Recuperação de Informação Jogos

Leonardo Ribeiro - *Crawler*
Eduardo Almeida - *Classificador*
Emanuel Silva - *Extrator*

Crawler

Robots.txt

1. O sistema antes de iniciar o Crawler visita o robots.txt do site.
2. O sistema gera um Regex conforme o robots.txt para evitar que páginas em especial não sejam visitadas.
3. Antes de visitar qualquer página o sistema verifica o Regex.

Heurísticas

Sistema de pontuação com base na quantidade de palavras chave que aparecem.

1. O sistema separa todos os links presentes documento html da página visitada.
2. No segundo passo o sistema separa todas as do documento
3. Após isso faz uma verificação de quais links estão presentes naquela div e atribui um ponto para todos os links dentro da página caso haja uma palavra chave dentro da (um ponto para cada palavra chave)
4. Antes de salvar o link + pontuação em uma estrutura de dados o sistema verifica se existe alguma palavra chave dentro do próprio link e atribui 100 pontos para cada correspondência.
5. Caso haja palavras chaves no link negativas no link só é adicionado um ponto

Resultados

Problemas enfrentados

1. Algumas páginas em especial como a Origin redireciona o link inicial, fazendo que com o primeiro link nunca seja visitado e dessa forma o sistema acaba entrando em loop infinito. (não resolvido)
2. Algumas páginas não tinham uma organização para o seus produtos em sua url, de forma que foi necessário o sistema de pontuação pelas para resolver o problema.
3. O site da steam em específico oferece a opção de acessar o mesmo anúncio de jogo em idiomas diferentes, de forma que anúncios do mesmo jogo aparecem em idiomas distintos.
4. Sites de produtos gerais também oferecem acessórios para jogos e que também se encontram na mesma categoria, de forma que foi necessário criar as palavras chaves negativas para verificar o que está sendo tratado na url.

Classificação

Classificação

1. Rotular exemplos positivos e negativos

2. Criar o conjunto de features usando feature selection

3. Treinar o classificador com uma ferramenta de ML

– Métodos: Naïve bayes, Decision tree (J48), SVM (SMO), Logistic regression, Multilayer perceptron

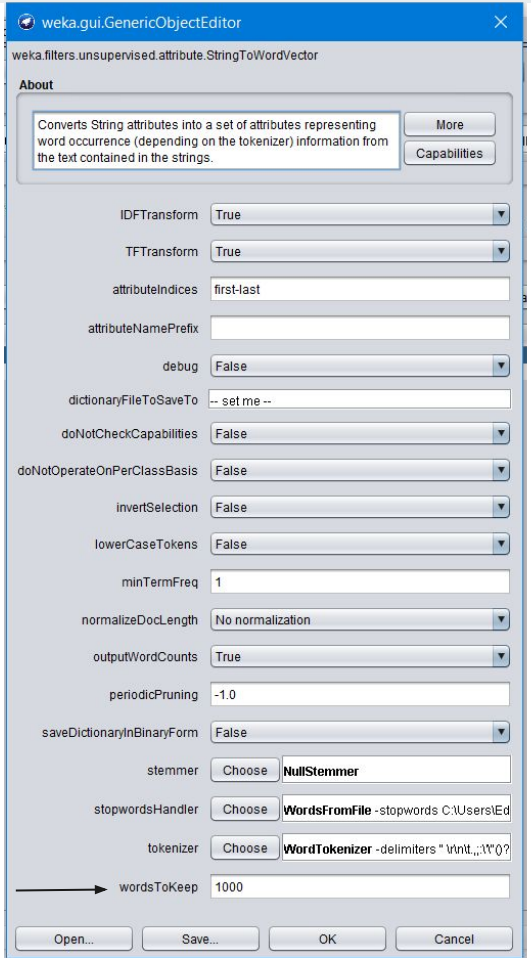
4. Comparar estratégias

Classificação

- Exemplos positivos e negativos
  - Folder
- Jsoup (Html -> Txt)
  - Código
- Weka TextDirectoryLoader (Txts -> Arff)
  - Código / WEKA “CLI”

Classificação

- Carregar Dataset (Link)
- Weka Filter: StringToWordVector
  - Stopwords



Mantendo 1000 palavras

Feature Selection

Correlation Based Feature Selection

Search Method: Ranker

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 1 @@class@@):
        Correlation Ranking Filter
Ranked attributes:
 0.652787855490286          1075 classificaÃ§Ã£o
 0.6522280837718224          566 jogo
 0.5776938278449559         1124 indicativa
 0.5135669666067229         1129 jogadores
 0.4984732587591599         1128 jogador
 0.4472805801113007         1148 mÃ-dia
 0.43826389044949426         481 game
 0.41297875006397533        1134 legenda
 0.40962732188824397         759 plataforma
 0.3993536257950106         1138 livre
 0.3746733700704161         1059 anos
 0.36805332713575933        1086 desenvolvedor
 0.3616493800866023          647 modelo
 0.35932071370747776        1212 violÃªncia
 0.3506810821959437          298 cor
 0.3407677821779824         1118 habilidades
 0.33885237991978606         505 gÃªnero
 0.33412896116524105         807 ps4
 0.32768692809440797        1144 movimento
 0.3248109299769483         1179 recomendado
```

Feature Selection

Information Gain Based Feature Selection

Search Method: Ranker

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 1 @@class@@):
        Information Gain Ranking Filter

Ranked attributes:
 0.4248     1075 classificaÃ§Ã£o
 0.4038      566 jogo
 0.3297     1124 indicativa
 0.2771     1129 jogadores
 0.247       759 plataforma
 0.2426     1128 jogador
 0.2178     1086 desenvolvedor
 0.1976      481 game
 0.192       176 aventura
 0.1873      505 gÃªnero
 0.1836     1134 legenda
 0.1826     1148 mÃ-dia
 0.1553      297 controle
 0.1526     1138 livre
 0.1345     1212 violÃªncia
 0.1278     1118 habilidades
 0.1245      485 games
 0.1225      647 modelo
 0.1212     1127 jogabilidade
 0.1212     1144 movimento
 0.1155      519 idiomas
 0.1092     1100 explore
 0.1016     1062 armas
 0.0996     1059 anos
 0.0952     1019 wi-fi
 0.0948     1090 disponÃ-veis
 0.0936     1179 recomendado
```

Feature Selection

- classificação
- jogadores
- desenvolvedor
- gênvero
- plataforma

No.		Name
1	☐	classificaÃ§Ã£o
2	☑	jogo
3	☑	indicativa
4	☐	jogadores
5	☐	plataforma
6	☑	jogador
7	☐	desenvolvedor
8	☑	game
9	☑	aventura
10	☐	gÃªnero
11	☑	legenda
12	☑	mÃ-dia
13	☑	controle
14	☑	livre
15	☑	violÃªncia
16	☑	habilidades
17	☑	games
18	☑	modelo
19	☑	jogabilidade
20	☑	movimento
21	☑	idiomas
22	☑	explore
23	☑	armas
24	☑	anos
25	☑	wi-fi
26	☑	disponÃ-veis
27	☑	recomendado
28	☑	usb

Naïve Bayes

```
=== Summary ===

Correctly Classified Instances         160               88.8889 %
Incorrectly Classified Instances        20               11.1111 %
Kappa statistic                          0.7778
Mean absolute error                      0.1113
Root mean squared error                  0.3181
Relative absolute error                 22.2649 %
Root relative squared error             63.6188 %
Total Number of Instances              180

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,867     0,089     0,907       0,867   0,886      0,779    0,961     0,949     Neg
               0,911     0,133     0,872       0,911   0,891      0,779    0,961     0,958     Pos
Weighted Avg.  0,889     0,111     0,890       0,889   0,889      0,779    0,961     0,954

=== Confusion Matrix ===

  a  b   <-- classified as
 78 12 |  a = Neg
  8 82 |  b = Pos
```

Log Completo

Decision Tree (J48)

```
=== Summary ===

Correctly Classified Instances         163               90.5556 %
Incorrectly Classified Instances        17                9.4444 %
Kappa statistic                          0.8111
Mean absolute error                      0.1494
Root mean squared error                  0.2733
Relative absolute error                 29.8762 %
Root relative squared error             54.6592 %
Total Number of Instances              180

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,856     0,044     0,951       0,856   0,901      0,815    0,942     0,916     Neg
               0,956     0,144     0,869       0,956   0,910      0,815    0,942     0,915     Pos
Weighted Avg.  0,906     0,094     0,910       0,906   0,905      0,815    0,942     0,916

=== Confusion Matrix ===

  a  b   <-- classified as
 77 13 |  a = Neg
  4 86 |  b = Pos
```

Log Completo

SMO

```
=== Summary ===

Correctly Classified Instances         160               88.8889 %
Incorrectly Classified Instances        20               11.1111 %
Kappa statistic                          0.7778
Mean absolute error                      0.1111
Root mean squared error                  0.3333
Relative absolute error                 22.2222 %
Root relative squared error             66.6667 %
Total Number of Instances              180

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,867     0,089     0,907       0,867   0,886      0,779    0,889     0,853     Neg
               0,911     0,133     0,872       0,911   0,891      0,779    0,889     0,839     Pos
Weighted Avg.  0,889     0,111     0,890       0,889   0,889      0,779    0,889     0,846

=== Confusion Matrix ===

  a  b   <-- classified as
 78 12 |  a = Neg
  8 82 |  b = Pos
```

Log Completo

Logistic Regression

```
=== Summary ===

Correctly Classified Instances         161               89.4444 %
Incorrectly Classified Instances        19               10.5556 %
Kappa statistic                          0.7889
Mean absolute error                      0.1627
Root mean squared error                  0.2856
Relative absolute error                 32.549  %
Root relative squared error             57.1133 %
Total Number of Instances              180

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,878     0,089     0,908       0,878   0,893      0,789    0,960     0,953     Neg
               0,911     0,122     0,882       0,911   0,896      0,789    0,960     0,956     Pos
Weighted Avg.  0,894     0,106     0,895       0,894   0,894      0,789    0,960     0,954

=== Confusion Matrix ===

  a  b   <-- classified as
 79 11 |  a = Neg
  8 82 |  b = Pos
```

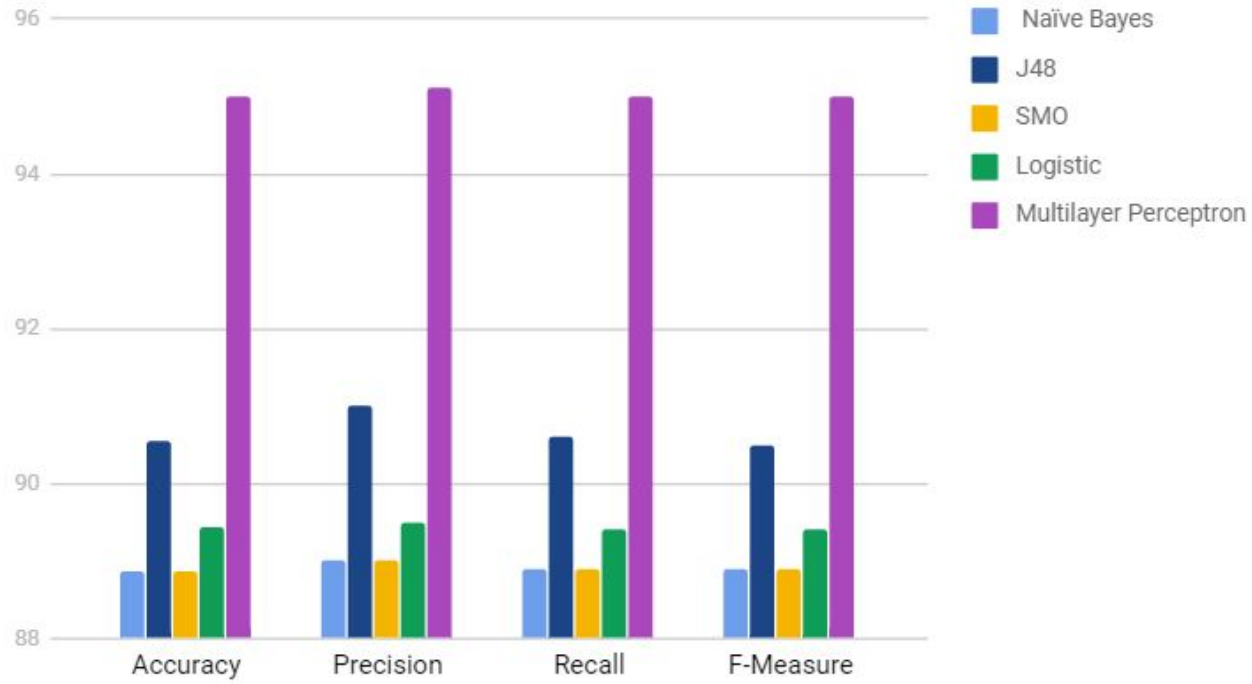
Log Completo

Multilayer Perceptron

```
=== Summary ===

Correctly Classified Instances         171               95      %
Incorrectly Classified Instances         9                5      %
Kappa statistic                          0.9
Mean absolute error                      0.0778
Root mean squared error                  0.1835
Relative absolute error                 15.5652 %
Root relative squared error             36.6969 %
Total Number of Instances              180

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0,978     0,078     0,926       0,978   0,951      0,901    0,984     0,971     Neg
                 0,922     0,022     0,976       0,922   0,949      0,901    0,984     0,984     Pos
Weighted Avg.    0,950     0,050     0,951       0,950   0,950      0,901    0,984     0,977

=== Confusion Matrix ===

  a  b   <-- classified as
 88  2 |  a = Neg
  7 83 |  b = Pos
```

Log Completo

Comparação

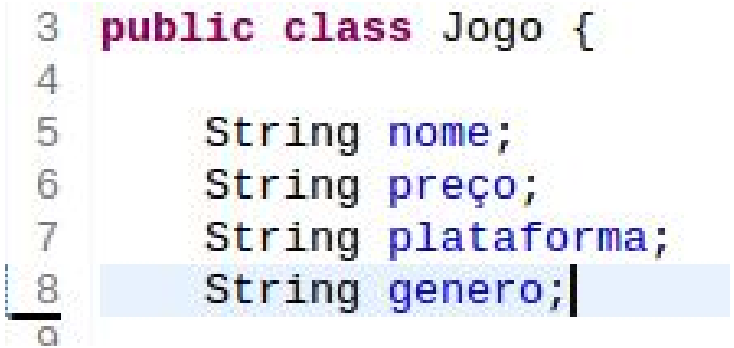

96
94
92
90
88
Accuracy
Precision
Recall
F-Measure
Naïve Bayes
J48
SMO
Logistic
Multilayer Perceptron

Extração

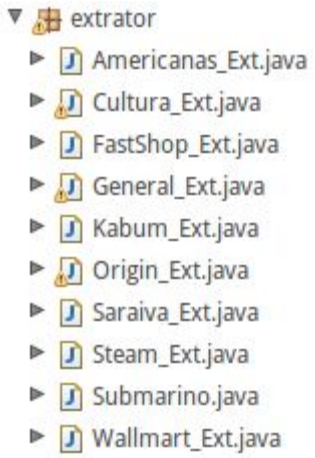
Extração

Base:

- 10 páginas;
- 4 pares por página:
    - Nome;
    - Preço;
    - Plataforma;
    - Gênero;



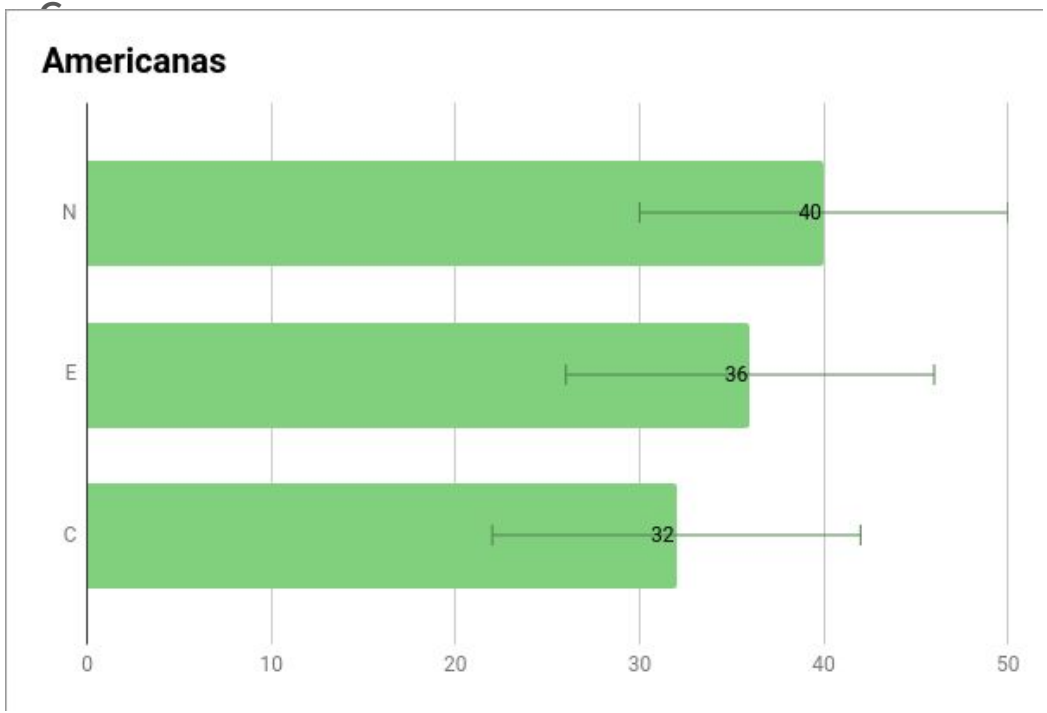
Processo de Extração - Wrappers Exclusivos




```
//Getting HTML from file
File input = new File("/home/emanuel/htmls/kabum/KaBuM5.html");
Document doc = Jsoup.parse(input, "UTF-8", "http://kabum.com/");

//Initializing database
Jogo jg = new Jogo();

//Getting name
Elements name = doc.getElementsByClass("titulo_det");
jg.setNome(name.text());

//Getting price
Elements price = doc.getElementsByClass("preco_normal");
jg.setPreço(price.text());

//Getting platform
Elements plat = doc.select("div:matches(id=\"shout_.+?\")");
jg.setPlataforma(plat.text());

//Getting genre
```

Resultados - Wrapper Americanas

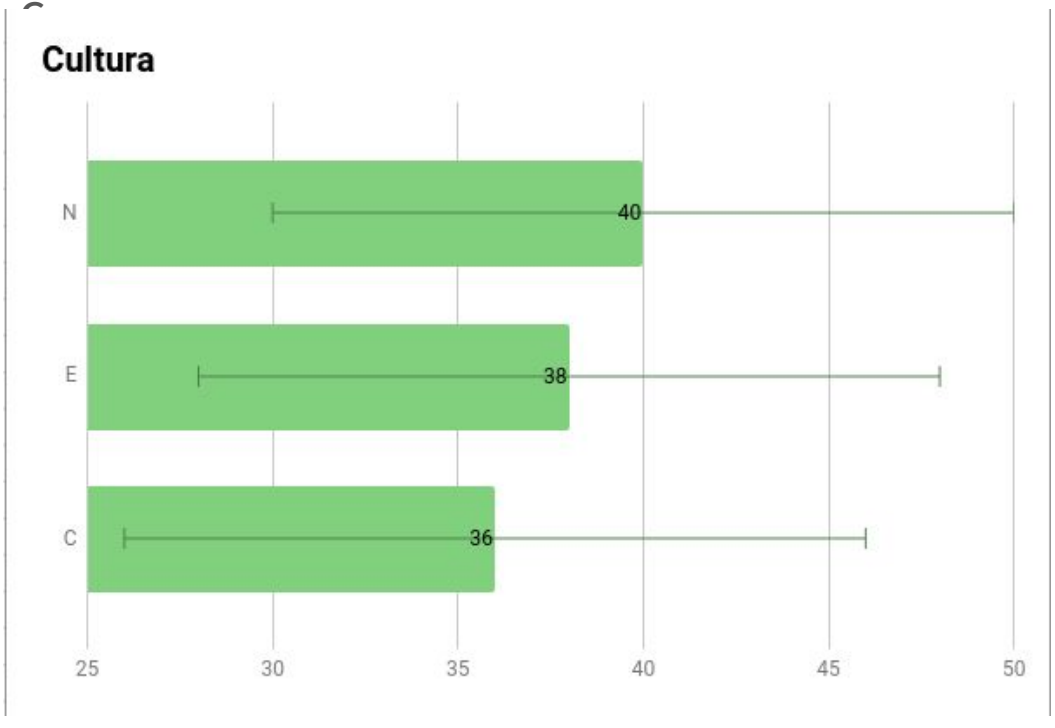
Total de Extrações Possíveis: N | Total de Pares Extraídos: E | Total de Pares Extraídos Corretamente: C



Recall = 0,8
Precision = 0,88
F-Measure = 0,83

Resultados - Wrapper Cultura

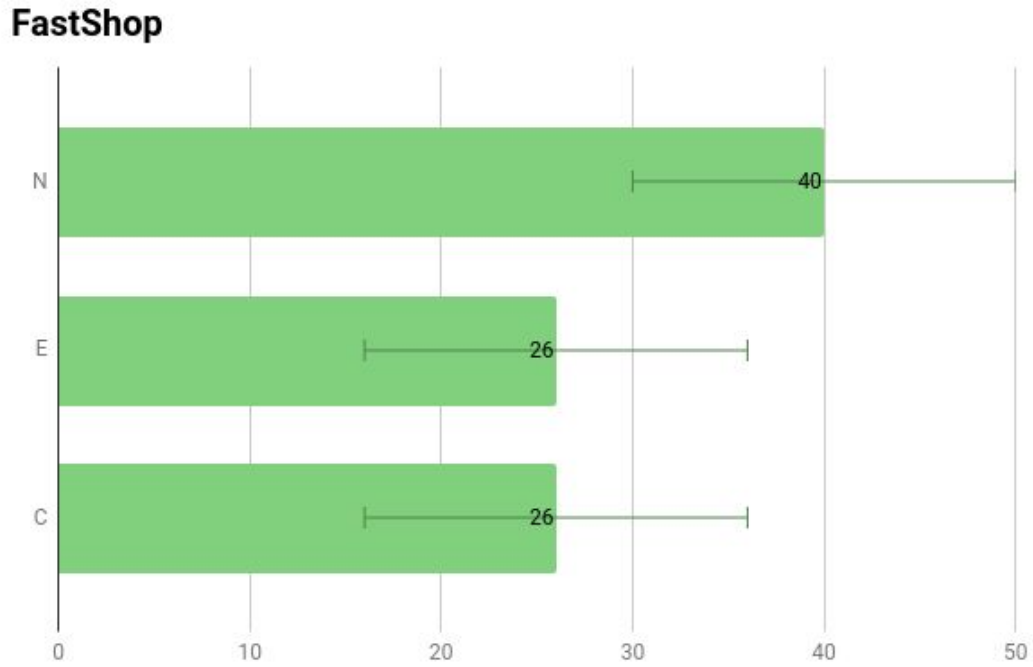
Total de Extrações Possíveis: N | Total de Pares Extraídos: E | Total de Pares Extraídos Corretamente: C



Recall = 0,9
Precision = 0,94
F-Measure = 0,92

Resultados - Wrapper FastShop

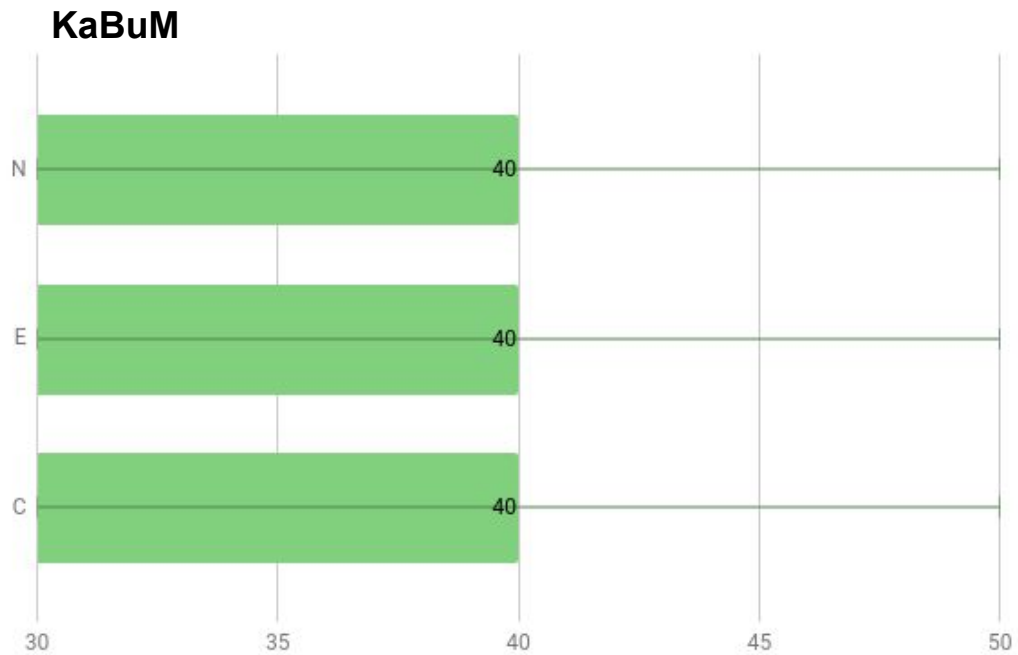
Total de Extrações Possíveis: N | Total de Pares Extraídos: E | Total de Pares Extraídos Corretamente:



Recall = 0,65
Precision = 1
F-Measure = 0,78

Resultados - Wrapper KaBuM

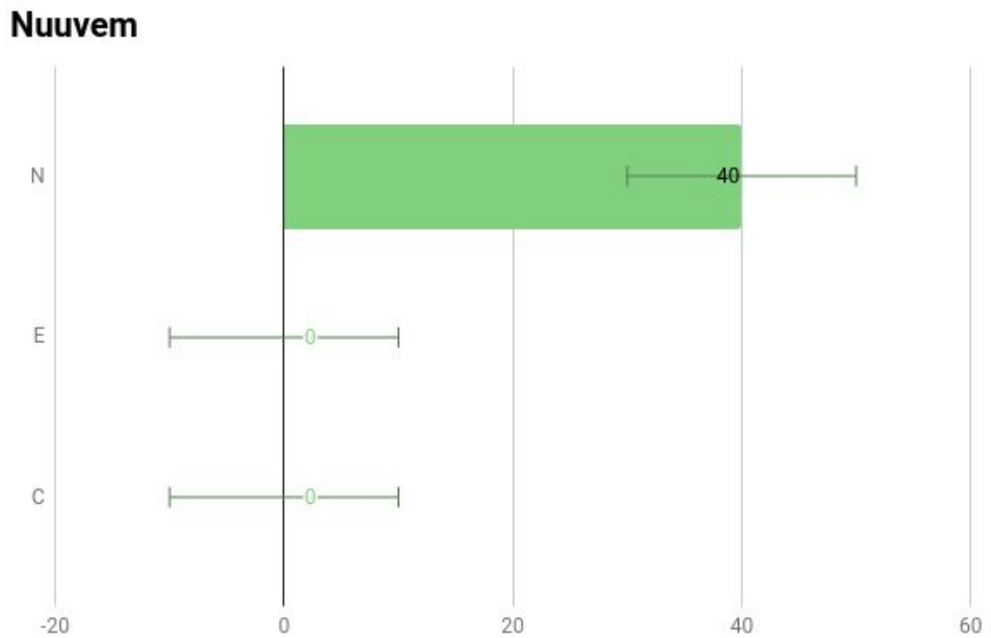
Total de Extrações Possíveis: N | Total de Pares Extraídos: E | Total de Pares Extraídos Corretamente: C



Recall = 1
Precision = 1
F-Measure = 1

Resultados - Wrapper Nuuvem

Total de Extrações Possíveis: N | Total de Pares Extraídos: E | Total de Pares Extraídos Corretamente: C

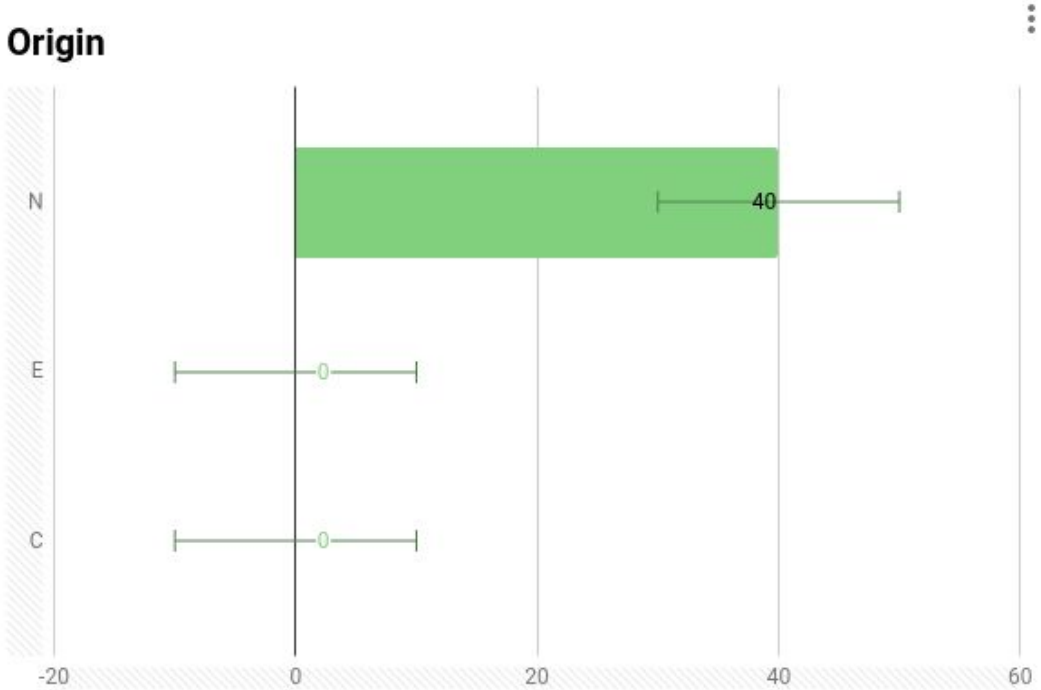


Recall = 0
Precision = -
F-Measure = -

Não foi possível capturar as páginas HTML.

Resultados - Wrapper Origin

Total de Extrações Possíveis: N | Total de Pares Extraídos: E | Total de Pares Extraídos Corretamente: C

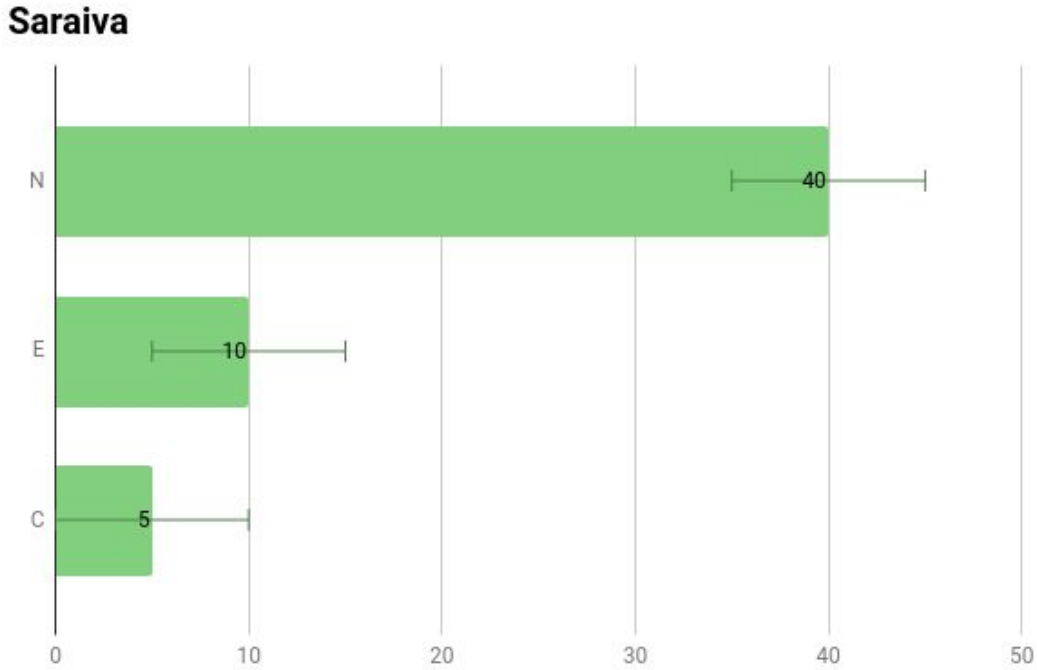


Recall = 0
Precision = -
F-Measure = -

Apesar de capturadas, ao abrir o HTML as páginas não eram carregadas. Provavelmente algum mecanismo de proteção de conteúdo.

Resultados - Wrapper Saraiva

Total de Extrações Possíveis: N | Total de Pares Extraídos: E | Total de Pares Extraídos Corretamente: C

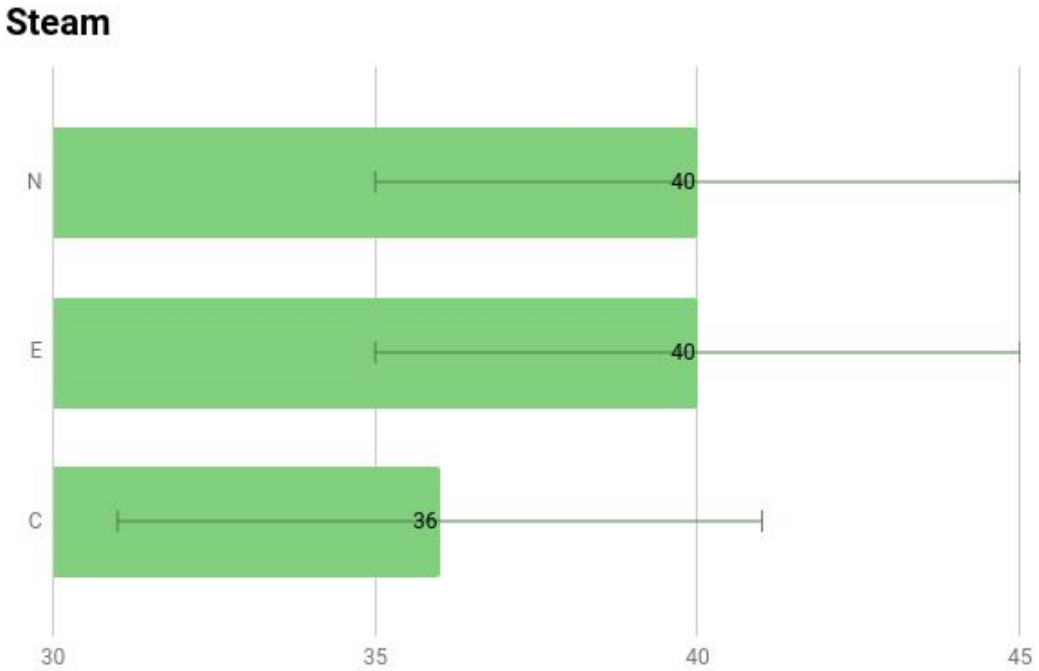


Recall = 0,125
Precision = 0,5
F-Measure = 0,2

A página utilizava *ReactIDs* que são gerados dinamicamente.

Resultados - Wrapper Steam

Total de Extrações Possíveis: N | Total de Pares Extraídos: E | Total de Pares Extraídos Corretamente:

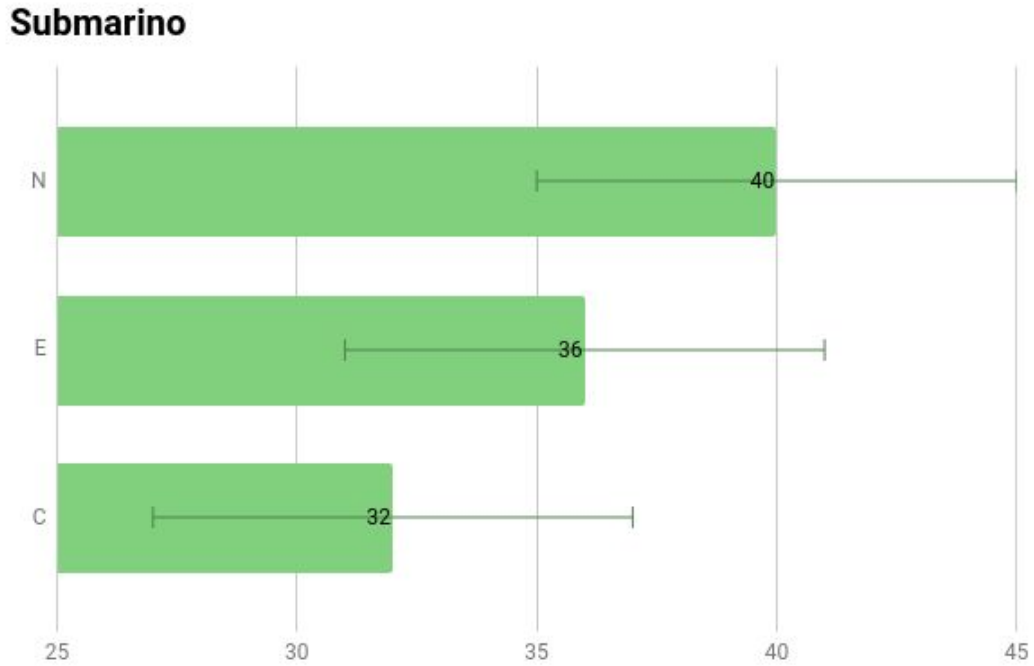


Recall = 0,9
Precision = 0,9
F-Measure = 0,9

3 dos 4 pares eram bem formatados com o contexto da página e eram sempre possíveis de extração. O quarto era gerado dinamicamente, e por ser textual, usar regex enviesaria os resultados.

Resultados - Wrapper Submarino

Total de Extrações Possíveis: N | Total de Pares Extraídos: E | Total de Pares Extraídos Corretamente:

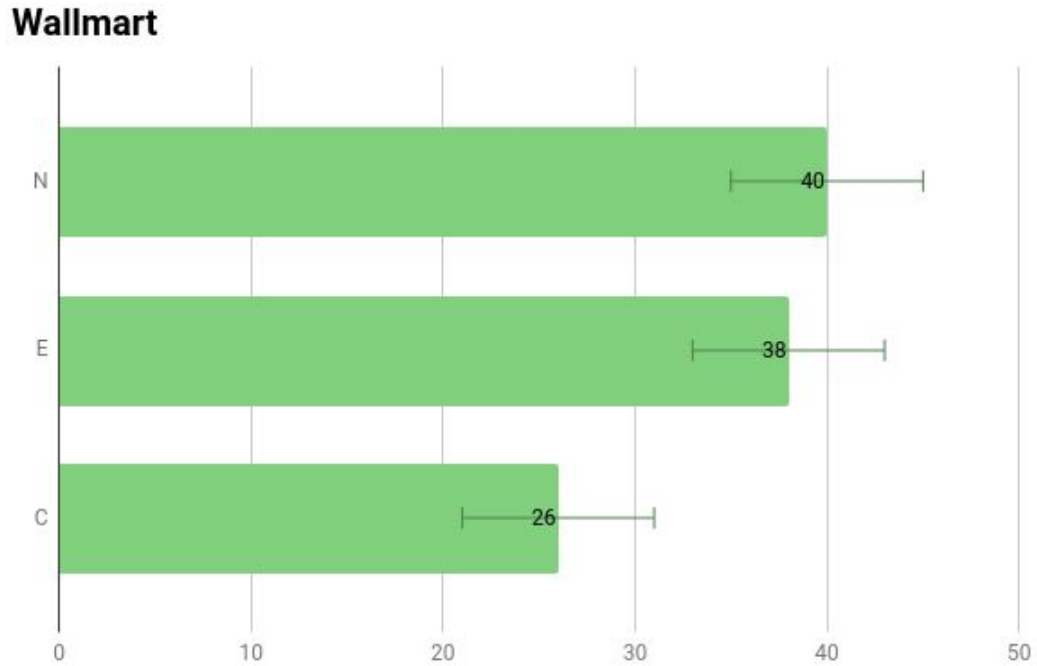


Recall = 0,8
Precision = 0,88
F-Measure = 0,83

Tentativa de usar DOM Tree para caminhar pelas estruturas do documento.

Resultados - Wrapper Wallmart

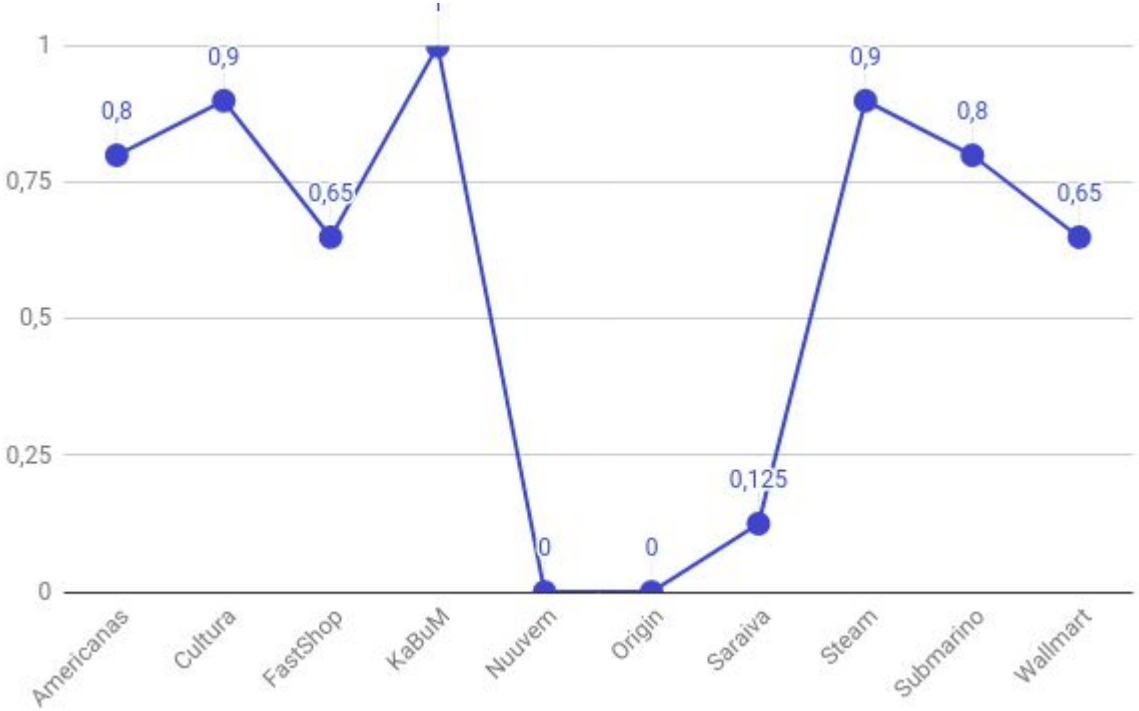
Total de Extrações Possíveis: N | Total de Pares Extraídos: E | Total de Pares Extraídos Corretamente:



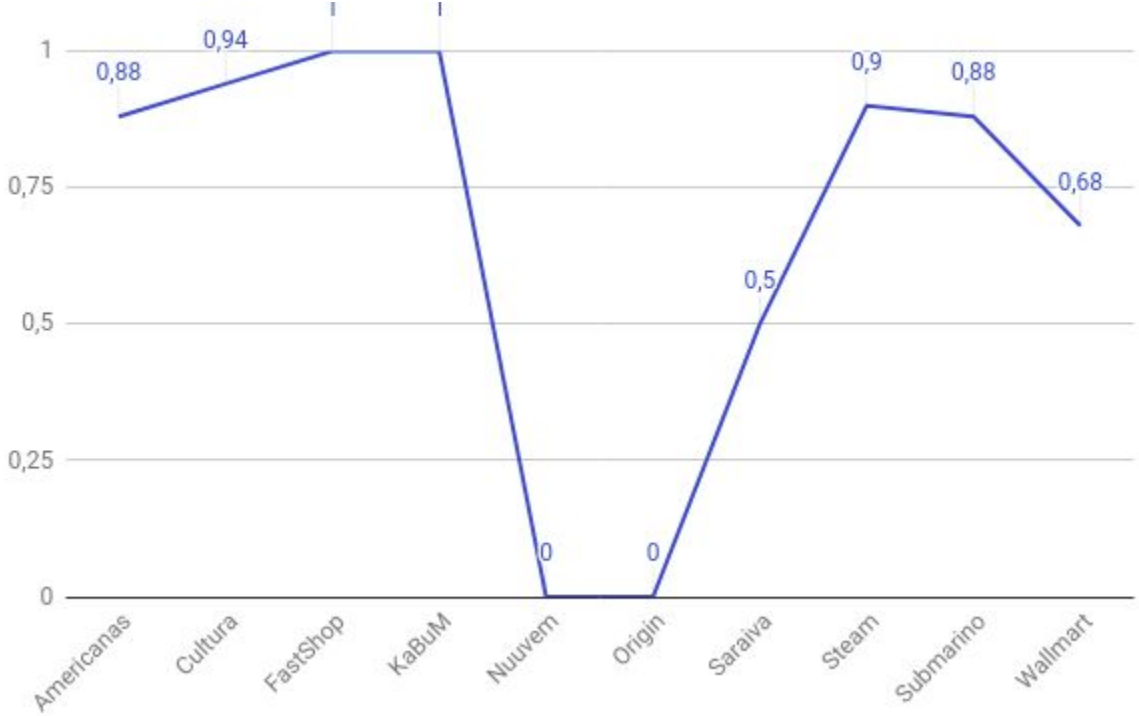
Recall = 0,65
Precision = 0,68
F-Measure = 0,66

Tentativa de usar DOM Tree para caminhar pelas estruturas do documento.

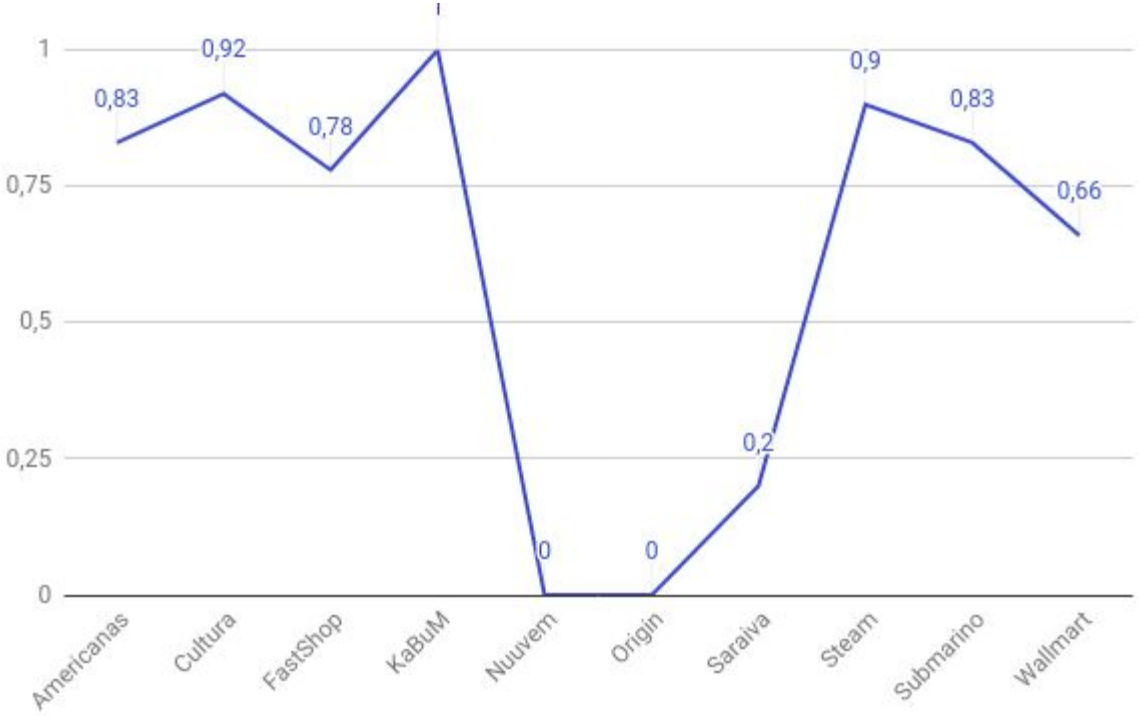
Resultado - Recall



Resultado - Precision



Resultado - F-Measure



Extrator Genérico

Estratégias tentadas:

- Navegação por DOM Tree;
- Navegação pelos elementos CSS + Expressões Regulares;
- Tentativas de captura através de Regex puro.

Resultados

- Em geral muito específico à alguns domínios;
- Resultados extremamente negativos de captura;
- Uso de Regex retornava muitos elementos, sendo necessário conhecimento do domínio para filtragem, o que foge ao princípio de extrator genérico.
- **Taxa de Recuperação para 400 pares = 40 pares encontrados. 10%.**