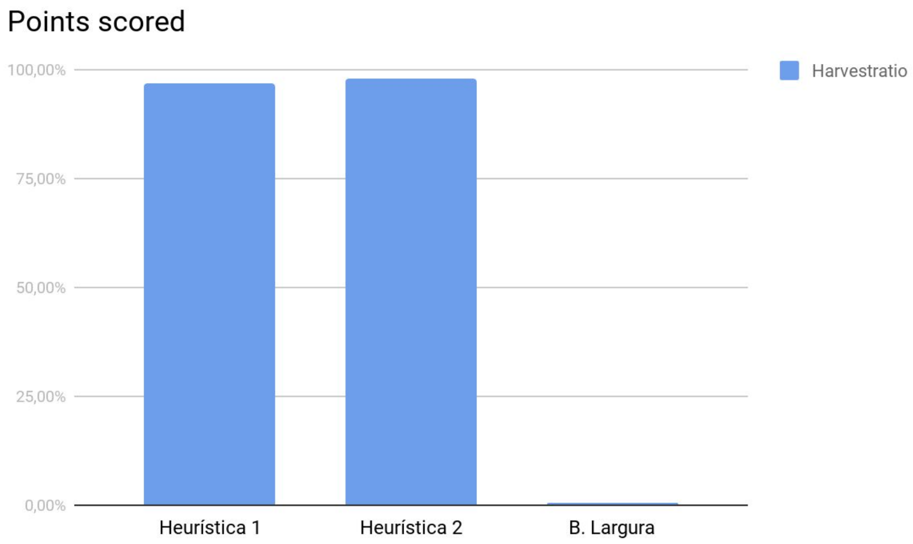
RI

Crawler

- Busca em largura
- Heurística
    1. (Pos_url*1 + Pos_ancora*2)-(Neg_url*1+Neg_ancora*5) >= 0
    2. Pos_url*1 + Pos_ancora*2 > 0

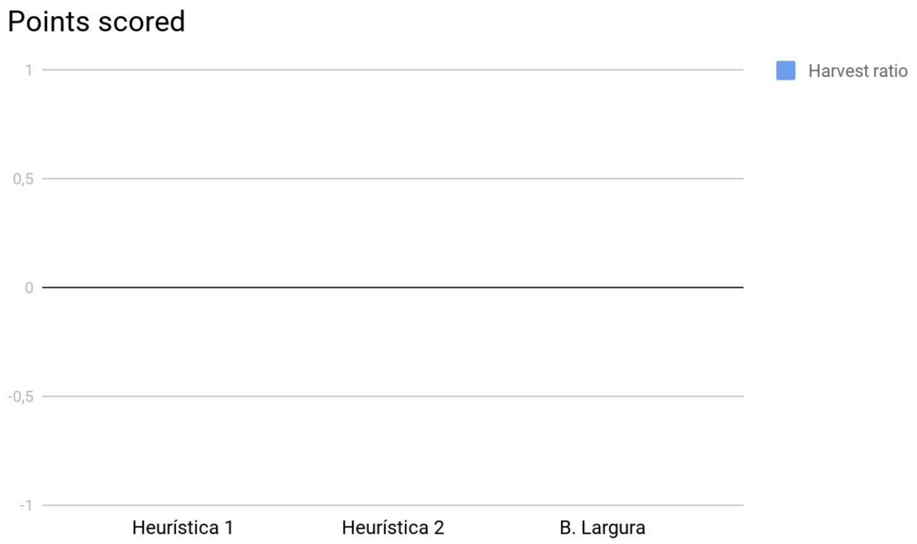
Crawler

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



www.extra.com.br

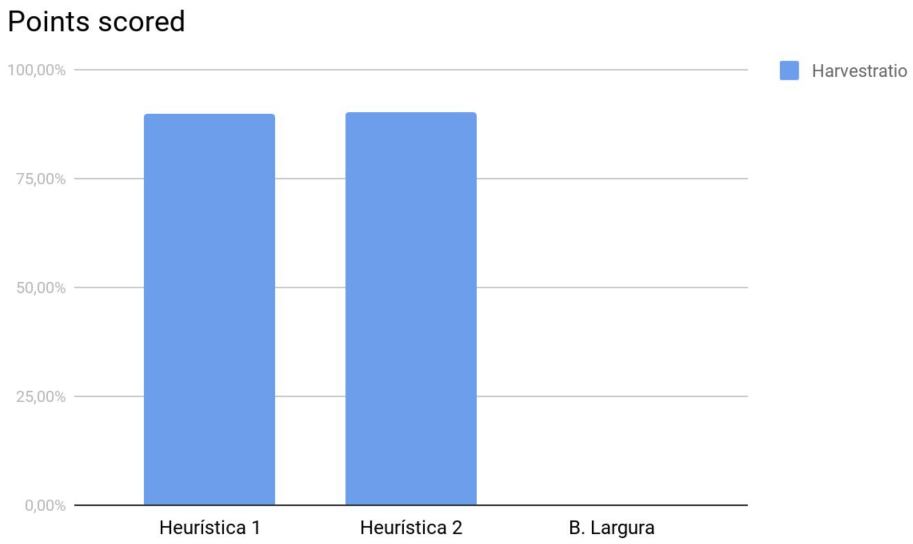
- Heurística 1 : 187 pages, 181 positivas
- Heurística 2 : 191 pages, 187 positivas
- Busca em Largura : 238 pages, 16 positivas



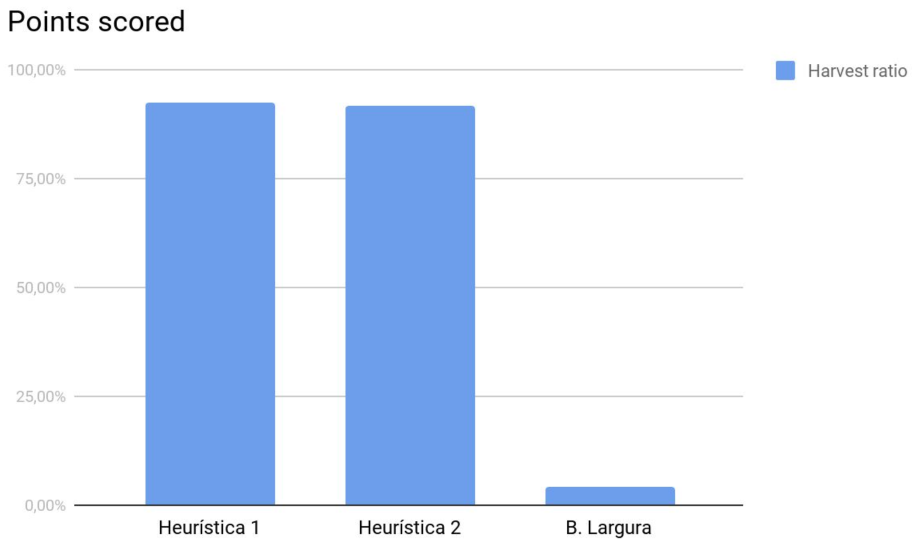
www.carrefour.com.br

- Heurística 1 : 591 pages, 0 positivas
- Heurística 2 : 0 pages, 0 positivas
- Busca em Largura : 258 pages, 0 positivas

www.pontofrio.com.br

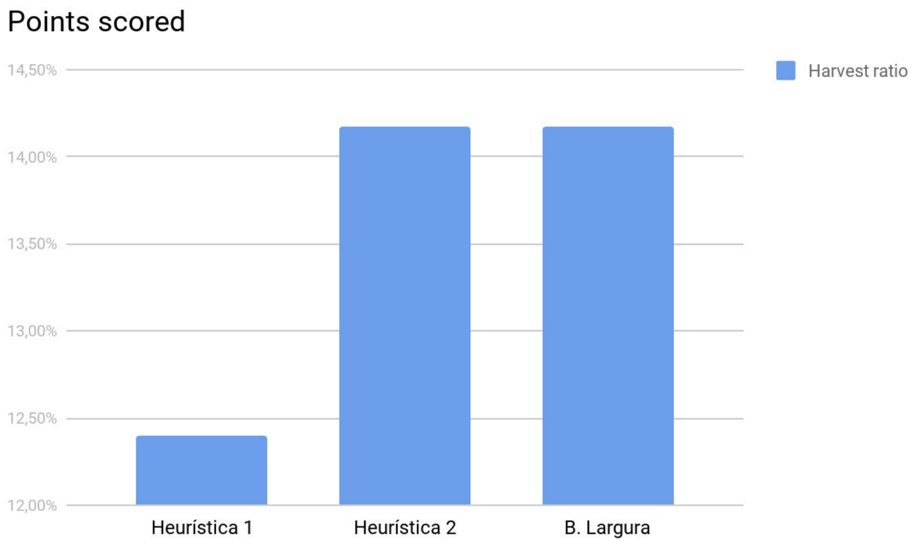


- Heurística 1 : 20 pages, 18 positivas
- Heurística 2 : 21 pages, 19 positivas
- Busca em Largura : 117 pages, 0 positivas



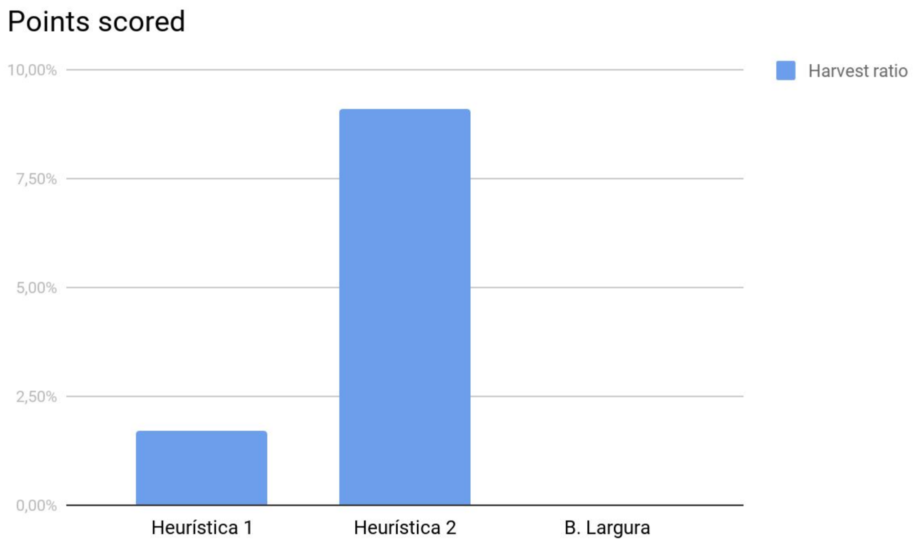
www.casasbahia.com.br

- Heurística 1 : 40 pages, 37 positivas
- Heurística 2 : 120 pages, 110 positivas
- Busca em Largura : 23 pages, 1 positivas



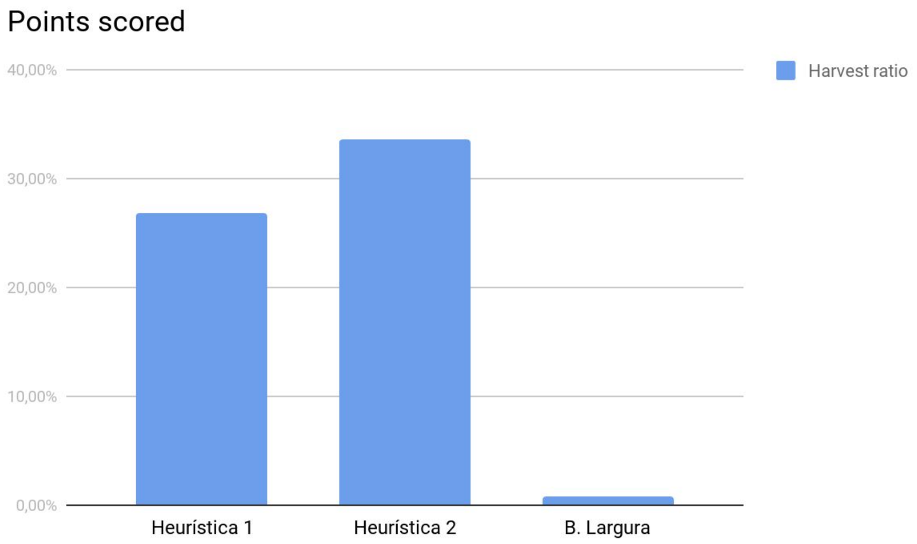
www.dell.com/br

- Heurística 1 : 153 pages, 19 positivas
- Heurística 2 : 134 pages, 19 positivas
- Busca em Largura : 134 pages, 19 positivas



www.submarino.com.br

- Heurística 1 : 60 pages, 1 positivas
- Heurística 2 : 11 pages, 1 positivas
- Busca em Largura : 131 pages, 0 positivas



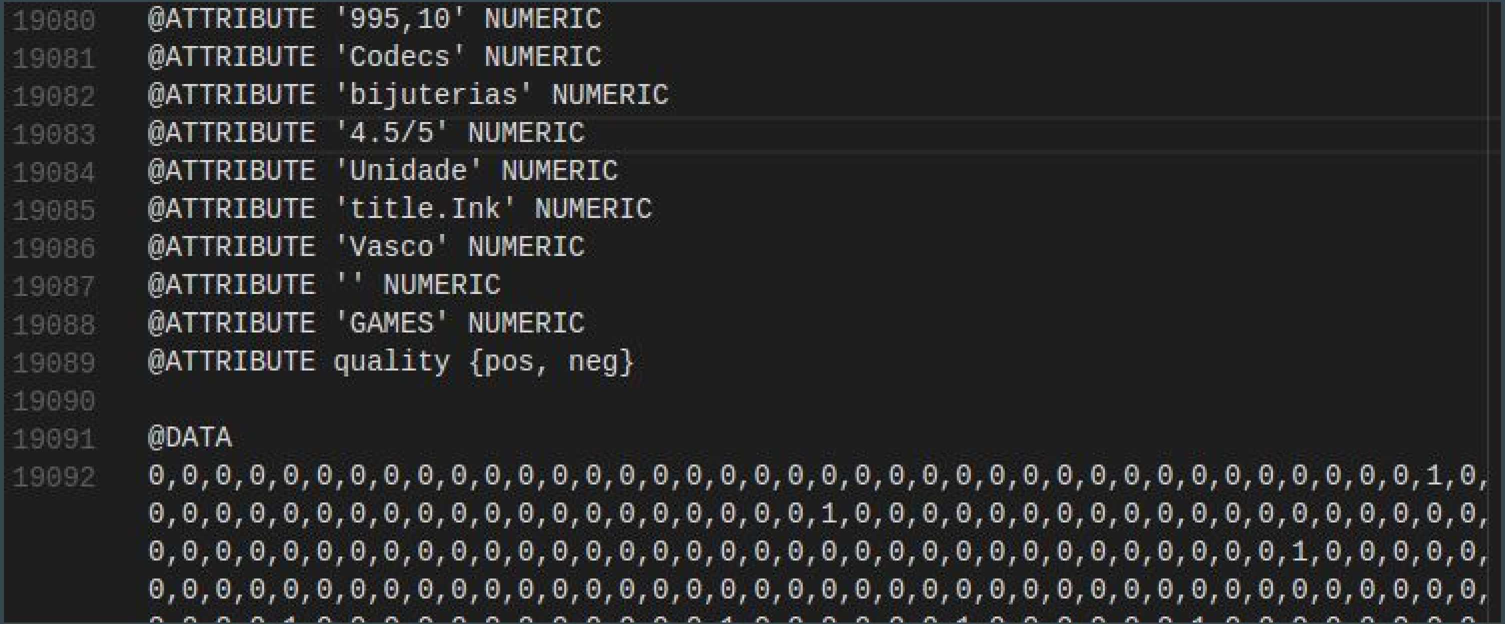
www.lojahp.com.br

- Heurística 1 : 213 pages, 57 positivas
- Heurística 2 : 220 pages, 74 positivas
- Busca em Largura : 254 pages, 2 positivas

americanas, magazineluiza e kabum: nem um dos crawlers foi capaz de pegar qualquer página, devido aos sites serem gerados por JavaScript✱

Classificador

Bag of words das 200 páginas

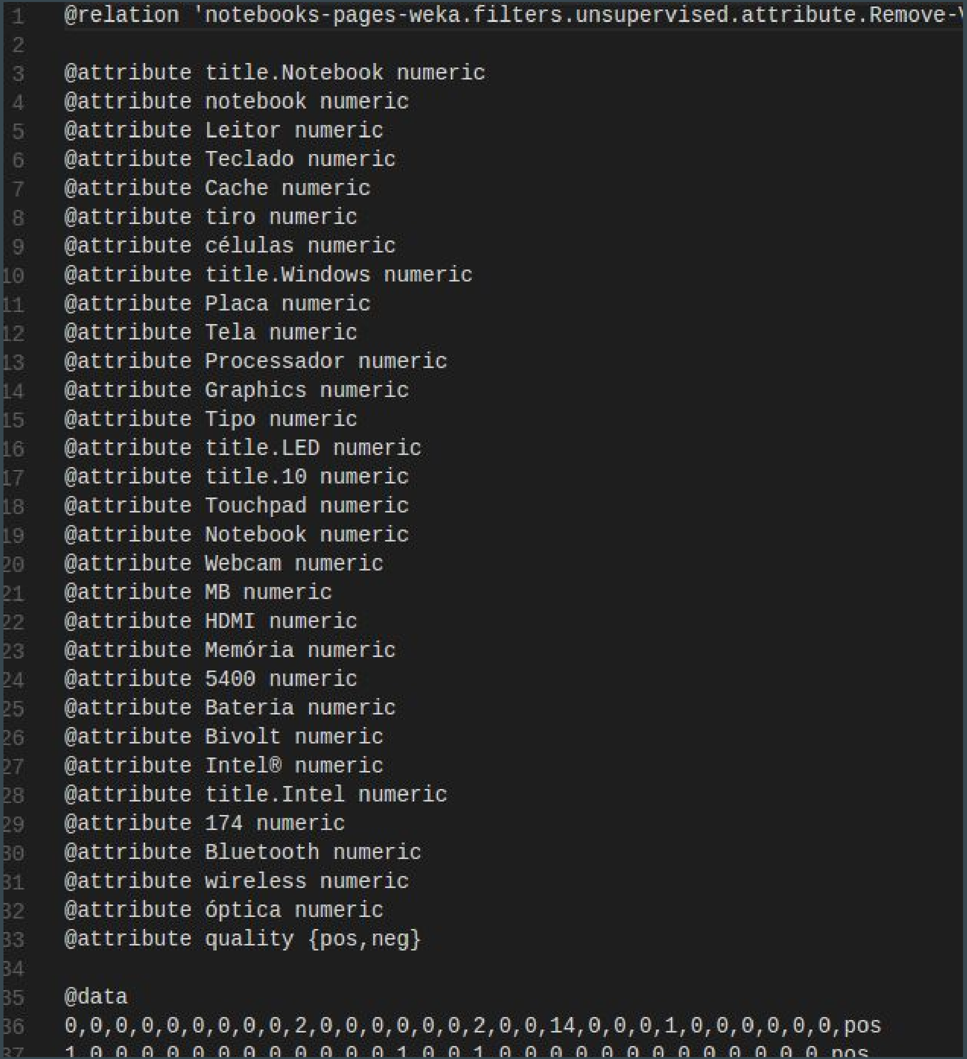


```
@ATTRIBUTE '995,10' NUMERIC
@ATTRIBUTE 'Codecs' NUMERIC
@ATTRIBUTE 'bijuterias' NUMERIC
@ATTRIBUTE '4.5/5' NUMERIC
@ATTRIBUTE 'Unidade' NUMERIC
@ATTRIBUTE 'title.Ink' NUMERIC
@ATTRIBUTE 'Vasco' NUMERIC
@ATTRIBUTE '' NUMERIC
@ATTRIBUTE 'GAMES' NUMERIC
@ATTRIBUTE quality {pos, neg}

@DATA
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,
```

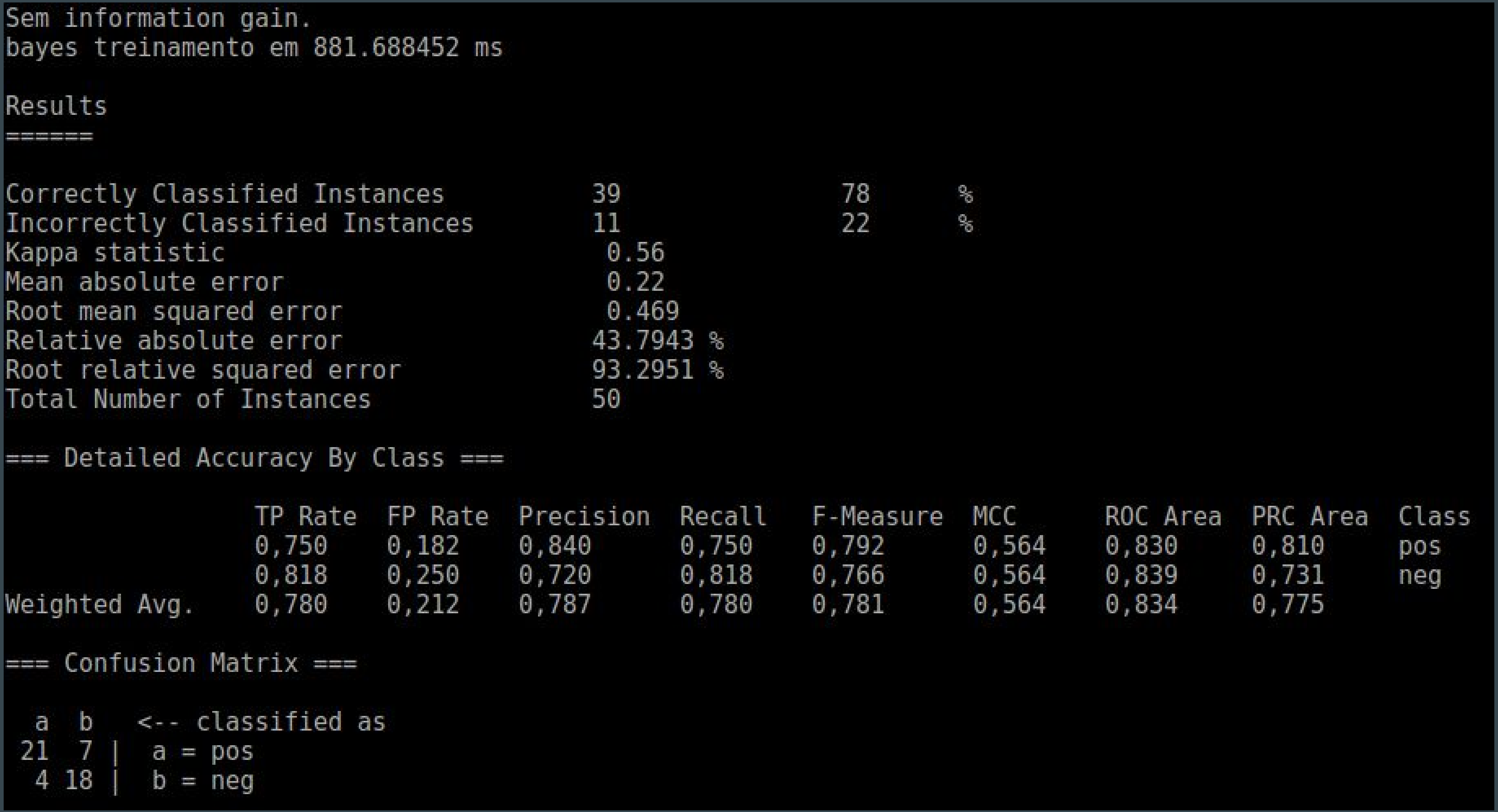
Classificador

Information Gain (30)

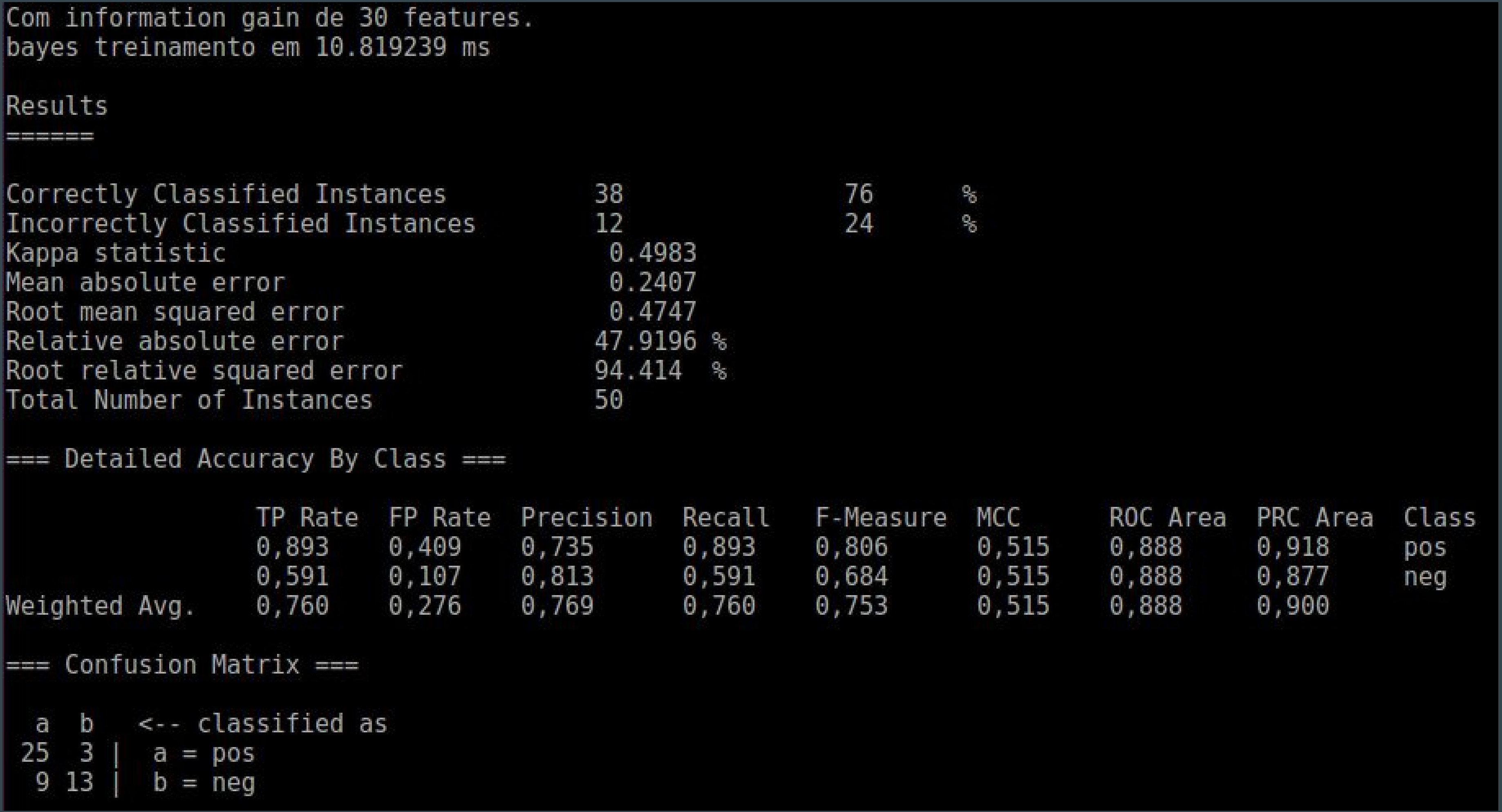
Stopwords não tem efeito



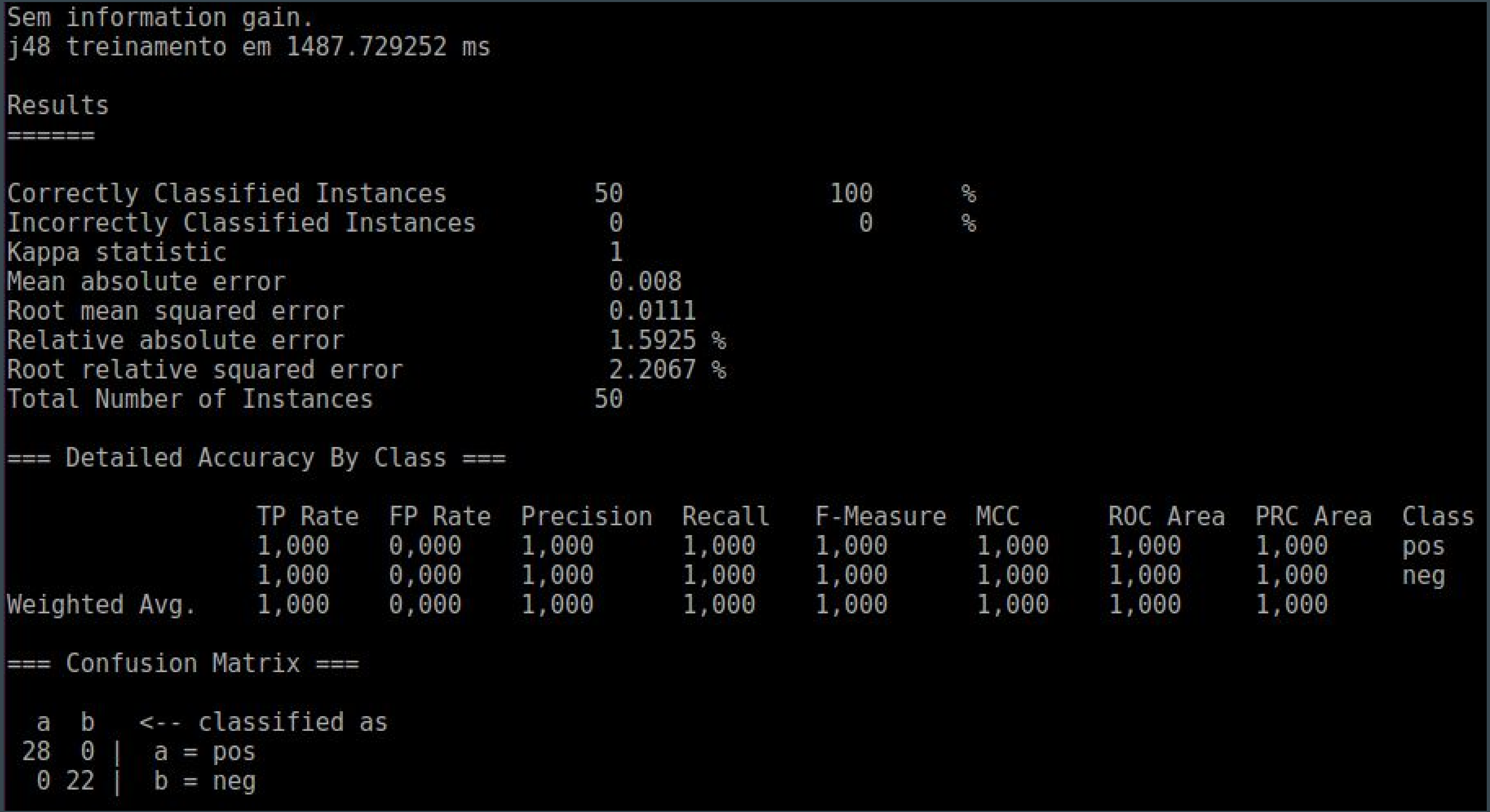
```
@relation 'notebooks-pages-weka.filters.unsupervised.attribute.Remove-

@attribute title.Notebook numeric
@attribute notebook numeric
@attribute Leitor numeric
@attribute Teclado numeric
@attribute Cache numeric
@attribute tiro numeric
@attribute células numeric
@attribute title.Windows numeric
@attribute Placa numeric
@attribute Tela numeric
@attribute Processador numeric
@attribute Graphics numeric
@attribute Tipo numeric
@attribute title.LED numeric
@attribute title.10 numeric
@attribute Touchpad numeric
@attribute Notebook numeric
@attribute Webcam numeric
@attribute MB numeric
@attribute HDMI numeric
@attribute Memória numeric
@attribute 5400 numeric
@attribute Bateria numeric
@attribute Bivolt numeric
@attribute Intel® numeric
@attribute title.Intel numeric
@attribute 174 numeric
@attribute Bluetooth numeric
@attribute wireless numeric
@attribute óptica numeric
@attribute quality {pos,neg}

@data
0,0,0,0,0,0,0,0,0,2,0,0,0,0,0,0,2,0,0,14,0,0,0,1,0,0,0,0,0,0,pos
1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,pos
```



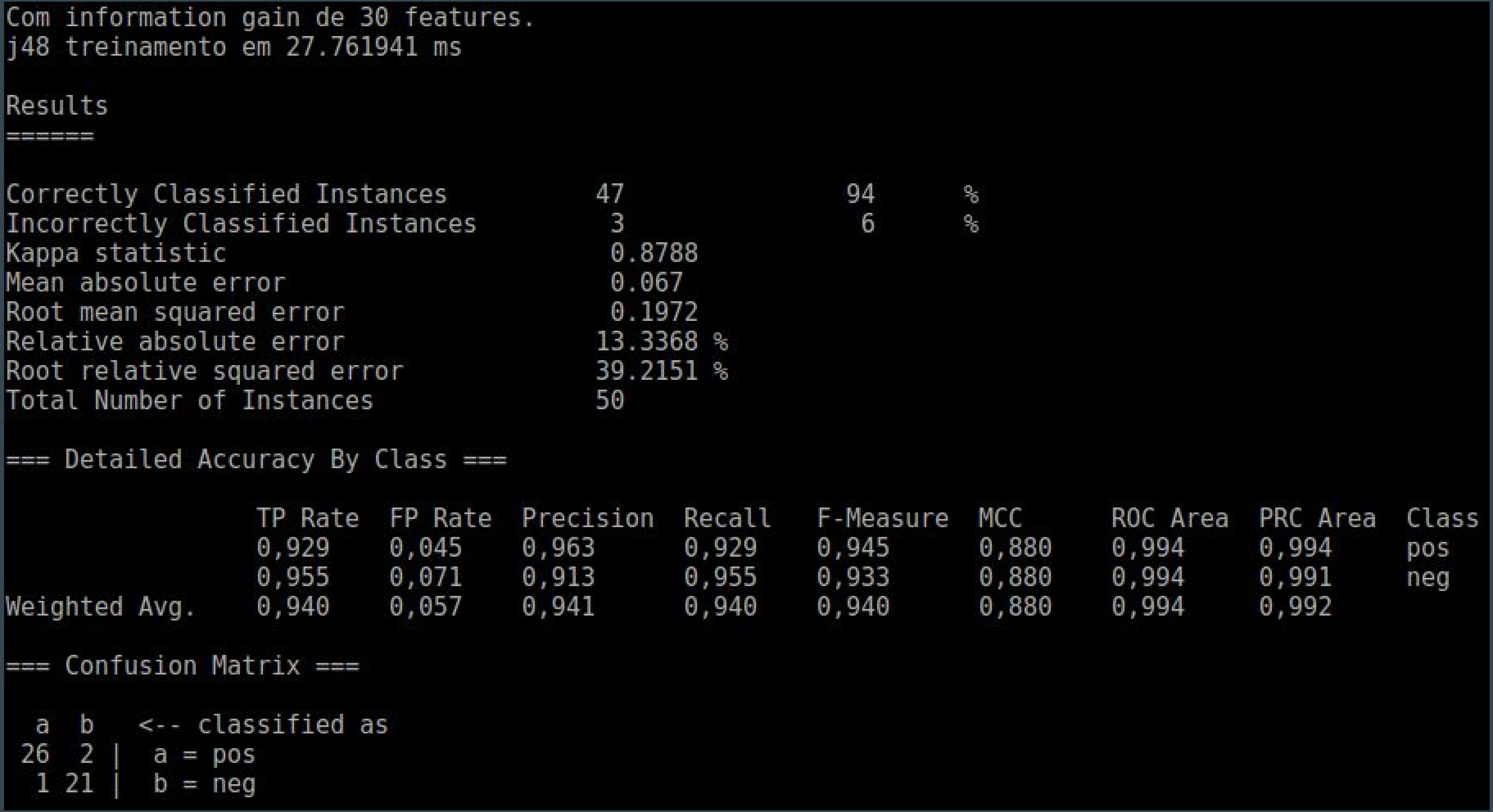
```
Sem information gain.
bayes treinamento em 881.688452 ms

Results
======

Correctly Classified Instances          39               78      %
Incorrectly Classified Instances        11               22      %
Kappa statistic                          0.56
Mean absolute error                      0.22
Root mean squared error                  0.469
Relative absolute error                 43.7943 %
Root relative squared error             93.2951 %
Total Number of Instances               50

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,750    0,182    0,840      0,750    0,792      0,564    0,830     0,810     pos
                 0,818    0,250    0,720      0,818    0,766      0,564    0,839     0,731     neg
Weighted Avg.    0,780    0,212    0,787      0,780    0,781      0,564    0,834     0,775

=== Confusion Matrix ===

  a  b   <-- classified as
 21  7 |  a = pos
  4 18 |  b = neg
```



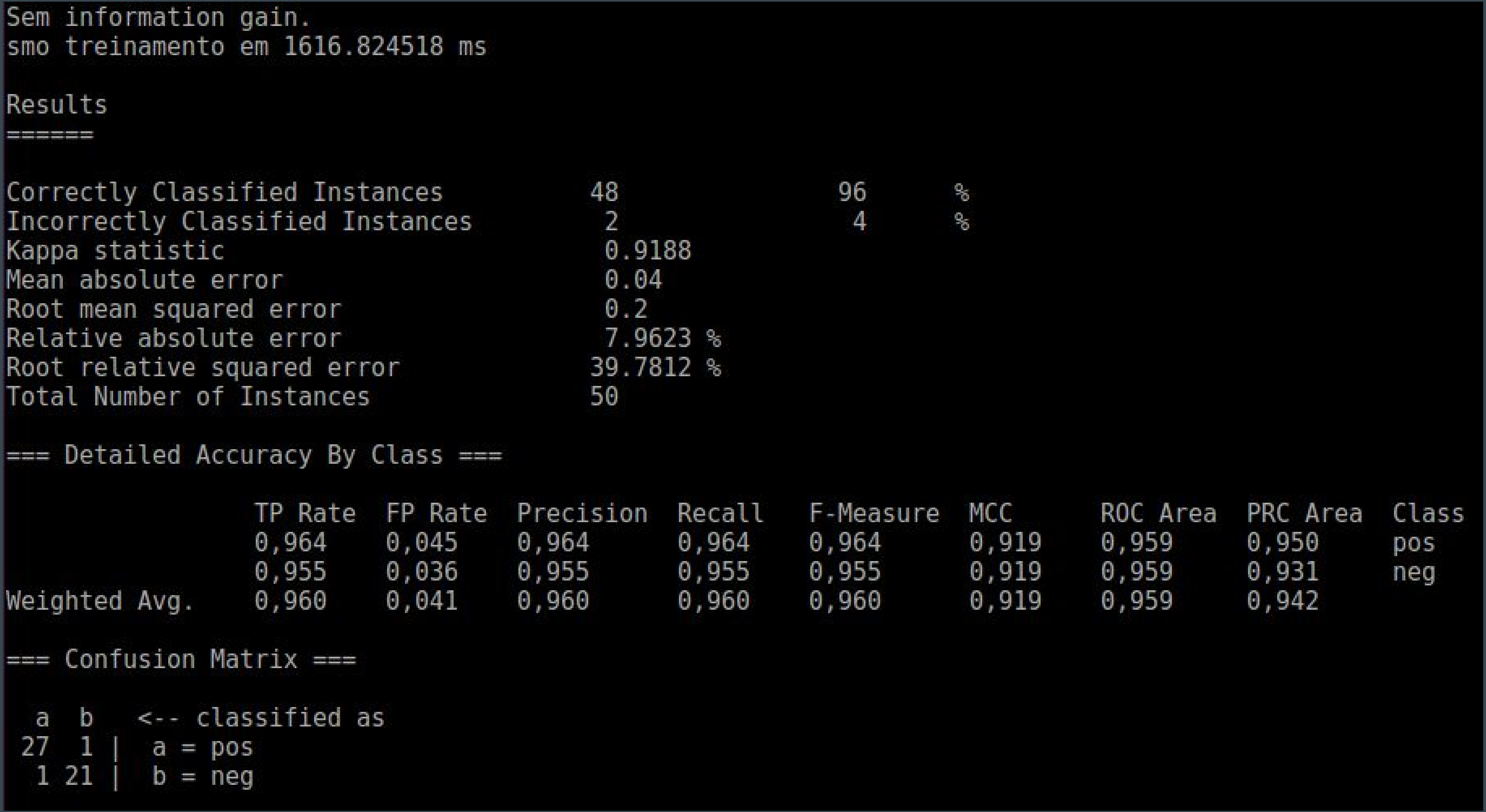
```
Com information gain de 30 features.
bayes treinamento em 10.819239 ms

Results
======

Correctly Classified Instances          38               76      %
Incorrectly Classified Instances        12               24      %
Kappa statistic                          0.4983
Mean absolute error                      0.2407
Root mean squared error                  0.4747
Relative absolute error                 47.9196 %
Root relative squared error             94.414  %
Total Number of Instances               50

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,893    0,409    0,735      0,893    0,806      0,515    0,888     0,918     pos
                 0,591    0,107    0,813      0,591    0,684      0,515    0,888     0,877     neg
Weighted Avg.    0,760    0,276    0,769      0,760    0,753      0,515    0,888     0,900

=== Confusion Matrix ===

  a  b   <-- classified as
 25  3 |  a = pos
  9 13 |  b = neg
```



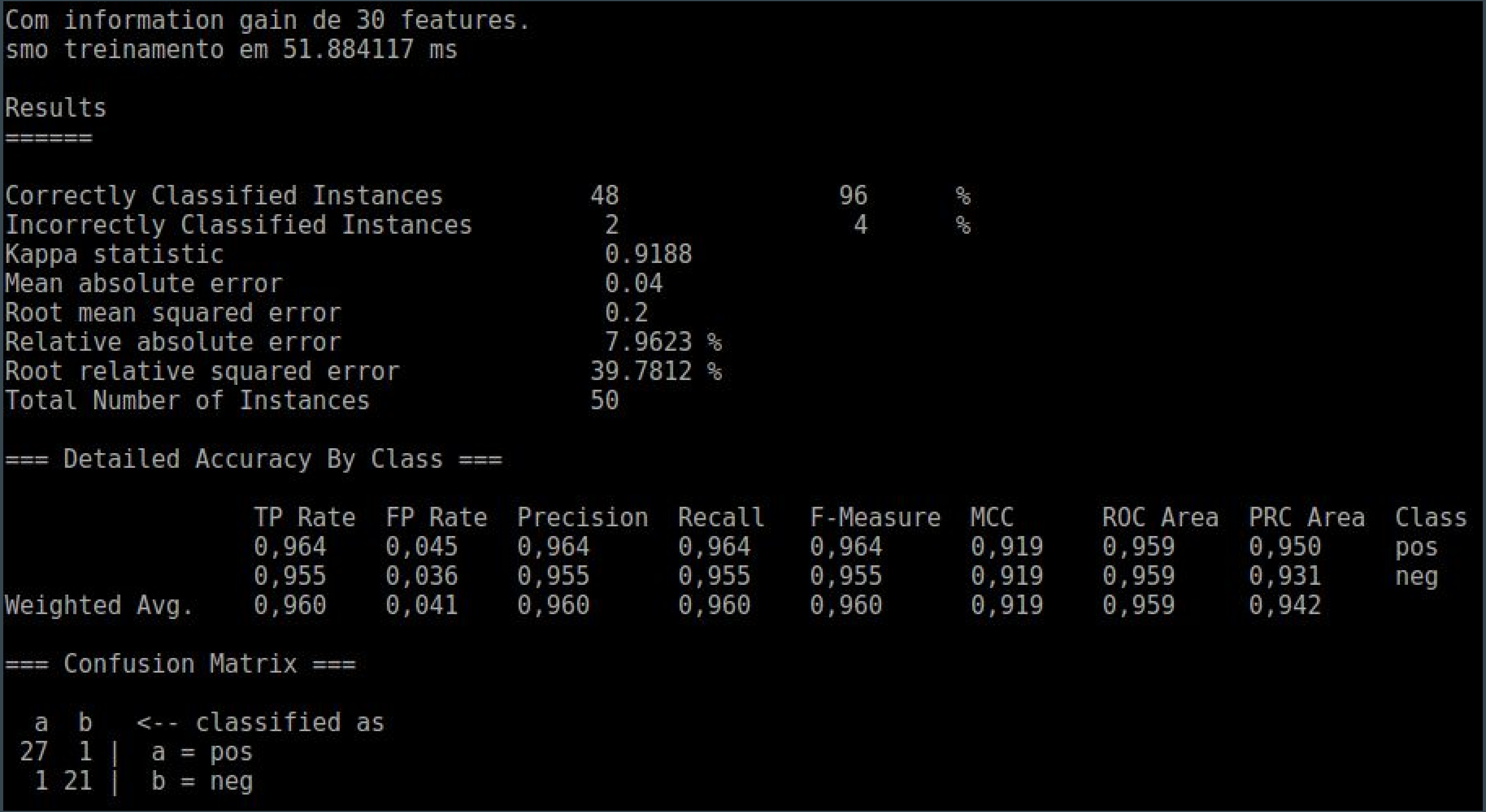
```
Sem information gain.
j48 treinamento em 1487.729252 ms

Results
=======

Correctly Classified Instances          50              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0.008
Root mean squared error                  0.0111
Relative absolute error                  1.5925 %
Root relative squared error              2.2067 %
Total Number of Instances               50

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     pos
                 1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     neg
Weighted Avg.    1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000

=== Confusion Matrix ===

  a  b   <-- classified as
 28  0 |  a = pos
  0 22 |  b = neg
```



```
Com information gain de 30 features.
j48 treinamento em 27.761941 ms

Results
=======

Correctly Classified Instances          47               94      %
Incorrectly Classified Instances         3                6      %
Kappa statistic                          0.8788
Mean absolute error                      0.067
Root mean squared error                  0.1972
Relative absolute error                 13.3368 %
Root relative squared error             39.2151 %
Total Number of Instances               50

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,929    0,045    0,963      0,929    0,945      0,880    0,994     0,994     pos
                 0,955    0,071    0,913      0,955    0,933      0,880    0,994     0,991     neg
Weighted Avg.    0,940    0,057    0,941      0,940    0,940      0,880    0,994     0,992

=== Confusion Matrix ===

  a  b   <-- classified as
 26  2 |  a = pos
  1 21 |  b = neg
```

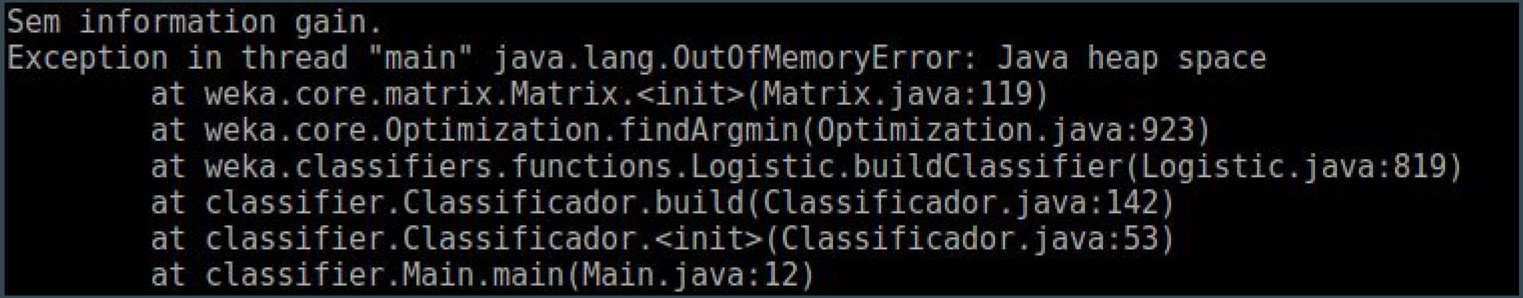


```
Sem information gain.
smo treinamento em 1616.824518 ms

Results
======

Correctly Classified Instances          48               96      %
Incorrectly Classified Instances         2                4      %
Kappa statistic                          0.9188
Mean absolute error                      0.04
Root mean squared error                  0.2
Relative absolute error                  7.9623 %
Root relative squared error             39.7812 %
Total Number of Instances               50

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,964    0,045    0,964      0,964    0,964      0,919    0,959     0,950     pos
                 0,955    0,036    0,955      0,955    0,955      0,919    0,959     0,931     neg
Weighted Avg.    0,960    0,041    0,960      0,960    0,960      0,919    0,959     0,942

=== Confusion Matrix ===

  a  b   <-- classified as
 27  1 |  a = pos
  1 21 |  b = neg
```



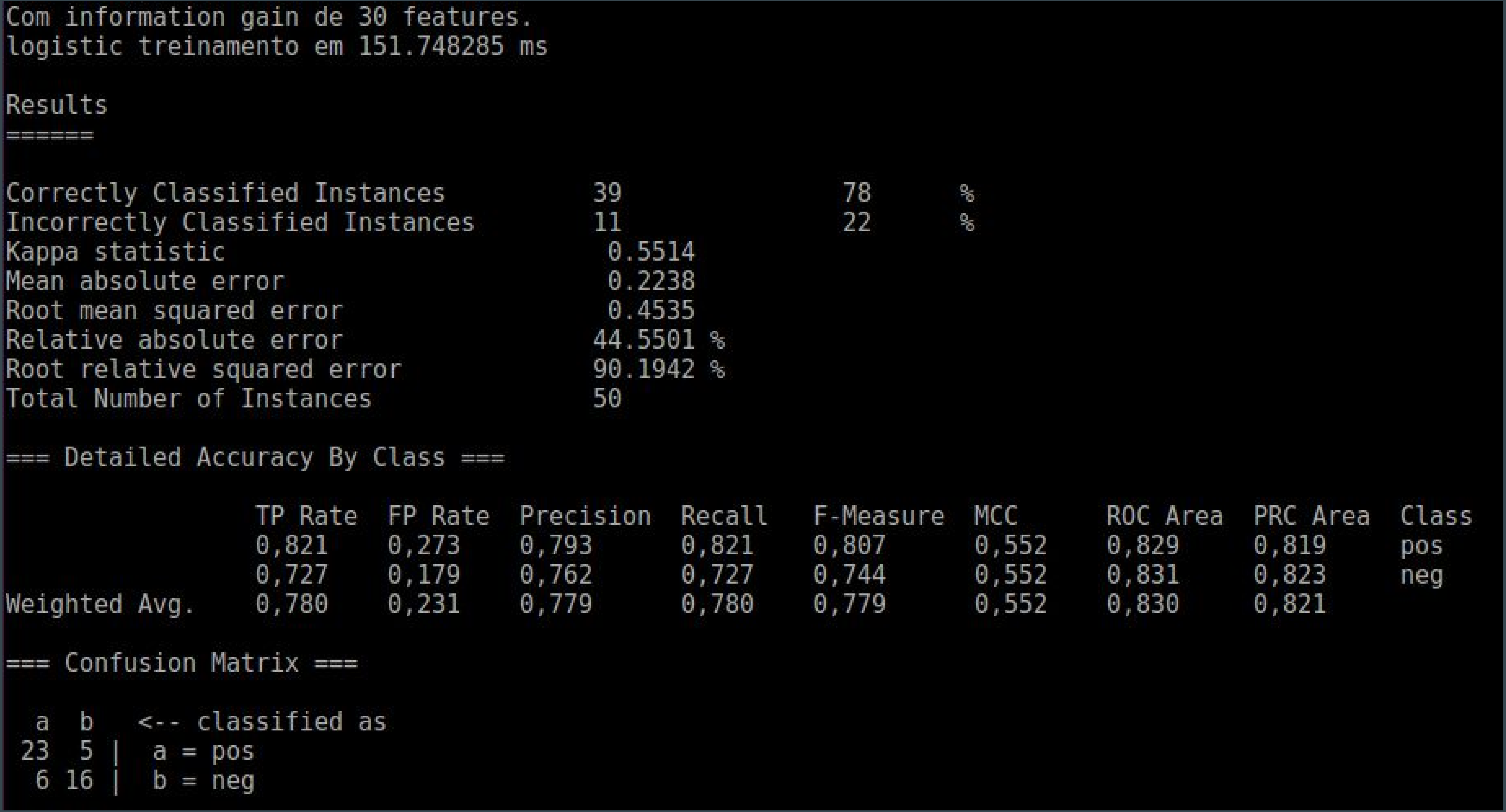
```
Com information gain de 30 features.
smo treinamento em 51.884117 ms

Results
======

Correctly Classified Instances          48               96      %
Incorrectly Classified Instances         2                4      %
Kappa statistic                          0.9188
Mean absolute error                      0.04
Root mean squared error                  0.2
Relative absolute error                  7.9623 %
Root relative squared error             39.7812 %
Total Number of Instances               50

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,964    0,045    0,964      0,964    0,964      0,919    0,959     0,950     pos
                 0,955    0,036    0,955      0,955    0,955      0,919    0,959     0,931     neg
Weighted Avg.    0,960    0,041    0,960      0,960    0,960      0,919    0,959     0,942

=== Confusion Matrix ===

  a  b   <-- classified as
 27  1 |  a = pos
  1 21 |  b = neg
```

Classificador - logistic sem information gain



```
Sem information gain.
Exception in thread "main" java.lang.OutOfMemoryError: Java heap space
        at weka.core.matrix.Matrix.<init>(Matrix.java:119)
        at weka.core.Optimization.findArgmin(Optimization.java:923)
        at weka.classifiers.functions.Logistic.buildClassifier(Logistic.java:819)
        at classifier.Classificador.build(Classificador.java:142)
        at classifier.Classificador.<init>(Classificador.java:53)
        at classifier.Main.main(Main.java:12)
```

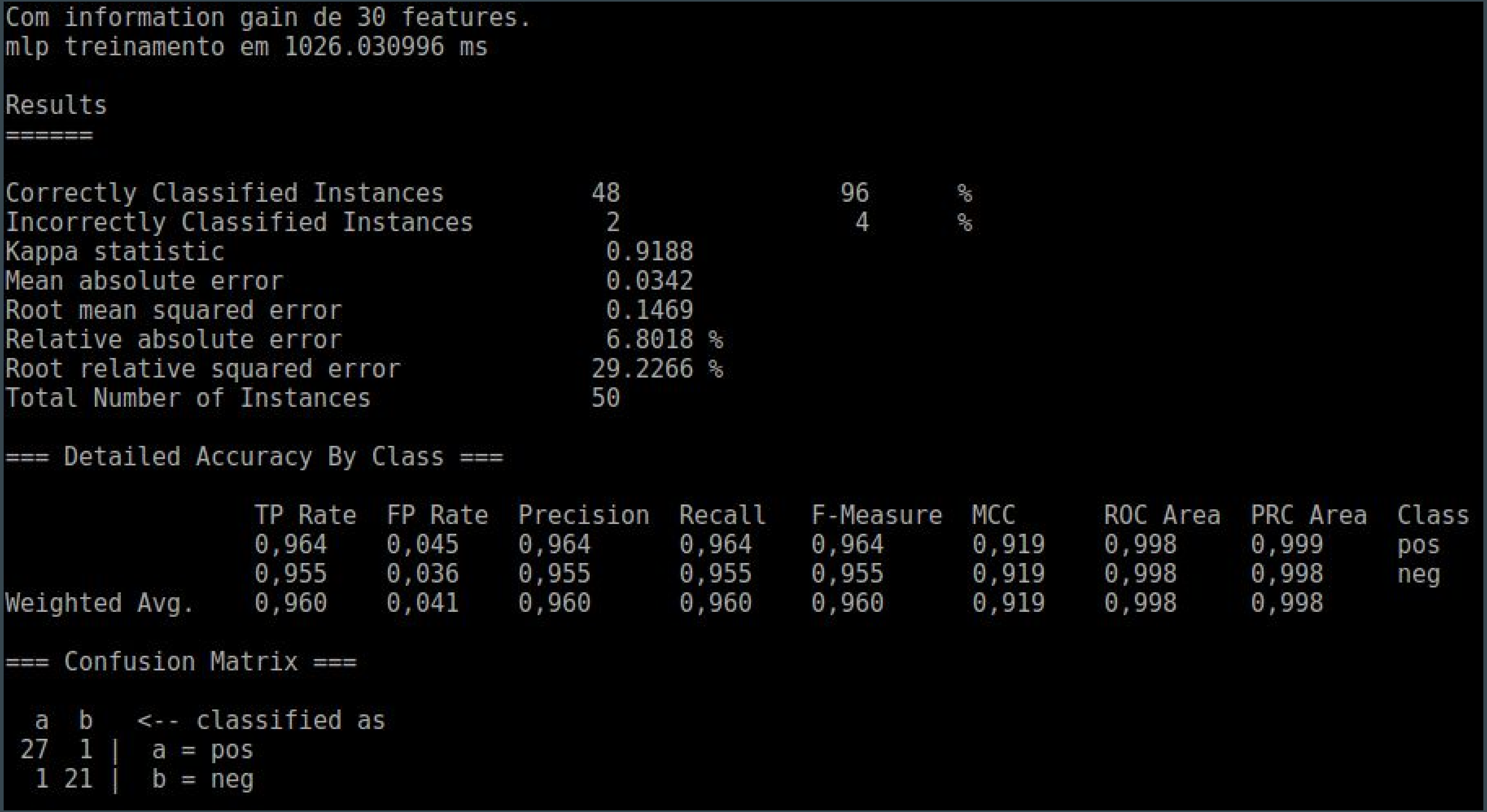
Mesmo aumentando o tamanho da heap do java, o programa demora muito (+5min) e não termina a execução.



```
Com information gain de 30 features.
logistic treinamento em 151.748285 ms

Results
======

Correctly Classified Instances          39               78      %
Incorrectly Classified Instances        11               22      %
Kappa statistic                          0.5514
Mean absolute error                      0.2238
Root mean squared error                  0.4535
Relative absolute error                 44.5501 %
Root relative squared error             90.1942 %
Total Number of Instances               50

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,821    0,273    0,793      0,821    0,807      0,552    0,829     0,819     pos
                 0,727    0,179    0,762      0,727    0,744      0,552    0,831     0,823     neg
Weighted Avg.    0,780    0,231    0,779      0,780    0,779      0,552    0,830     0,821

=== Confusion Matrix ===

  a  b   <-- classified as
 23  5 |  a = pos
  6 16 |  b = neg
```

Classificador - MultilayerPerceptron sem information gain

Rodou mais de 15 min e não terminou ...



```
Com information gain de 30 features.
mlp treinamento em 1026.030996 ms

Results
======

Correctly Classified Instances          48               96      %
Incorrectly Classified Instances         2                4      %
Kappa statistic                          0.9188
Mean absolute error                      0.0342
Root mean squared error                  0.1469
Relative absolute error                  6.8018 %
Root relative squared error             29.2266 %
Total Number of Instances               50

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,964    0,045    0,964      0,964    0,964      0,919    0,998     0,999     pos
                 0,955    0,036    0,955      0,955    0,955      0,919    0,998     0,998     neg
Weighted Avg.    0,960    0,041    0,960      0,960    0,960      0,919    0,998     0,998

=== Confusion Matrix ===

  a  b   <-- classified as
 27  1 |  a = pos
  1 21 |  b = neg
```

Classificador

	Bayes	j48	smo	logistic	mlp
Tudo	A - 78% P - 0.84 R - 0.75 T - 882ms	A - 100% P - 1 R - 1 T - 1488ms	A - 96% P - 0.964 R - 0.964 T - 1617ms	x	x
Information gain (30)	A - 73% P - 0.735 R - 0.893 R - 11ms	A - 94% P - 0.963 R - 0.929 T - 28ms	A - 96% P - 0.964 R - 0.964 T - 52ms	A - 78% P - 0.793 R - 0.821 T - 152ms	A - 96% P - 0.964 R - 0.964 T - 1026ms

A - Accuracy em rel. “pos”
P - Precision em rel. “pos”
R - Recall em rel. “pos”
T - Tempo de treinamento

Classificador

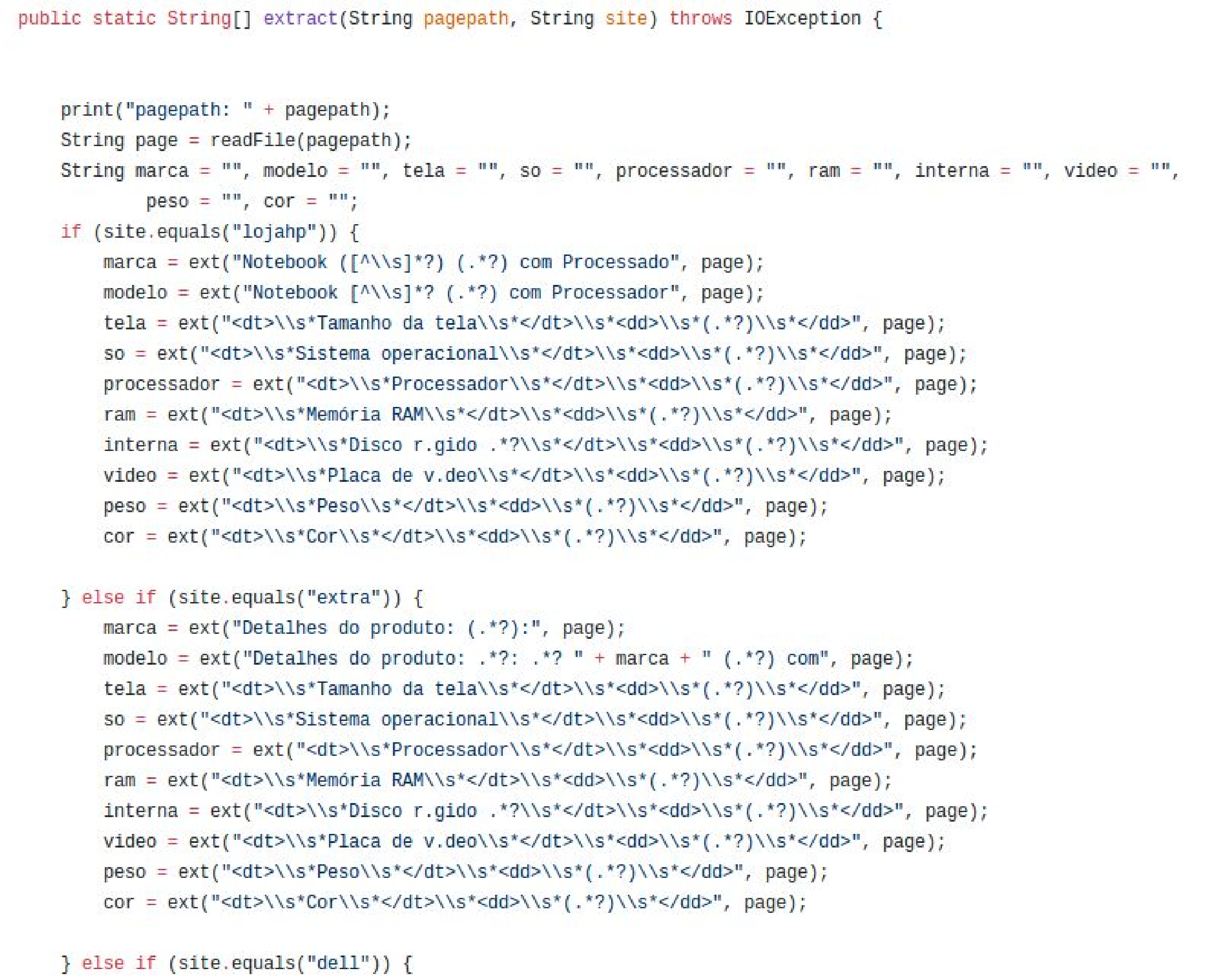
Bayes é o mais rápido e de menor precisão para a classe dos positivos, também acerta mais sem information gain

Bayes e Logistic não são bons em acerto quando comparados aos outros.

MultilayerPerceptron é o mais lento, porém tem boas taxas de acerto.

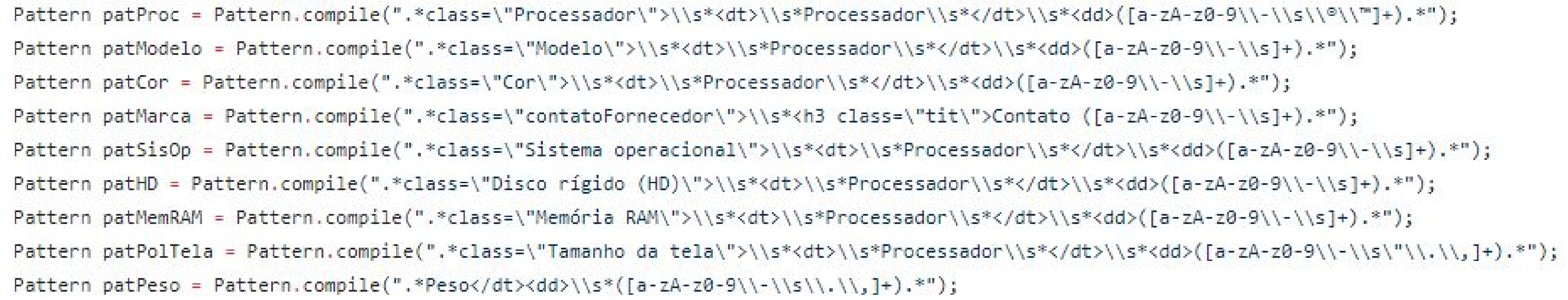
J48 e smo parecem ser os que melhor se aplicam para os casos testados.

Extrator



```
public static String[] extract(String pagepath, String site) throws IOException {


    print("pagepath: " + pagepath);
    String page = readFile(pagepath);
    String marca = "", modelo = "", tela = "", so = "", processador = "", ram = "", interna = "", video = "",
            peso = "", cor = "";
    if (site.equals("lojahp")) {
        marca = ext("Notebook ([^\\s]*?) (.*?) com Processado", page);
        modelo = ext("Notebook [^\\s]*? (.*?) com Processador", page);
        tela = ext("<dt>\\s*Tamanho da tela\\s*</dt>\\s*<dd>\\s*(.*?)\\s*</dd>", page);
        so = ext("<dt>\\s*Sistema operacional\\s*</dt>\\s*<dd>\\s*(.*?)\\s*</dd>", page);
        processador = ext("<dt>\\s*Processador\\s*</dt>\\s*<dd>\\s*(.*?)\\s*</dd>", page);
        ram = ext("<dt>\\s*Memória RAM\\s*</dt>\\s*<dd>\\s*(.*?)\\s*</dd>", page);
        interna = ext("<dt>\\s*Disco r.gido .*?\\s*</dt>\\s*<dd>\\s*(.*?)\\s*</dd>", page);
        video = ext("<dt>\\s*Placa de v.deo\\s*</dt>\\s*<dd>\\s*(.*?)\\s*</dd>", page);
        peso = ext("<dt>\\s*Peso\\s*</dt>\\s*<dd>\\s*(.*?)\\s*</dd>", page);
        cor = ext("<dt>\\s*Cor\\s*</dt>\\s*<dd>\\s*(.*?)\\s*</dd>", page);

    } else if (site.equals("extra")) {
        marca = ext("Detalhes do produto: (.*?):", page);
        modelo = ext("Detalhes do produto: .*?: .*? " + marca + " (.*?) com", page);
        tela = ext("<dt>\\s*Tamanho da tela\\s*</dt>\\s*<dd>\\s*(.*?)\\s*</dd>", page);
        so = ext("<dt>\\s*Sistema operacional\\s*</dt>\\s*<dd>\\s*(.*?)\\s*</dd>", page);
        processador = ext("<dt>\\s*Processador\\s*</dt>\\s*<dd>\\s*(.*?)\\s*</dd>", page);
        ram = ext("<dt>\\s*Memória RAM\\s*</dt>\\s*<dd>\\s*(.*?)\\s*</dd>", page);
        interna = ext("<dt>\\s*Disco r.gido .*?\\s*</dt>\\s*<dd>\\s*(.*?)\\s*</dd>", page);
        video = ext("<dt>\\s*Placa de v.deo\\s*</dt>\\s*<dd>\\s*(.*?)\\s*</dd>", page);
        peso = ext("<dt>\\s*Peso\\s*</dt>\\s*<dd>\\s*(.*?)\\s*</dd>", page);
        cor = ext("<dt>\\s*Cor\\s*</dt>\\s*<dd>\\s*(.*?)\\s*</dd>", page);

    } else if (site.equals("dell")) {
```

Extrator



```
Pattern patProc = Pattern.compile(".*class=\"Processador\">\\s*<dt>\\s*Processador\\s*</dt>\\s*<dd>([a-zA-z0-9\\-\\s\\®\\™]+).*");
Pattern patModelo = Pattern.compile(".*class=\"Modelo\">\\s*<dt>\\s*Processador\\s*</dt>\\s*<dd>([a-zA-z0-9\\-\\s]+).*");
Pattern patCor = Pattern.compile(".*class=\"Cor\">\\s*<dt>\\s*Processador\\s*</dt>\\s*<dd>([a-zA-z0-9\\-\\s]+).*");
Pattern patMarca = Pattern.compile(".*class=\"contatoFornecedor\">\\s*<h3 class=\"tit\">Contato ([a-zA-z0-9\\-\\s]+).*");
Pattern patSisOp = Pattern.compile(".*class=\"Sistema operacional\">\\s*<dt>\\s*Processador\\s*</dt>\\s*<dd>([a-zA-z0-9\\-\\s]+).*");
Pattern patHD = Pattern.compile(".*class=\"Disco rígido (HD)\">\\s*<dt>\\s*Processador\\s*</dt>\\s*<dd>([a-zA-z0-9\\-\\s]+).*");
Pattern patMemRAM = Pattern.compile(".*class=\"Memória RAM\">\\s*<dt>\\s*Processador\\s*</dt>\\s*<dd>([a-zA-z0-9\\-\\s]+).*");
Pattern patPolTela = Pattern.compile(".*class=\"Tamanho da tela\">\\s*<dt>\\s*Processador\\s*</dt>\\s*<dd>([a-zA-z0-9\\-\\s\"\\.\\,]+).*");
Pattern patPeso = Pattern.compile(".*Peso</dt><dd>\\s*([a-zA-z0-9\\-\\s\\.\\,]+).*");
```

Extrator

- Total de extrações possíveis: N = 20
- Total de pares extraídos pelo sistema: E = 20
- Total de pares extraídos corretamente: C = 10

Extrator

- Recall = 0.5
- Precision = 0.5
- F-Measure = 0.5