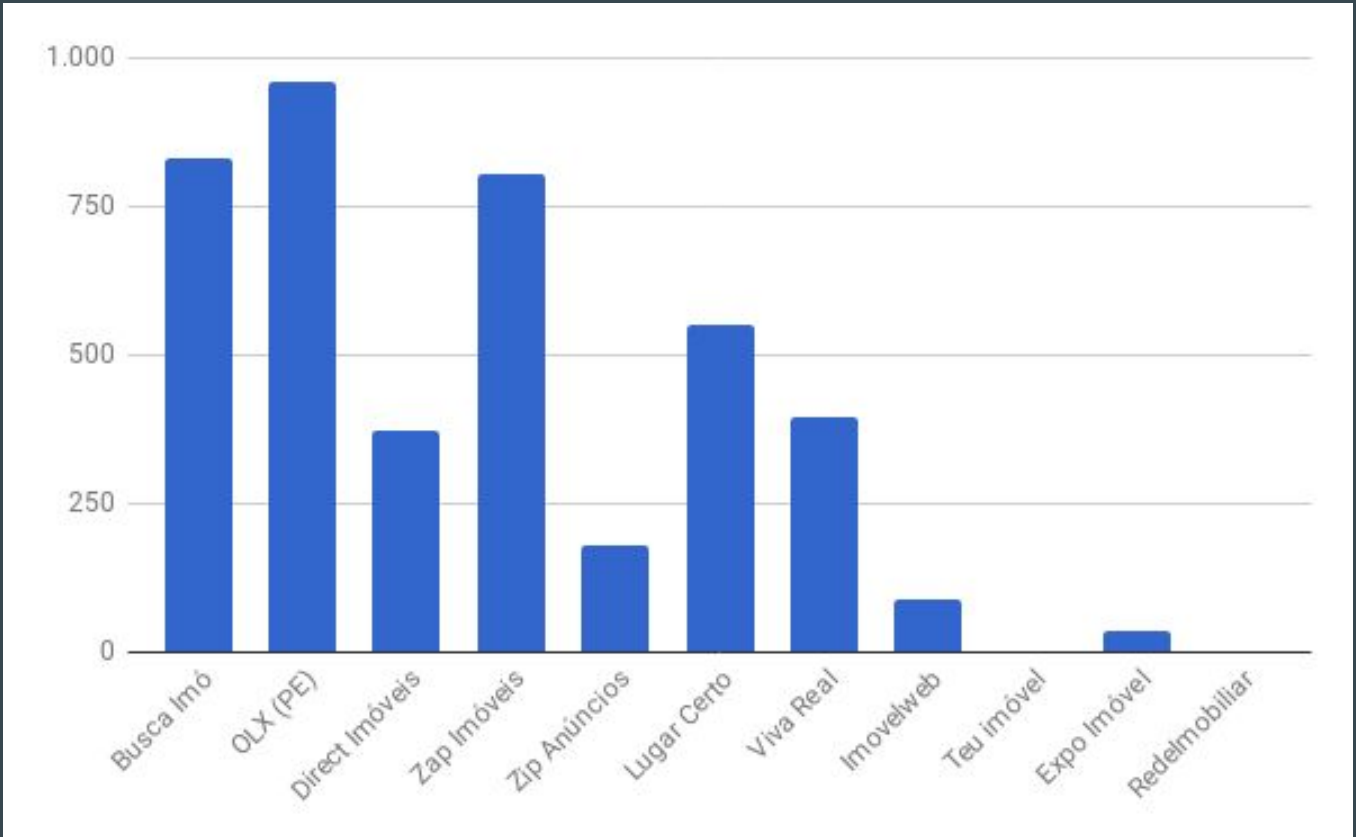
Tarefa 1.2 : Busca Heurística

Função Heurística - Naive Bayes Classifier



- Feita uma base de links manualmente (pos, neg) - 300 links;
- Acurácia = 0.59;
- Primeira heurística: pega links com >70% de ser relevante - “Morre de fome”, fronteiro acaba e não completa a base;
- Segunda heurística: pega links com >60% de ser relevante - mesmo problema;
- Terceira heurística: pega links com >10% - consegue completar as bases.

Tarefa 1.2 : Busca Heurística - resultados

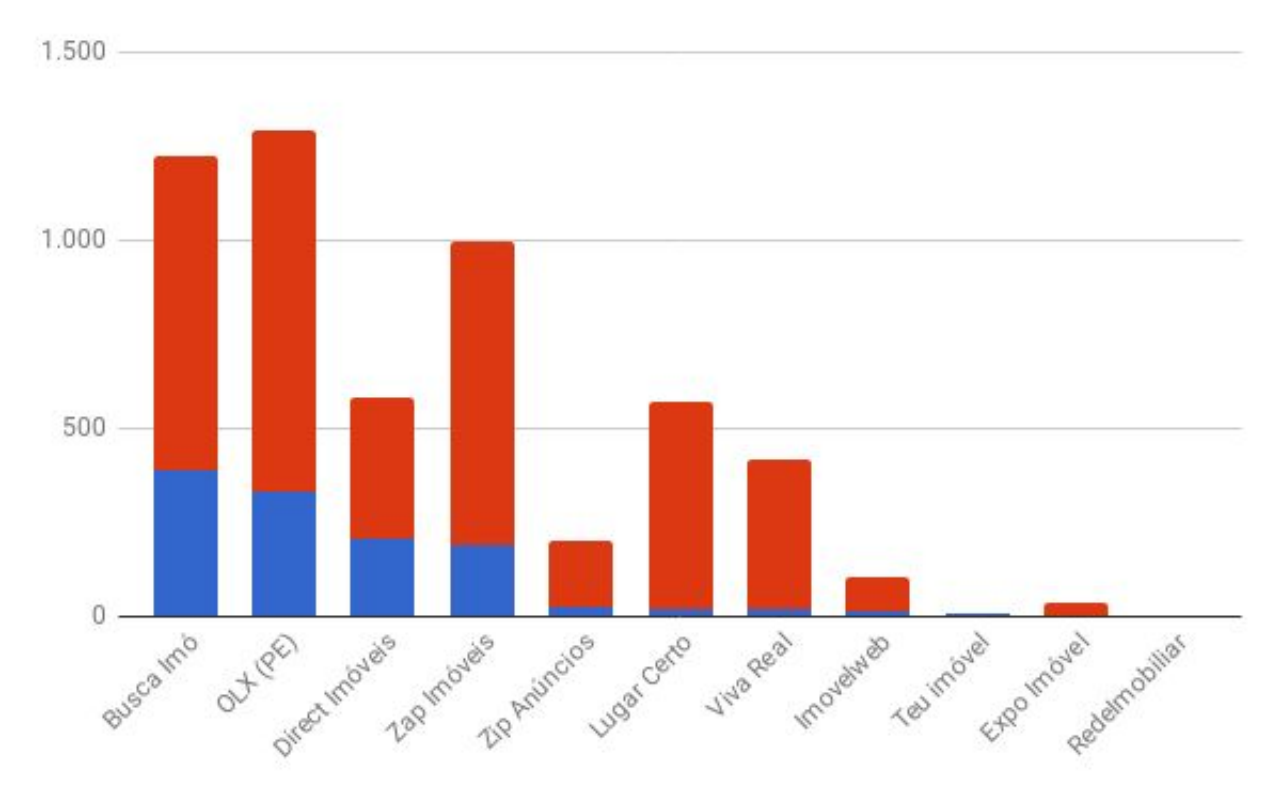


Tarefa 1.2 : Busca Heurística - resultados

Domínio	Harvest Ratio
Busca Imóveis	0.832
OLX (PE)	0.961
Direct Imóveis	0.372
Zap Imóveis	0.806
Zip Anúncios	0.180
Lugar Certo	0.550

Domínio	Harvest Ratio
Viva Real	0.395
Imovelweb	0.090
Teu imóvel	0.0
Expo Imóvel	0.035
RedeImobiliariasecovi	0.0

Tarefa 1.2 : Comparação entre métodos



Tarefa 1.2 : Comparação entre métodos

Domínio	HR (Busca cega)	HR (Heurística)	Df
Busca Imóveis	0.390	0.832	0.442
OLX (PE)	0.331	0.961	0.630
Direct Imóveis	0.209	0.372	0.163
Zap Imóveis	0.192	0.806	0.614
Zip Anúncios	0.024	0.180	0.156
Lugar Certo	0.021	0.550	0.529

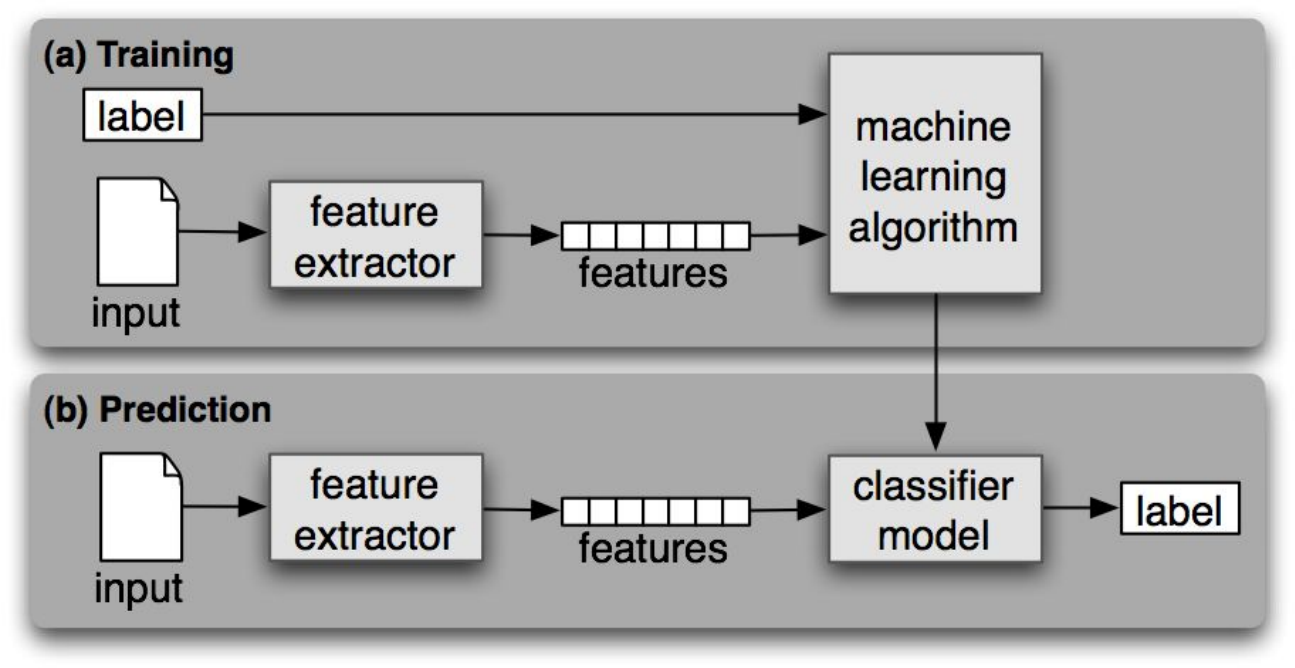
Domínio	HR (Busca cega)	HR (Heurística)	Df
Viva Real	0.020	0.395	0.375
Imovelweb	0.014	0.090	0.076
Teu imóvel	0.007	0.0	- 0.007
Expo Imóvel	0.003	0.035	0.032
RedeImobiliariasecovi	0.0	0.0	0.0

Tarefa 1.2 : Observações e melhorias

- Sites bem estruturados tiveram desempenho melhor para ambas as buscas;

- Possível solução para aumentar precisão do caso heurístico, melhorar a base de treinamento da função heurística ou classificador de links completo.

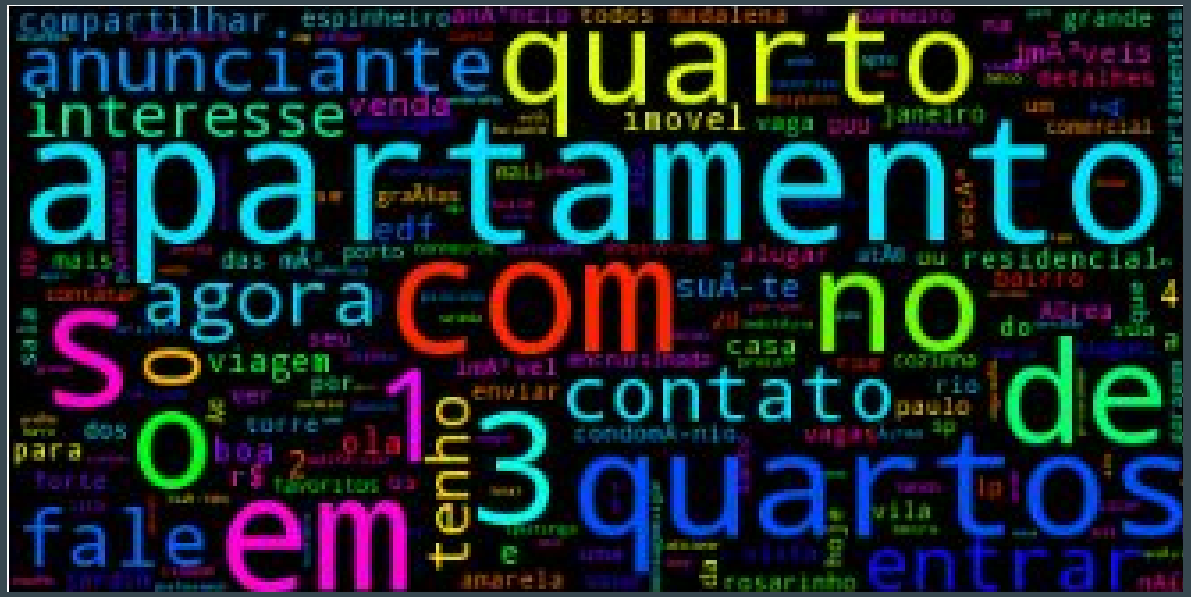
Tarefa 2 : Detectar Páginas com Instâncias



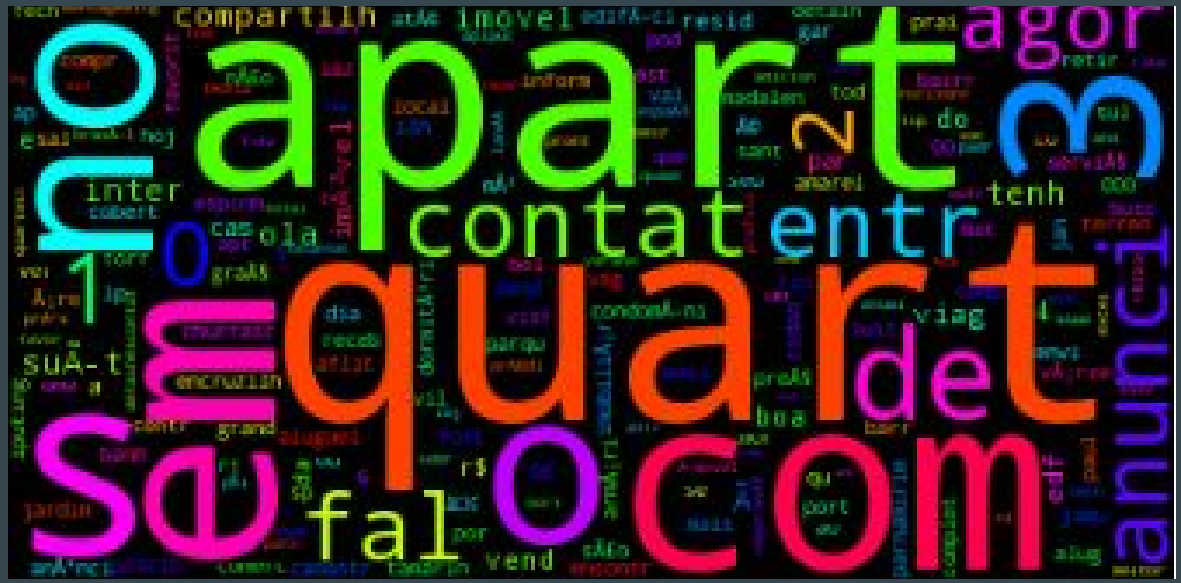
2.1 Rotular Exemplos Positivos e Negativos

2.2 Criar Conjunto de Features (Bag of Words) :

Filtro em caracteres especiais + Palavras minúsculas - (20565 palavras)

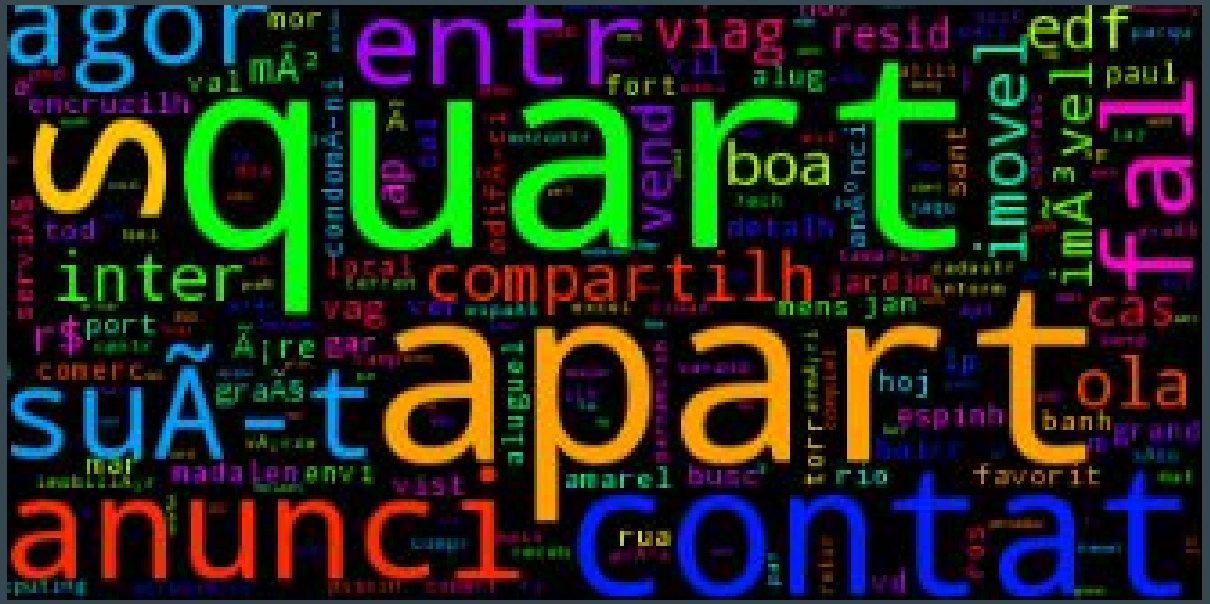


Filtro em caracteres especiais + Palavras minúsculas + Stemming - (17002 palavras)

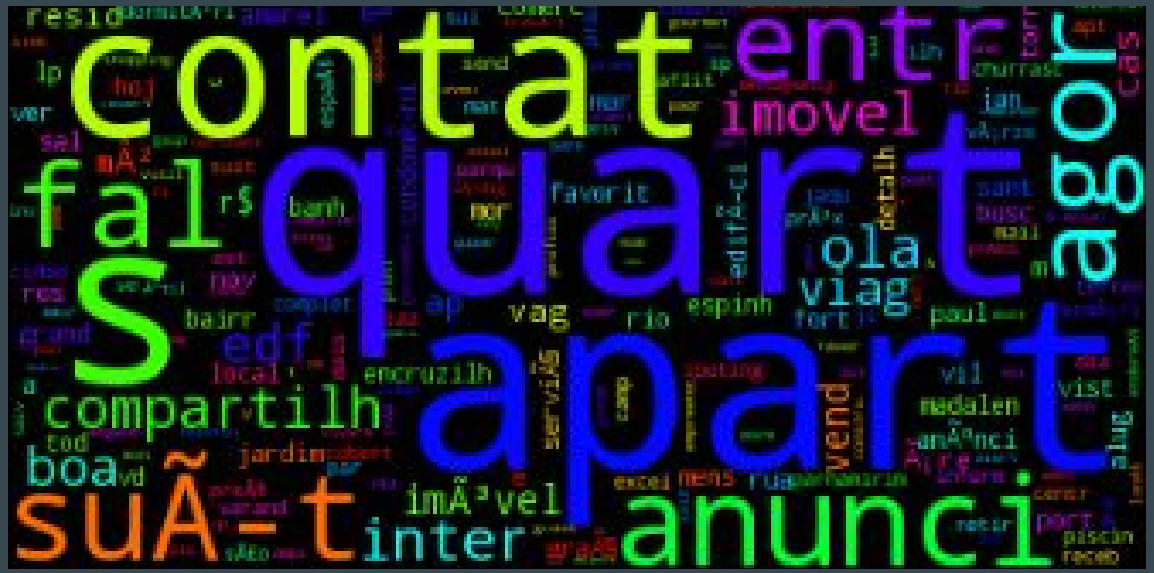


Filtro em caracteres especiais e stop words + Palavras minúsculas - (20437 palavras)

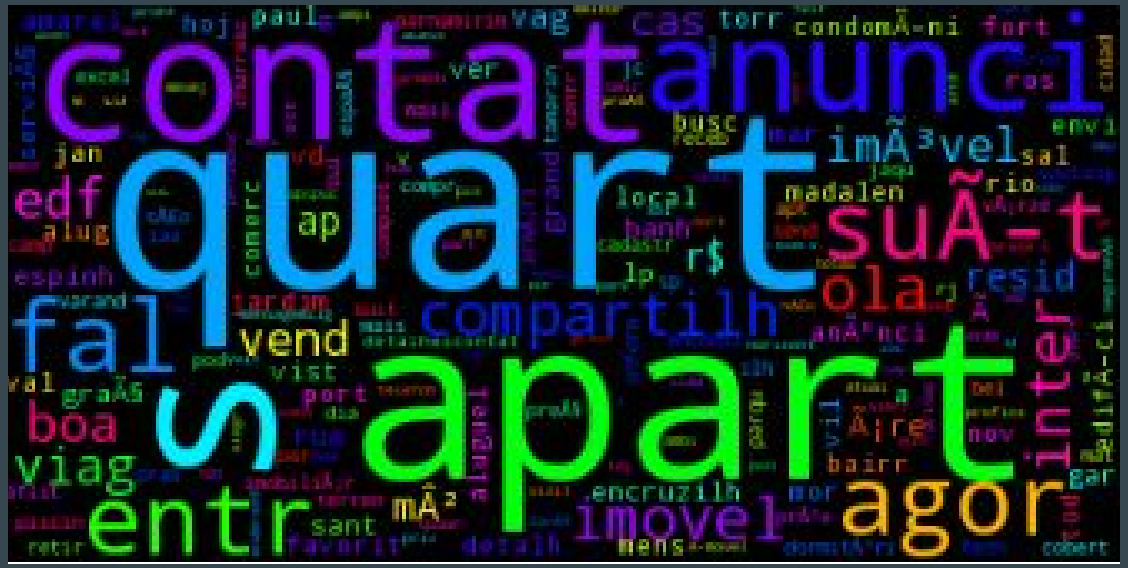
Filtro em caracteres especiais, números e stop words + Palavras minúsculas + Stemming (8658 palavras)



Filtro em caracteres especiais, números e stop words + Palavras minúsculas + Stemming (1000 melhores palavras)



Filtro em caracteres especiais, números e stop words + Palavras minúsculas + Stemming (1000 palavras mais frequentes)



2.3 Treinar Classificadores

Cross-Validation: 10 folders

Grid Search

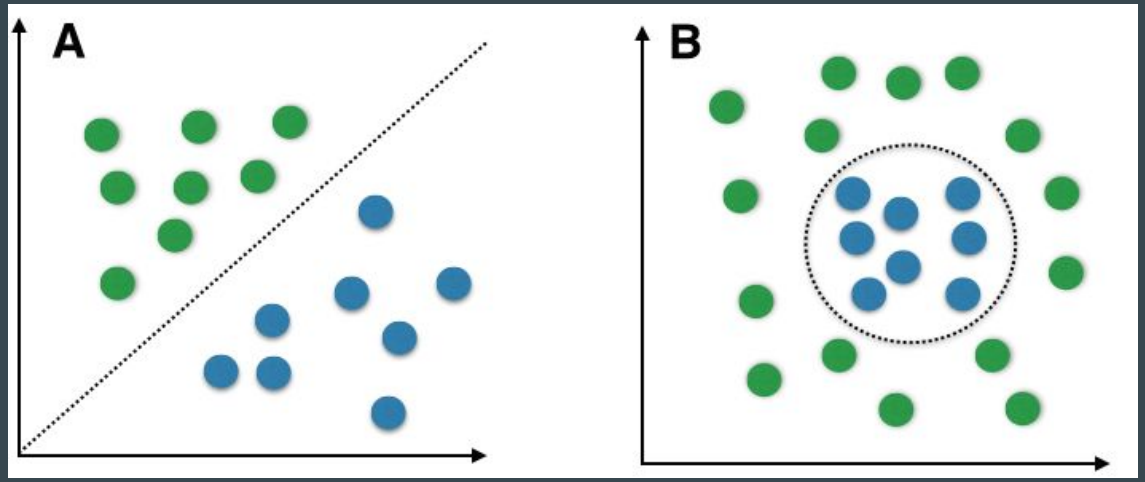
Accuracy

Precision

Recall

F-Measure

Training Time



Logistic Regression

	Accuracy	Precision	Recall	F-Measure	Training Time (seconds)
bw1	0.56 (+/- 0.31)	0.53 (+/- 0.41)	0.77 (+/- 0.58)	0.6278461538	0.192956
bw2	0.52 (+/- 0.38)	0.42 (+/- 0.50)	0.63 (+/- 0.81)	0.504	0.0704541
bw3	0.60 (+/- 0.33)	0.49 (+/- 0.52)	0.77 (+/- 0.78)	0.5988888889	0.0612788
bw4	0.57 (+/- 0.27)	0.52 (+/- 0.45)	0.65 (+/- 0.68)	0.5777777778	0.13944
bw5	0.59 (+/- 0.21)	0.53 (+/- 0.40)	0.66 (+/- 0.61)	0.5878991597	0.022089
bw6	0.56 (+/- 0.30)	0.49 (+/- 0.47)	0.64 (+/- 0.71)	0.5550442478	0.027739

Naive Bayes

	Accuracy	Precision	Recall	F-Measure	Training Time (seconds)
bw1	0.58 (+/- 0.34)	0.41 (+/- 0.65)	0.47 (+/- 0.87)	0.4379545455	0.0606251
bw2	0.55 (+/- 0.23)	0.36 (+/- 0.63)	0.44 (+/- 0.81)	0.396	0.047797
bw3	0.56 (+/- 0.32)	0.36 (+/- 0.66)	0.44 (+/- 0.80)	0.396	0.061635
bw4	0.55 (+/- 0.19)	0.38 (+/- 0.65)	0.36 (+/- 0.70)	0.3697297297	0.0277941
bw5	0.66 (+/- 0.31)	0.59 (+/- 0.67)	0.59 (+/- 0.64)	0.59	0.0030601
bw6	0.57 (+/- 0.25)	0.43 (+/- 0.76)	0.38 (+/- 0.72)	0.4034567901	0.00308299

Support Vector Machine

	Accuracy	Precision	Recall	F-Measure	Training Time (seconds)
bw1	0.66 (+/- 0.28)	0.67 (+/- 0.36)	0.89 (+/- 0.36)	0.7644871795	0.789962
bw2	0.62 (+/- 0.24)	0.57 (+/- 0.45)	0.79 (+/- 0.70)	0.6622058824	0.589598
bw3	0.70 (+/- 0.31)	0.68 (+/- 0.32)	0.95 (+/- 0.20)	0.7926380368	0.72193
bw4	0.64 (+/- 0.31)	0.59 (+/- 0.48)	0.80 (+/- 0.68)	0.6791366906	0.316355
bw5	0.48 (+/- 0.08)	0.00 (+/- 0.00)	0.00 (+/- 0.00)	0	0.037677
bw6	0.49 (+/- 0.03)	0.00 (+/- 0.00)	0.00 (+/- 0.00)	0	0.038074

Multilayer Perceptron

	Accuracy	Precision	Recall	F-Measure	Training Time (seconds)
bw1	0.62 (+/- 0.30)	0.53 (+/- 0.58)	0.70 (+/- 0.76)	0.6032520325	1.0273
bw2	0.52 (+/- 0.16)	0.41 (+/- 0.43)	0.77 (+/- 0.79)	0.5350847458	1.56981
bw3	0.73 (+/- 0.37)	0.64 (+/- 0.72)	0.67 (+/- 0.75)	0.6546564885	2.9102
bw4	0.58 (+/- 0.28)	0.57 (+/- 0.54)	0.72 (+/- 0.65)	0.6362790698	1.63997
bw5	0.61 (+/- 0.20)	0.55 (+/- 0.61)	0.56 (+/- 0.78)	0.554954955	0.33602
bw6	0.64 (+/- 0.35)	0.63 (+/- 0.72)	0.55 (+/- 0.79)	0.5872881356	0.294669

Random Forest

	Accuracy	Precision	Recall	F-Measure	Training Time (seconds)
bw1	0.89 (+/- 0.28)	0.90 (+/- 0.60)	0.78 (+/- 0.56)	0.8357142857	2.79431
bw2	0.89 (+/- 0.24)	1.00 (+/- 0.00)	0.78 (+/- 0.49)	0.8764044944	2.44361
bw3	0.89 (+/- 0.29)	0.89 (+/- 0.60)	0.78 (+/- 0.59)	0.8313772455	2.62891
bw4	0.89 (+/- 0.30)	0.99 (+/- 0.05)	0.79 (+/- 0.62)	0.8787640449	2.14304
bw5	0.89 (+/- 0.30)	0.87 (+/- 0.61)	0.83 (+/- 0.61)	0.8495294118	2.01905
bw6	0.87 (+/- 0.31)	0.87 (+/- 0.59)	0.77 (+/- 0.66)	0.8169512195	1.79734

Ensemble

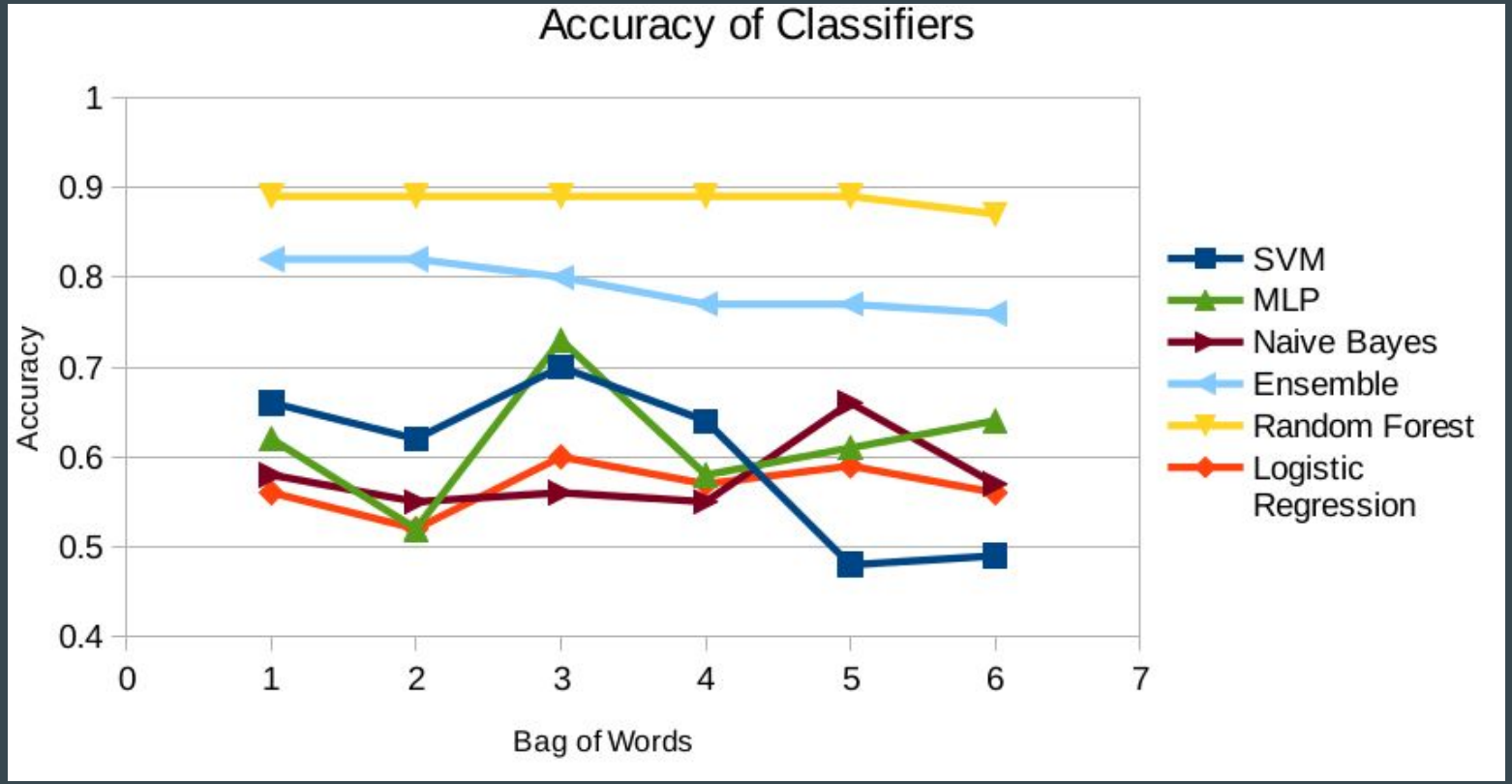
	Accuracy	Precision	Recall	F-Measure	Training Time (seconds)
bw1	0.82 (+/- 0.35)	0.80 (+/- 0.80)	0.64 (+/- 0.69)	0.7111111111	3.66417
bw2	0.82 (+/- 0.37)	0.80 (+/- 0.80)	0.64 (+/- 0.73)	0.7111111111	4.03468
bw3	0.80 (+/- 0.34)	0.79 (+/- 0.79)	0.61 (+/- 0.70)	0.6884285714	5.42888
bw4	0.77 (+/- 0.34)	0.90 (+/- 0.60)	0.54 (+/- 0.68)	0.675	3.82493
bw5	0.77 (+/- 0.38)	0.80 (+/- 0.80)	0.53 (+/- 0.75)	0.637593985	2.2365
bw6	0.76 (+/- 0.37)	0.79 (+/- 0.79)	0.53 (+/- 0.75)	0.6343939394	2.08828

2.4 Comparar Estratégias

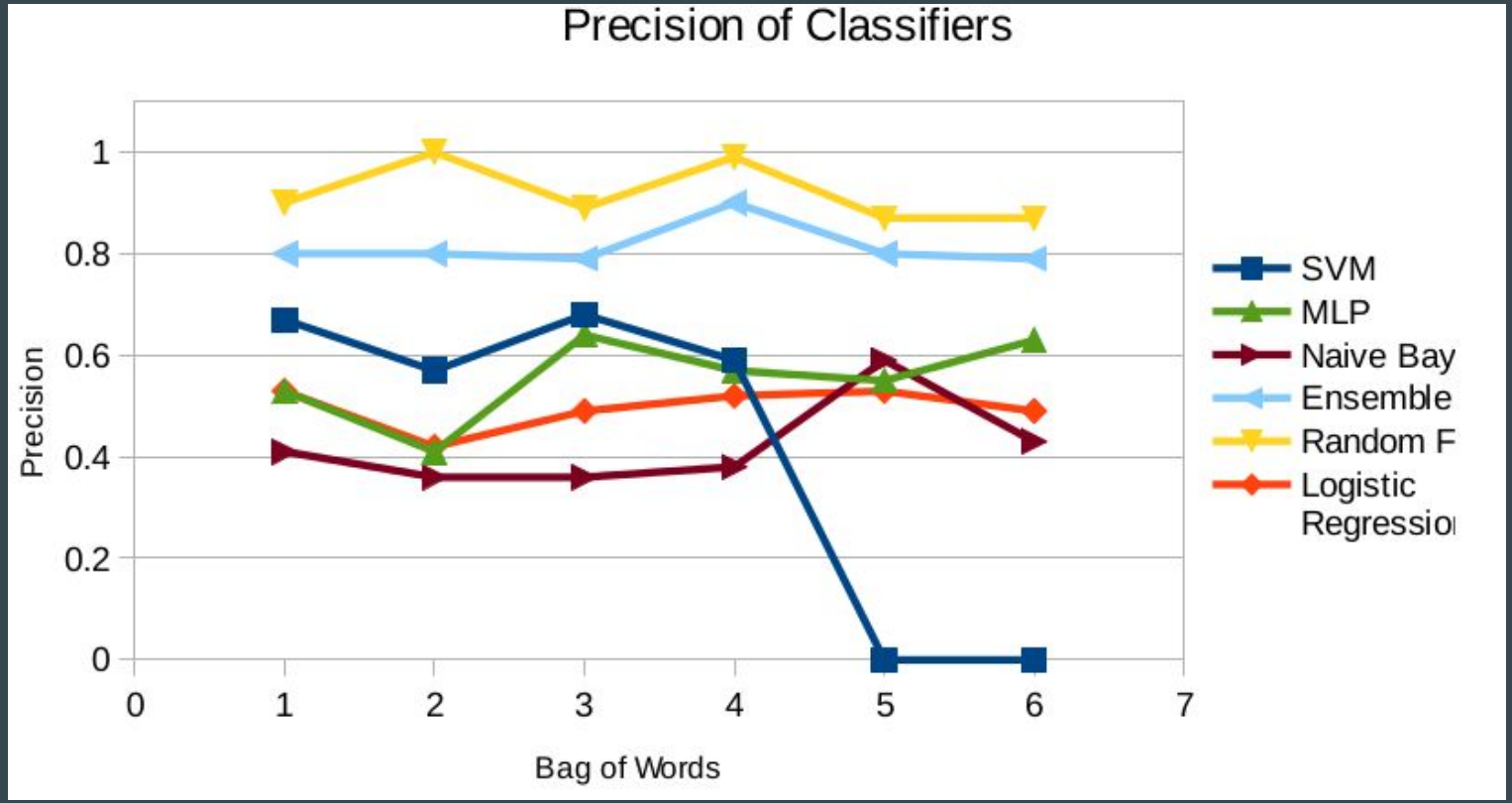
Critérios:

1. Accuracy
2. Precision [tp / tp + fp]
3. Recall [tp / tp + fn]
4. F-measure (2 * (Recall * Precision) / (Recall + Precision)]
5. Training Time

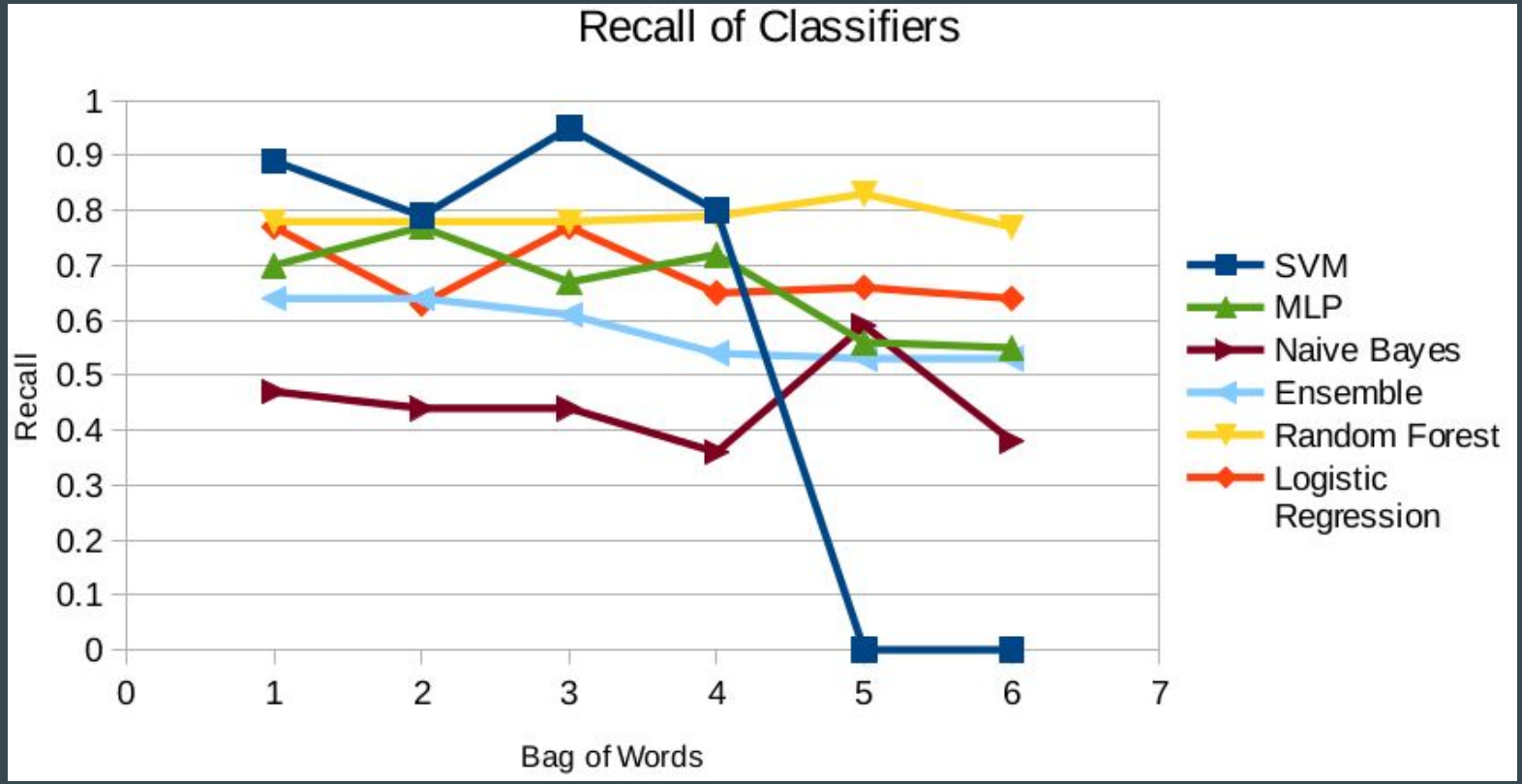
2.4.1 Accuracy



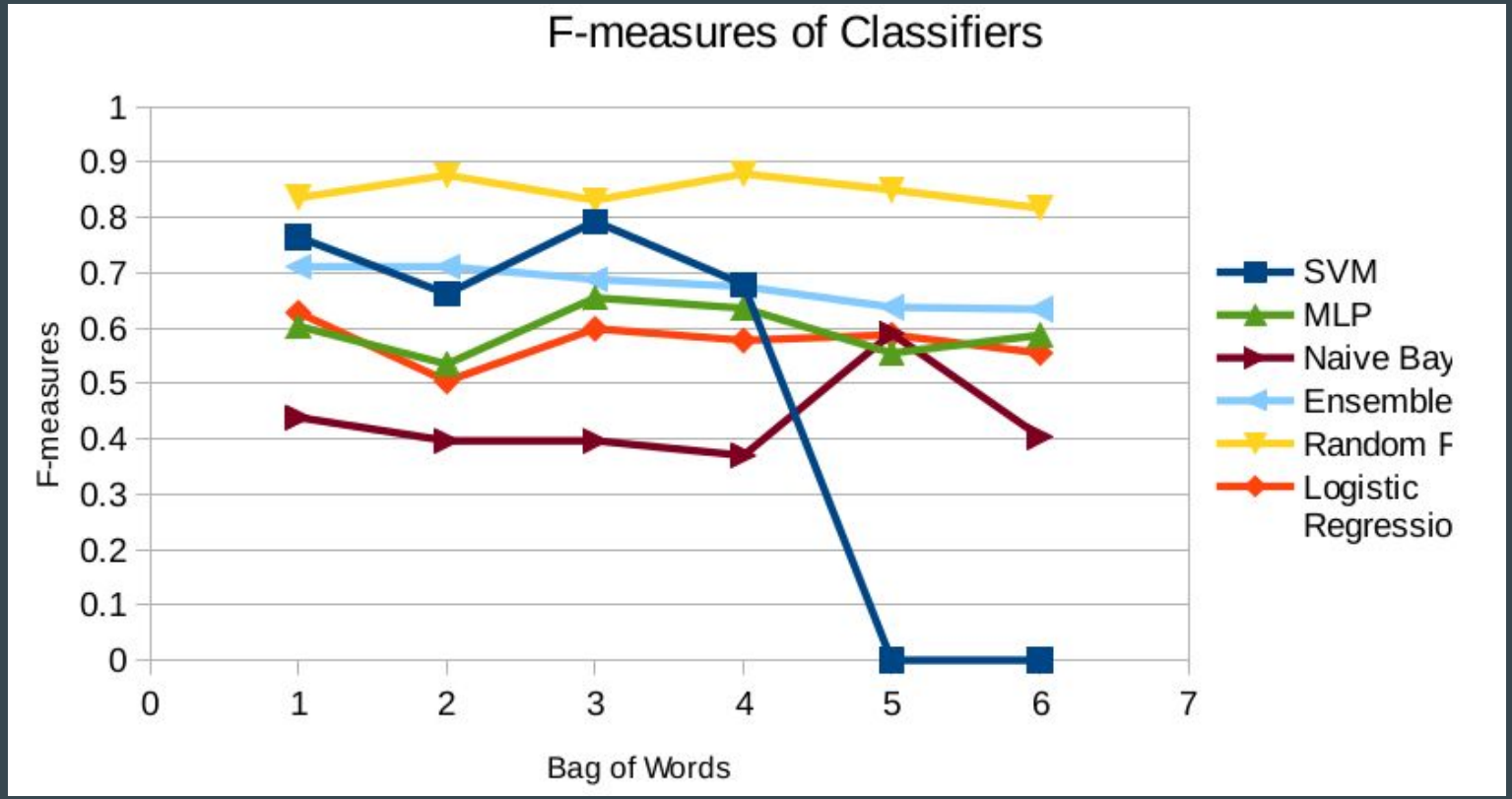
2.4.2 Precision



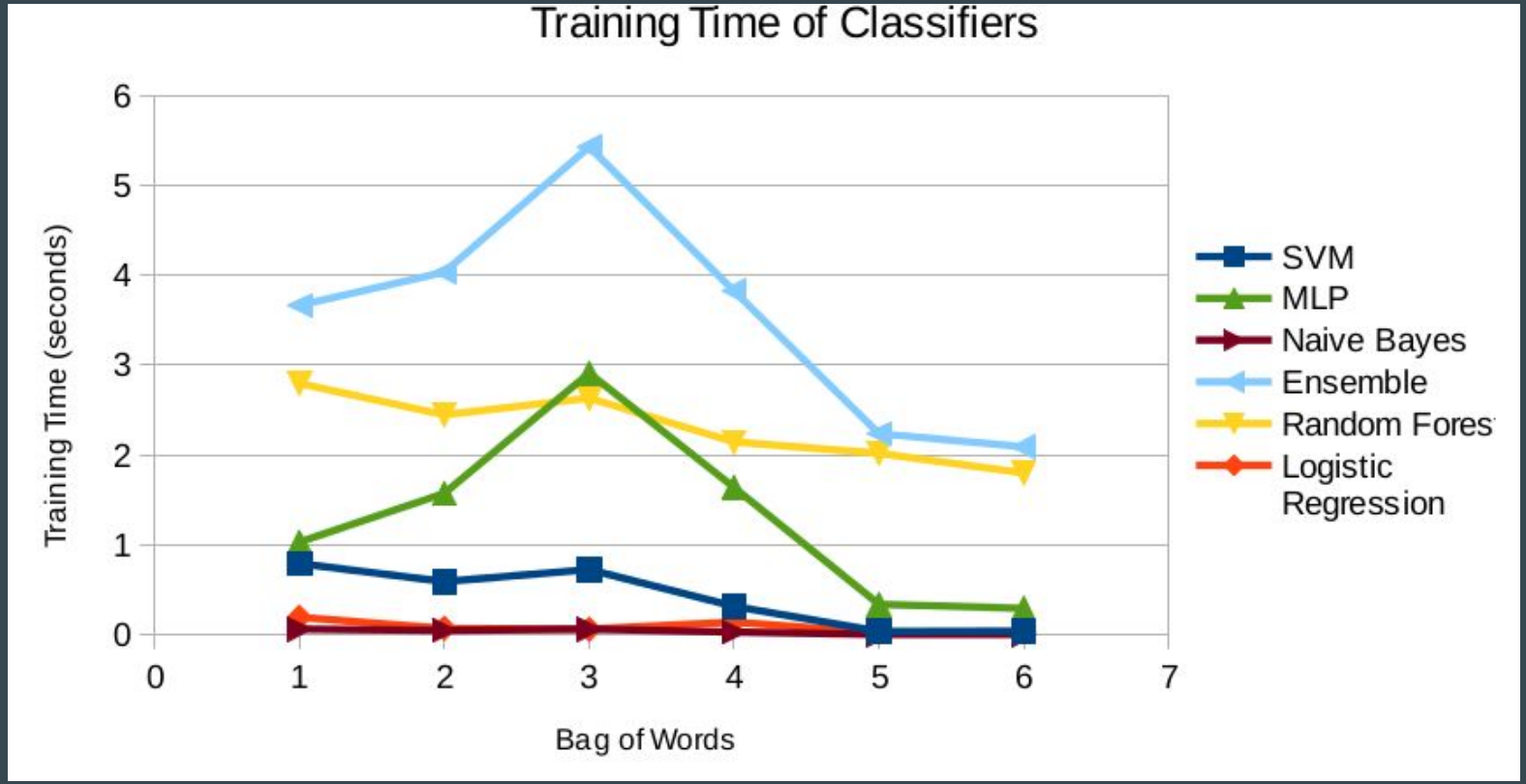
2.4.3 Recall


Recall of Classifiers
Recall
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
1
2
3
4
5
6
7
Bag of Words
SVM
MLP
Naive Bayes
Ensemble
Random Forest
Logistic Regression

2.4.4 F-measure



2.4.5 Training Time

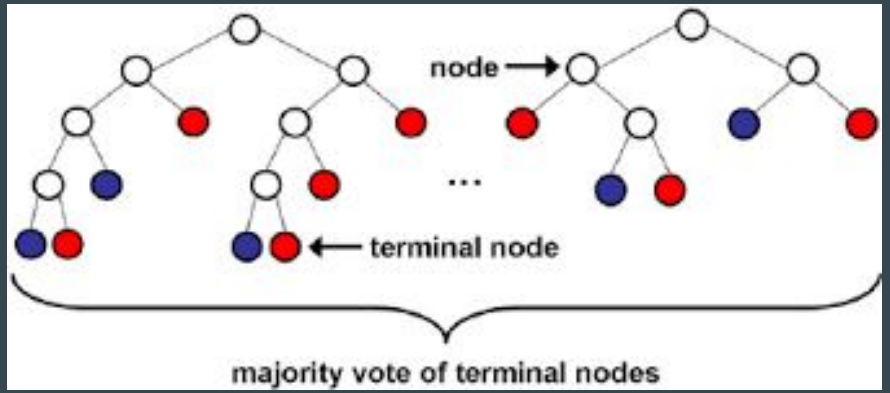


2.5 Conclusão

Random Forest

Filtro em caracteres especiais, números e stop words

+ Palavras minúsculas + Stemming (8658 palavras)



Parte 3 : Extrair Instâncias com seus Valores e Atributos

Wrapper - Tarefas

1. Criar um wrapper para cada site
2. Criar uma solução que funcione em todos os sites do domínio
3. Comparar estratégias

Abordagens

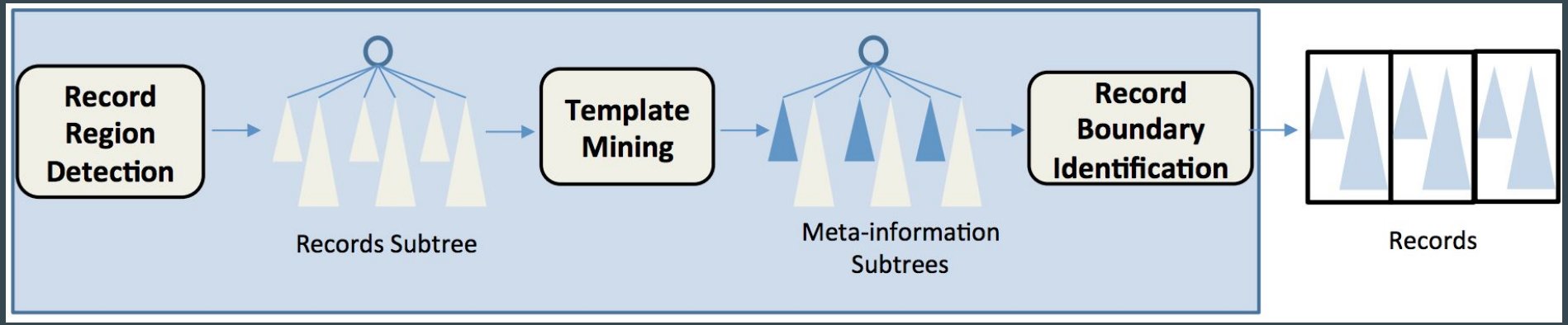
- Simples:
  - Construir um wrapper para cada site
    - Pouca supervisão
    - Independente
    - Estrutura varia
- Composta:
  - Construir um único wrapper para o domínio
    - Dois passos

Detecção da Região do Registro - Simples


```
▼<ul class="unstyled container">
  ::before
  ▼<li>
    ▼<h3> == $0
        "2 "
        <span class="text-info icone-quartos">quartos</span>
      </h3>
    </li>
  ▼<li>
    ▼<h3>
        "1 "
        <span class="text-info icone-vagas">vaga</span>
      </h3>
    </li>
  ▶<li>…</li>
```

Detecção da Região do Registro - Composta

- Encontrar sub-árvore contendo os registros
- Extrair dados dos nós filhos



Extração

- A partir de uma base de 10 sites de cada domínio já classificados anteriormente obteve os seguintes resultados usando a abordagem simples:

Resultados

	Zapimoveis	Expoimoveis	OLX	Directimoveis	Imovelavenda
N	80	100	80	100	100
E	53	67	48	85	120
C	53	66	48	84	107
Recall(C/N)	0.6625	0.66	0.6	0.84	1.07
Precision(C/E)	1.0	0.98	1.0	0.98	0.89

Resultados

	Imovelweb	Lugarcerto	Redeimoveispe	Teuimovel	Vivareal
N	100	70	80	100	100
E	92	60	59	66	78
C	86	60	59	66	78
Recall(C/N)	0.86	0.857	0.737	0.66	0.78
Precision(C/E)	0.93	1.0	1.0	1.0	1.0

Extração

- A partir de uma base de 10 sites de cada domínio já classificados anteriormente obteve os seguintes resultados usando a abordagem composta:

Resultados

	Regex_extractor
N	950
E	641
C	587
Recall(C/N)	0.62
Precision(C/E)	0.91