Recuperação de Informação

Câmeras Digitais

•••

Bruno Cavalcanti
Guilherme Henrique
Pedro Henrique

Domínios

- Sony
- Newegg
- Nikon
- Dpreview
- Currys
- Sigmaphoto
- Ricoh
- Visions
- Canon
- WexPhotographic

Crawler

•••

Bruno Cavalcanti

Jsoup

CrawlerBFS

Crawler Classifier

WEKA
The University
of Waikato

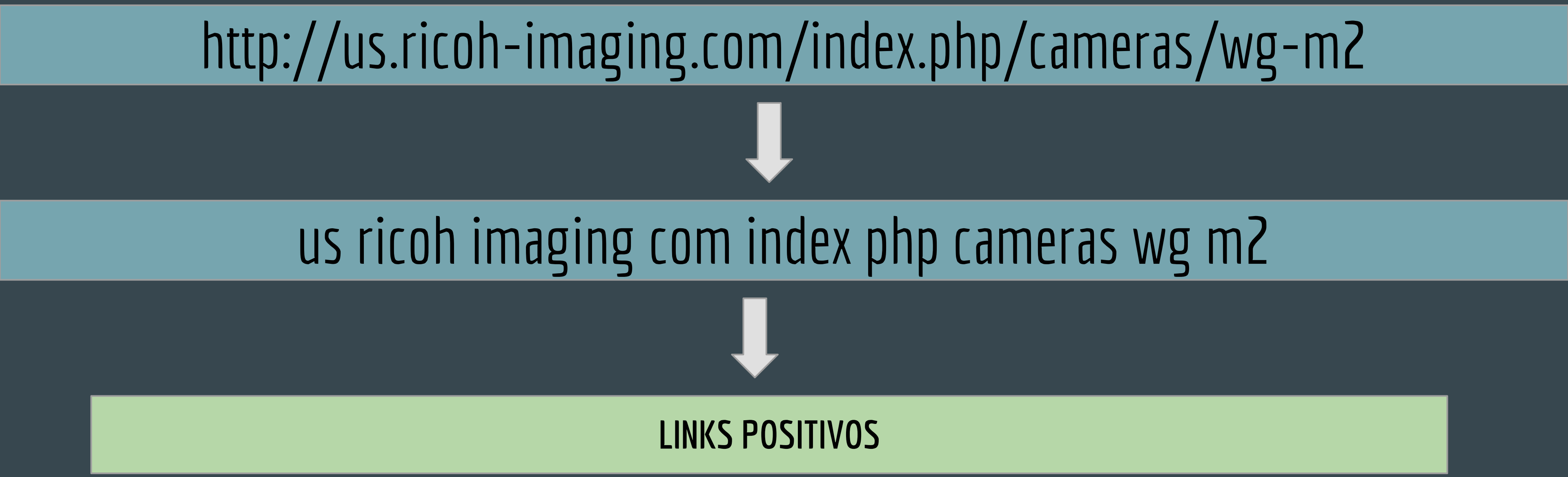
LINKS POSITIVOS

LINKS MUITO PRÓXIMOS

LINKS PRÓXIMOS

LINKS DISTANTES

conjuntos de treinamento

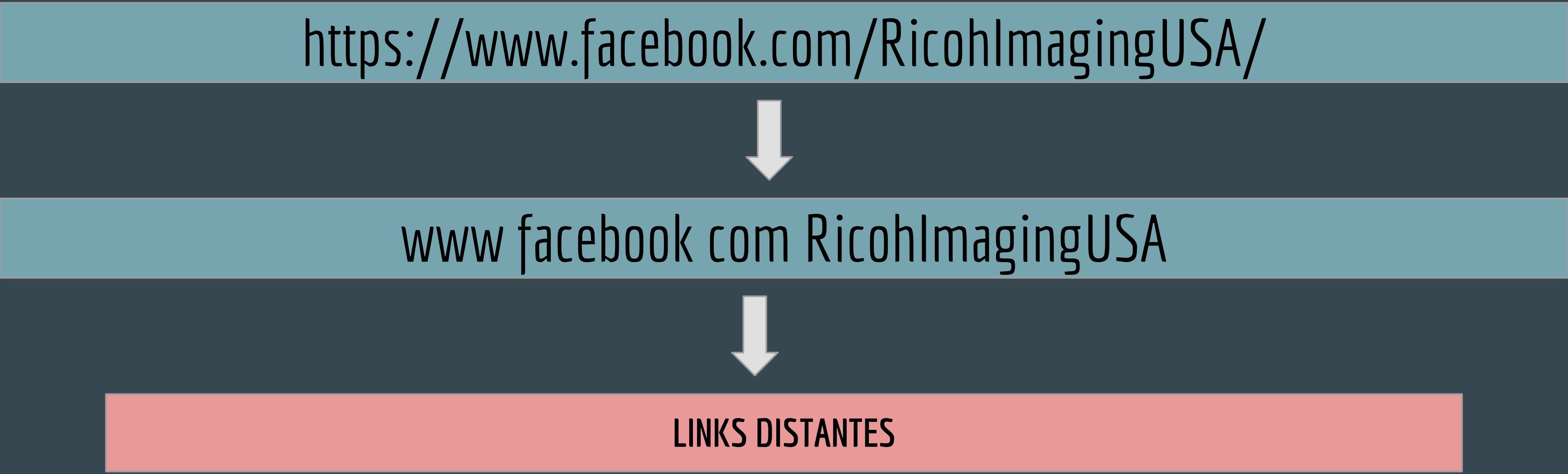


http://us.ricoh-imaging.com/index.php/cameras/wg-m2

us ricoh imaging com index php cameras wg m2

LINKS POSITIVOS

tratamento dos links



https://www.facebook.com/RicohImagingUSA/

www facebook com RicohImagingUSA

LINKS DISTANTES

tratamento dos links

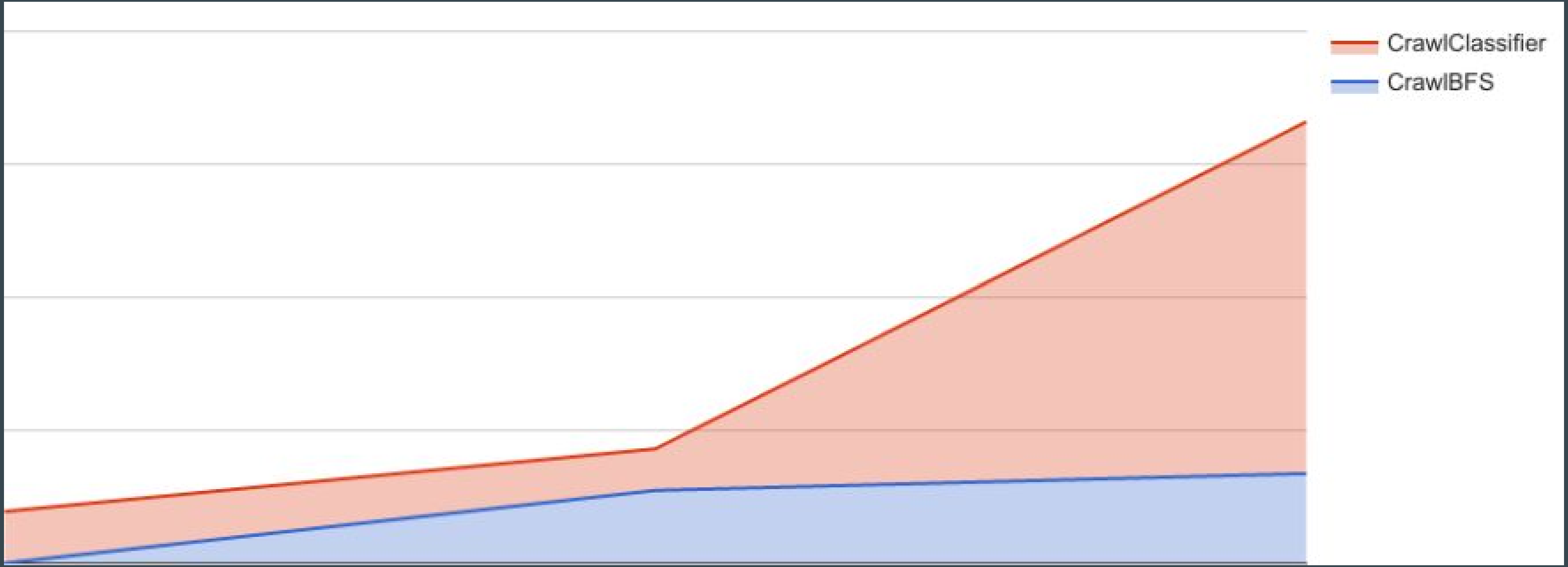
newegg com Product Product aspx Item N88E16837084665

newegg com Product Product aspx Item N12E1683370437

difícil classificação em alguns domínios

	CrawlBFS	CrawlClassifier
5min	9.3457947E-4	0.07780847
10min	0.0010958904	0.06266596
15min	0.0013495276	0.052870676

harvest ratio



harvest ratio

Classifier

•••

Pedro Henrique

Páginas classificadas manualmente

140 páginas negativas (divididas em outros conjuntos para o classificador do crawler)

134 páginas positivas

Extração de atributos das páginas

Vetor de atributos formado de todas palavras distintas de todas as páginas

Mais de 10000 palavras

Cada página é transformada em uma instância onde os valores do vetor são a frequência da palavra na página

Classificadores

7 classificadores diferentes

NaiveBayes, J48, SVM (SMO), Logistic, MLP, KNN(Ibk), RandomForest

7 configurações de filtros diferentes

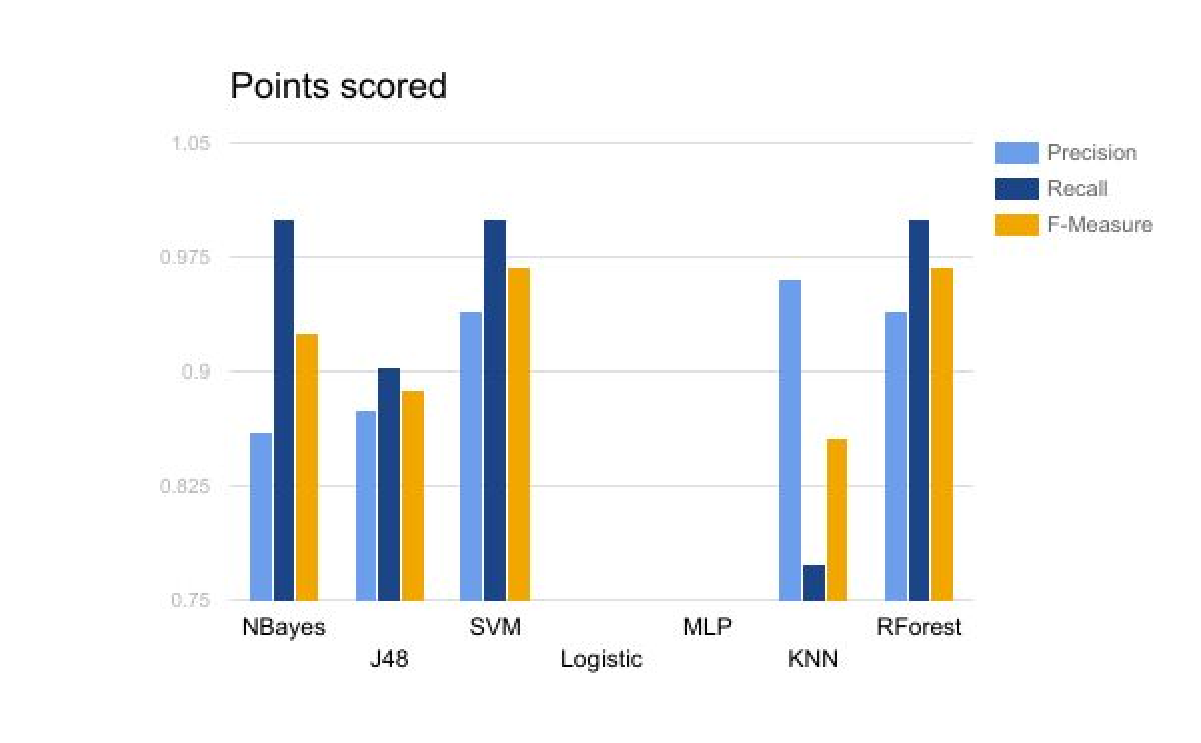
Sem filtros, Maior frequência (10, 50, 100), Ganho de informação (10, 50, 100)

Logistic e MLP não foram treinados sem filtro

47 classificadores treinados no total

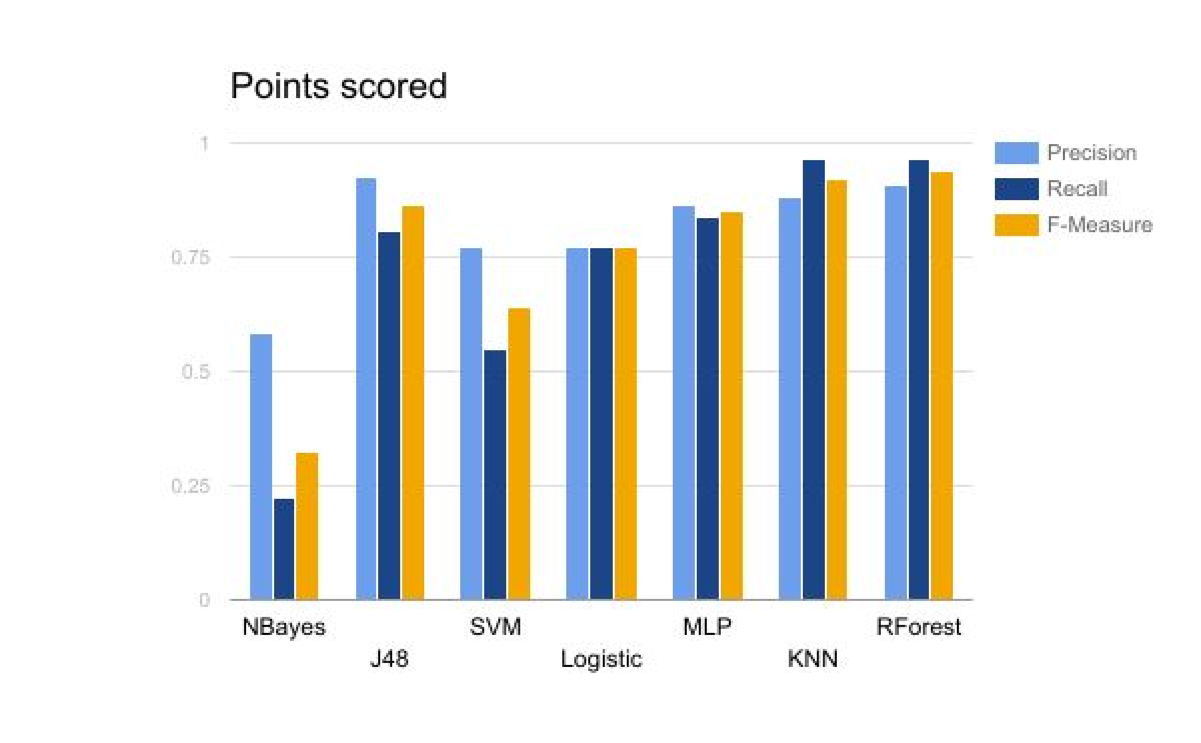
Classificadores

Sem filtros



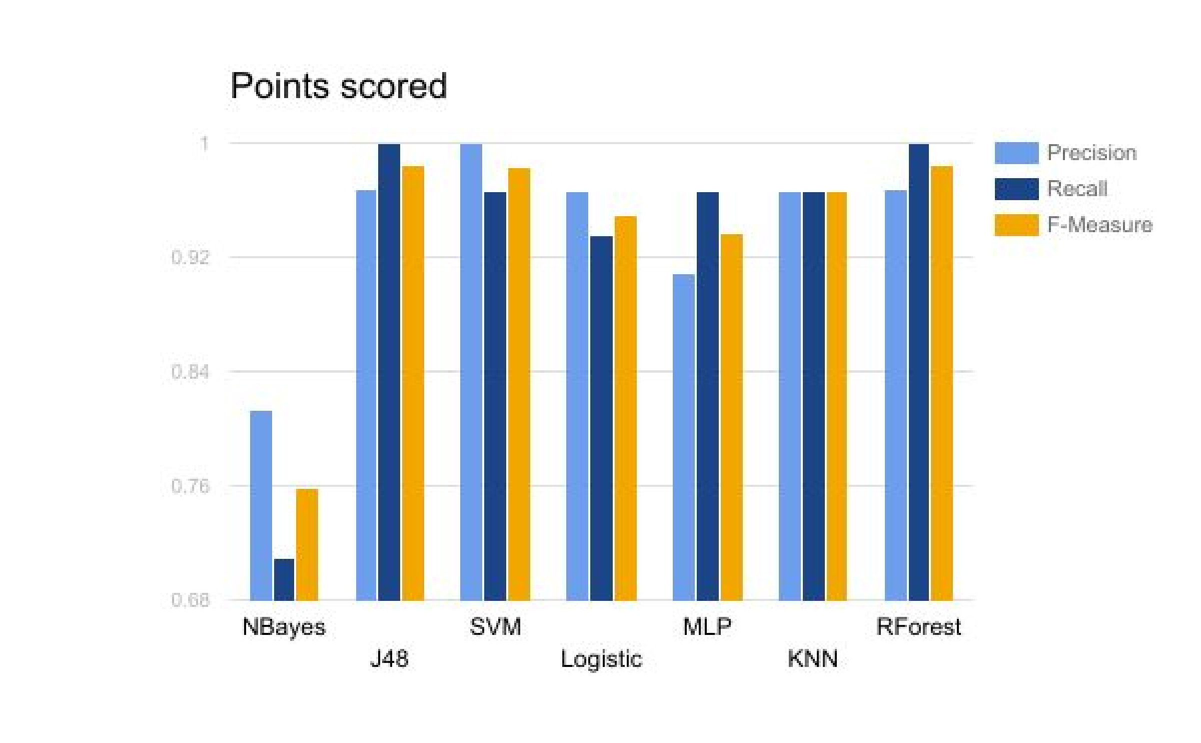
Classificadores

Maior frequência 10



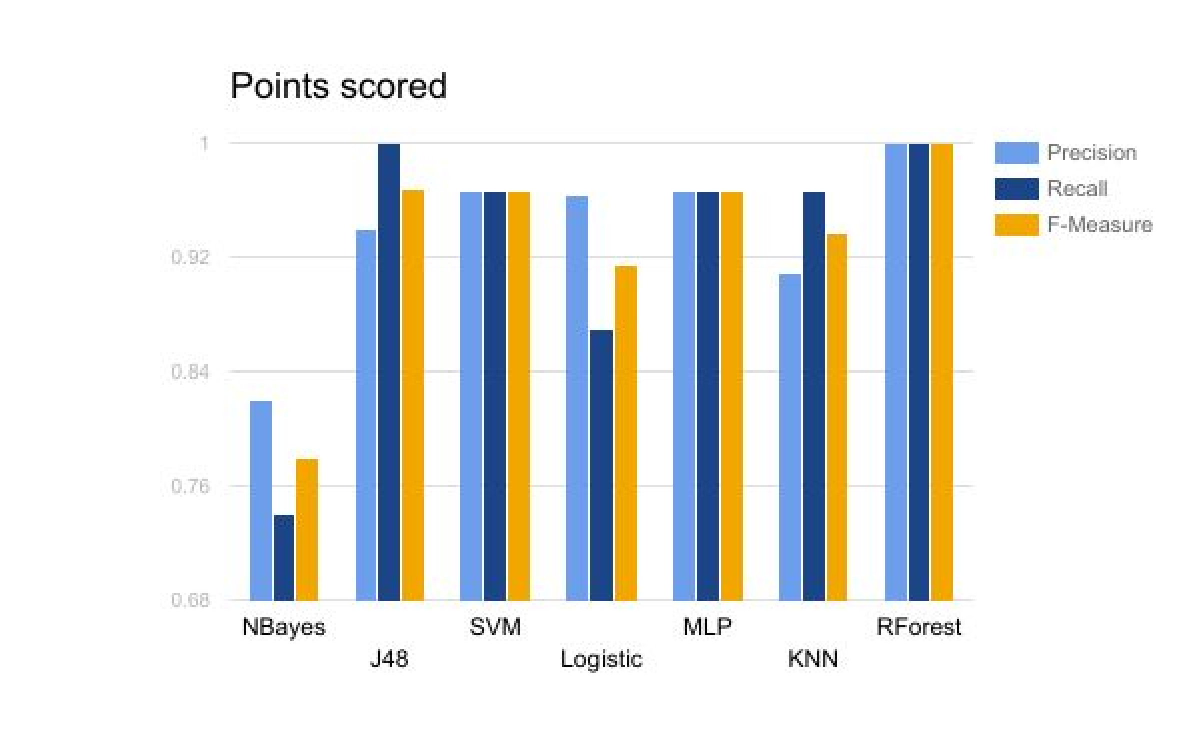
Classificadores

Maior frequência 50



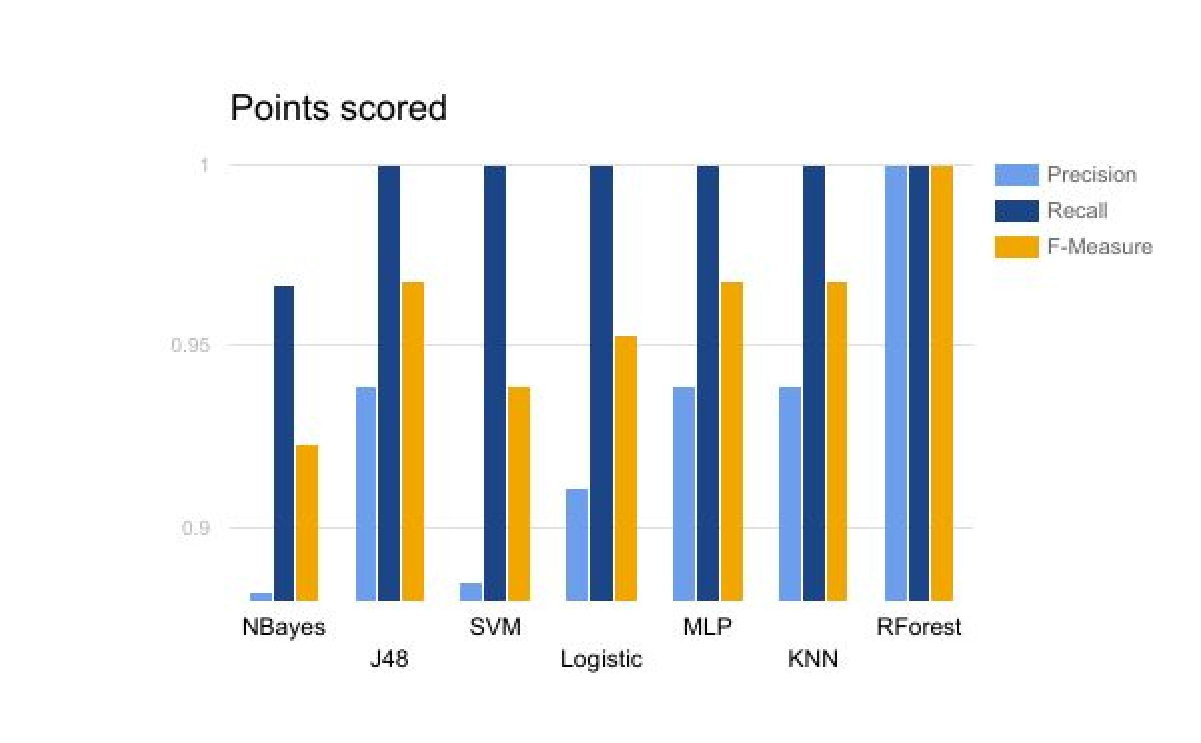
Classificadores

Maior frequência 100



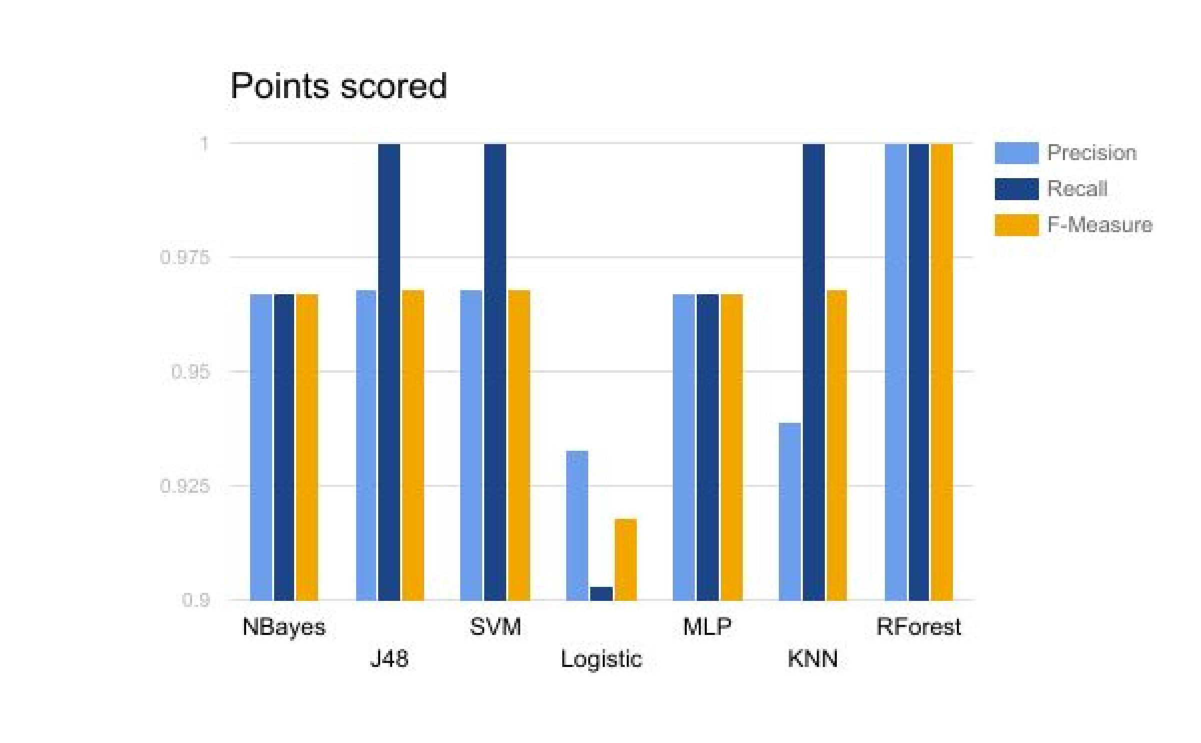
Classificadores

Ganho de Informação 10



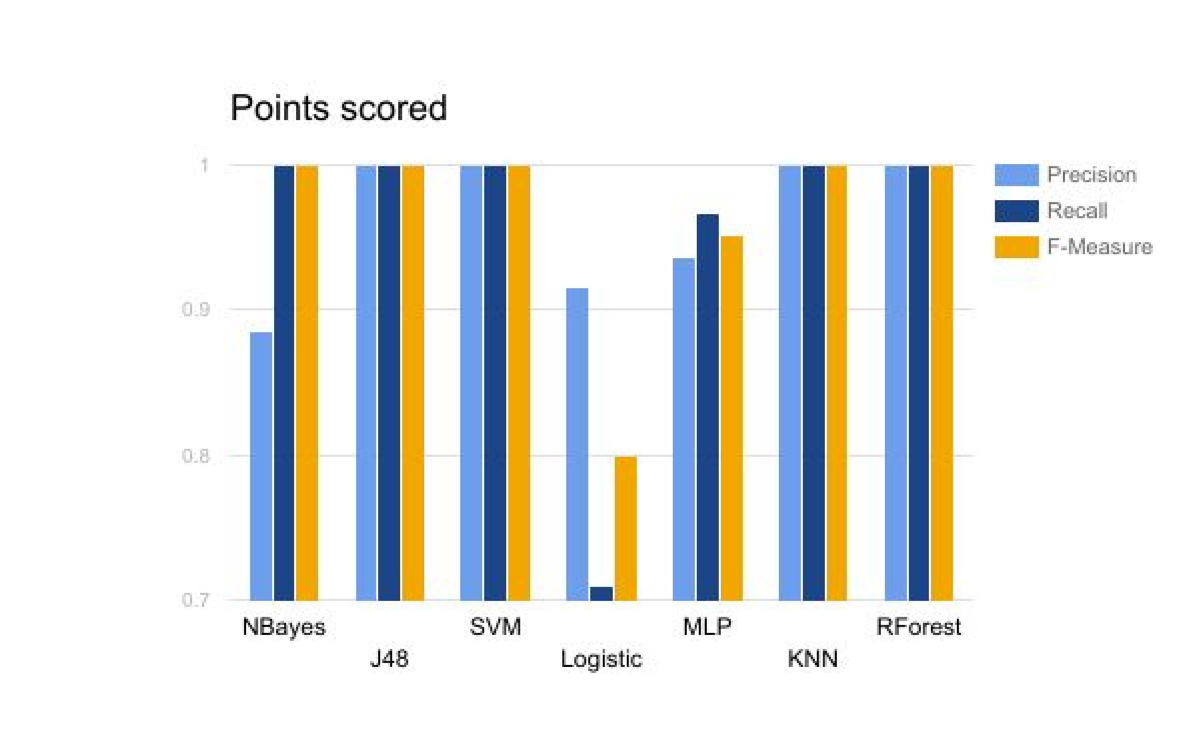
Classificadores

Ganho de Informação 50

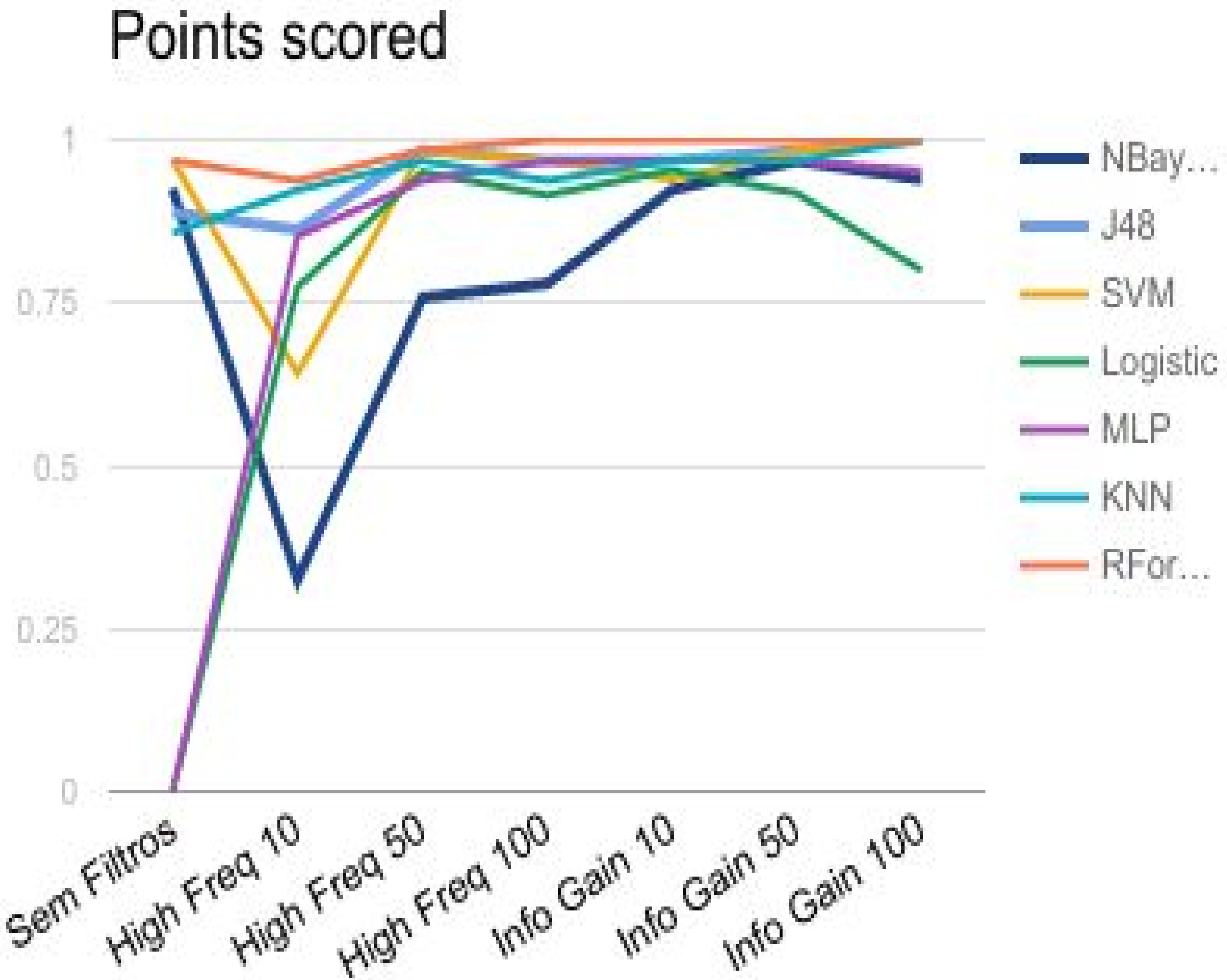


Classificadores

Ganho de Informação 100



Classificadores F-Measure



Extractor

•••

Guilherme Henrique

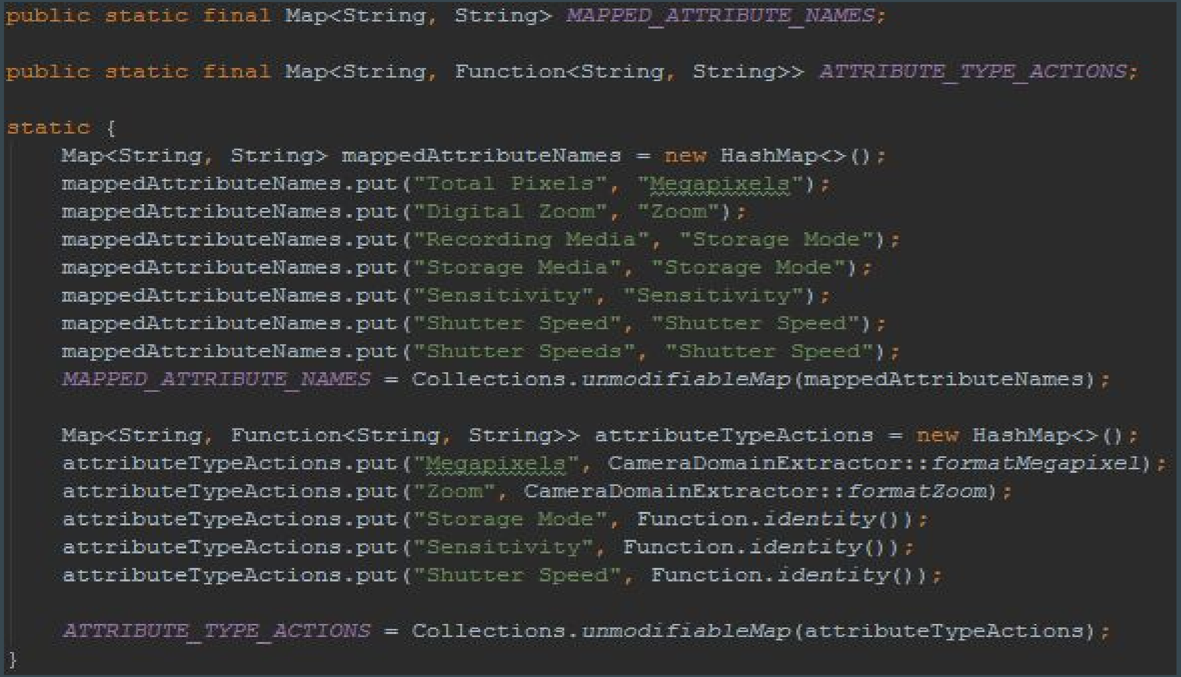
Atributos alvo da extração

- **Price**
- **Name**
- Megapixel
- Zoom
- Storage Mode
- Sensitivity
- Shutter Speed
- Sensor Size

Abordagens

- Específica
  - Atributos específicos com os valores
    - titles, trs, tls, divs, dls...
    - Cada extrator específico possui um mapa de possíveis atributos
- Geral
  - Busca em largura na DOM tree, verificando se o nó possui dois filhos
    - Se possuir, verifica se é um atributo que está na hash
      - Se está, o armazena
      - Senão, o descarta
    - A hash do extrator geral é a composição de todas as hashes dos extratores específicos

Mapa de atributos exemplo (Canon)



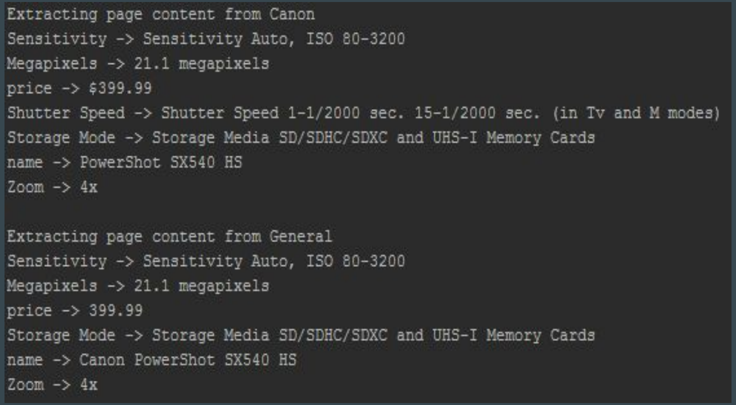
```java
public static final Map<String, String> MAPPED_ATTRIBUTE_NAMES;

public static final Map<String, Function<String, String>> ATTRIBUTE_TYPE_ACTIONS;

static {
    Map<String, String> mappedAttributeNames = new HashMap<>();
    mappedAttributeNames.put("Total Pixels", "Megapixels");
    mappedAttributeNames.put("Digital Zoom", "Zoom");
    mappedAttributeNames.put("Recording Media", "Storage Mode");
    mappedAttributeNames.put("Storage Media", "Storage Mode");
    mappedAttributeNames.put("Sensitivity", "Sensitivity");
    mappedAttributeNames.put("Shutter Speed", "Shutter Speed");
    mappedAttributeNames.put("Shutter Speeds", "Shutter Speed");
    MAPPED_ATTRIBUTE_NAMES = Collections.unmodifiableMap(mappedAttributeNames);

    Map<String, Function<String, String>> attributeTypeActions = new HashMap<>();
    attributeTypeActions.put("Megapixels", CameraDomainExtractor::formatMegapixel);
    attributeTypeActions.put("Zoom", CameraDomainExtractor::formatZoom);
    attributeTypeActions.put("Storage Mode", Function.identity());
    attributeTypeActions.put("Sensitivity", Function.identity());
    attributeTypeActions.put("Shutter Speed", Function.identity());

    ATTRIBUTE_TYPE_ACTIONS = Collections.unmodifiableMap(attributeTypeActions);
}
```

Cálculo das métricas

- Todas as instâncias positivas do classificador
- Métricas
  - Precision
  - Recall
  - F-Measure

Canon

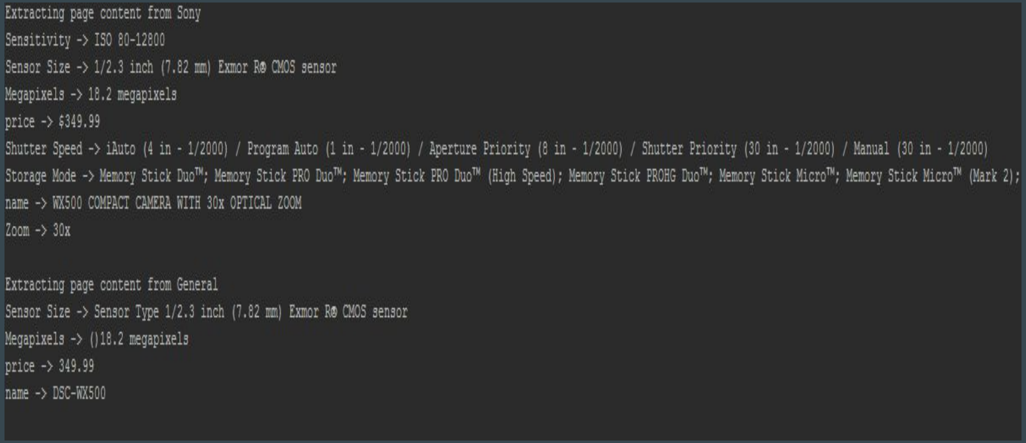
- 7 de 8 atributos extraídos
  - Não possui informações sobre o tamanho do sensor (sensor size)
- Métricas
  - Específico:
    - Precision: 1
    - Recall : 0.7678571429
    - F-Measure: 0.8686868687
  - Geral
    - Precision: 1
    - Recall: 0.53125
    - F-Measure: 0.6938775510204082



```
Extracting page content from Canon
Sensitivity -> Sensitivity Auto, ISO 80-3200
Megapixels -> 21.1 megapixels
price -> $399.99
Shutter Speed -> Shutter Speed 1-1/2000 sec. 15-1/2000 sec. (in Tv and M modes)
Storage Mode -> Storage Media SD/SDHC/SDXC and UHS-I Memory Cards
name -> PowerShot SX540 HS
Zoom -> 4x

Extracting page content from General
Sensitivity -> Sensitivity Auto, ISO 80-3200
Megapixels -> 21.1 megapixels
price -> 399.99
Storage Mode -> Storage Media SD/SDHC/SDXC and UHS-I Memory Cards
name -> Canon PowerShot SX540 HS
Zoom -> 4x
```

Sony

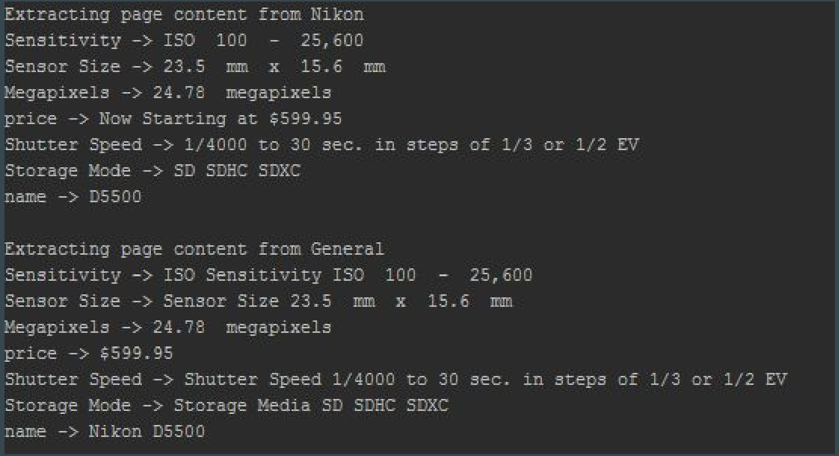
- 8 de 8 atributos extraídos
- Métricas
  - Específico:
    - Precision: 1
    - Recall : 0.8333333333333334
    - F-Measure: 0.9090909090
  - Geral
    - Precision: 1
    - Recall: 0.46875
    - F-Measure: 0.63829787234042



```
Extracting page content from Sony
Sensitivity -> ISO 80-12800
Sensor Size -> 1/2.3 inch (7.82 mm) Exmor R® CMOS sensor
Megapixels -> 18.2 megapixels
price -> $349.99
Shutter Speed -> iAuto (4 in - 1/2000) / Program Auto (1 in - 1/2000) / Aperture Priority (8 in - 1/2000) / Shutter Priority (30 in - 1/2000) / Manual (30 in - 1/2000)
Storage Mode -> Memory Stick Duo™; Memory Stick PRO Duo™; Memory Stick PRO Duo™ (High Speed); Memory Stick PROHG Duo™; Memory Stick Micro™; Memory Stick Micro™ (Mark 2);
name -> WX500 COMPACT CAMERA WITH 30x OPTICAL ZOOM
Zoom -> 30x

Extracting page content from General
Sensor Size -> Sensor Type 1/2.3 inch (7.82 mm) Exmor R® CMOS sensor
Megapixels -> ()18.2 megapixels
price -> 349.99
name -> DSC-WX500
```

Nikon

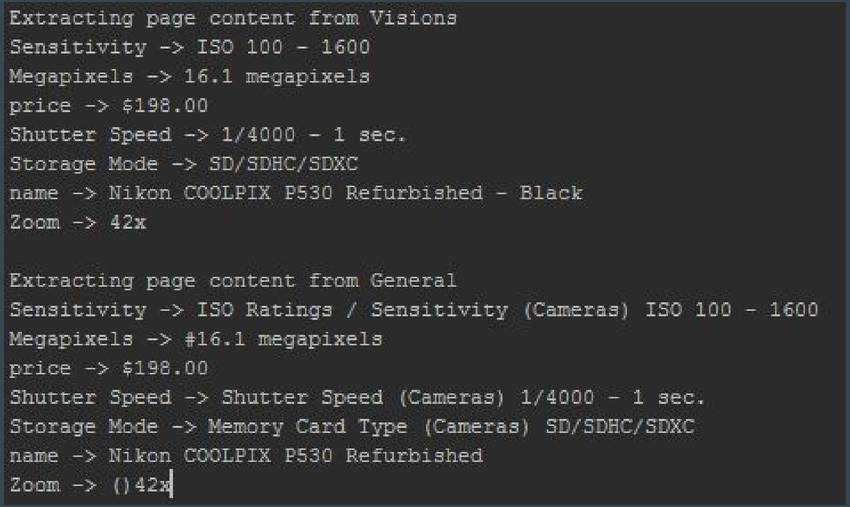
- 7 de 8 atributos extraídos
    - Não possui informações sobre de Zoom
- Métricas
    - Específico:
        - Precision: 1
        - Recall: 0.8392857142857143
        - F-measure: 0.9126213592233009
    - Geral
        - Precision: 1
        - Recall: 0.8392857142857143
        - F-measure: 0.9126213592233009



```
Extracting page content from Nikon
Sensitivity -> ISO  100  -  25,600
Sensor Size -> 23.5  mm  x  15.6  mm
Megapixels -> 24.78  megapixels
price -> Now Starting at $599.95
Shutter Speed -> 1/4000 to 30 sec. in steps of 1/3 or 1/2 EV
Storage Mode -> SD SDHC SDXC
name -> D5500

Extracting page content from General
Sensitivity -> ISO Sensitivity ISO  100  -  25,600
Sensor Size -> Sensor Size 23.5  mm  x  15.6  mm
Megapixels -> 24.78  megapixels
price -> $599.95
Shutter Speed -> Shutter Speed 1/4000 to 30 sec. in steps of 1/3 or 1/2 EV
Storage Mode -> Storage Media SD SDHC SDXC
name -> Nikon D5500
```

Visions

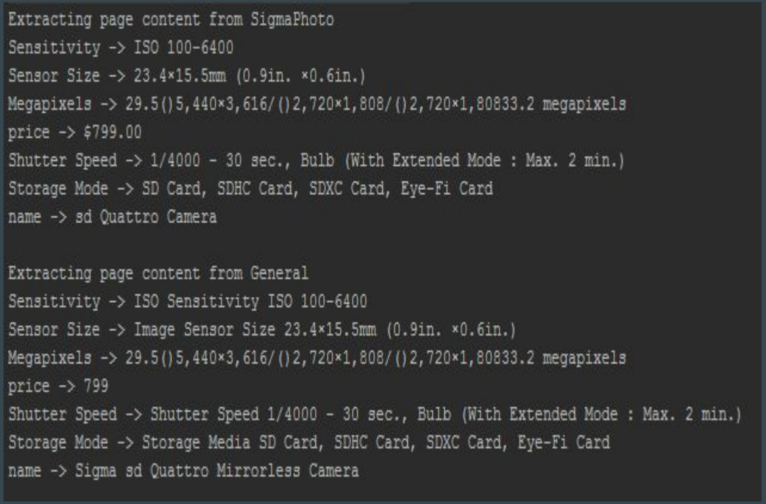
- 7 de 8 atributos extraídos
  - Não possui informações sobre o tamanho do sensor (sensor size)
- Métricas
  - Específico:
    - Precision: 1
    - Recall: 1
    - F-measure: 1
  - Geral
    - Precision: 1
    - Recall: 0.875
    - F-measure: 0.9333333333333333



```
Extracting page content from Visions
Sensitivity -> ISO 100 - 1600
Megapixels -> 16.1 megapixels
price -> $198.00
Shutter Speed -> 1/4000 - 1 sec.
Storage Mode -> SD/SDHC/SDXC
name -> Nikon COOLPIX P530 Refurbished - Black
Zoom -> 42x

Extracting page content from General
Sensitivity -> ISO Ratings / Sensitivity (Cameras) ISO 100 - 1600
Megapixels -> #16.1 megapixels
price -> $198.00
Shutter Speed -> Shutter Speed (Cameras) 1/4000 - 1 sec.
Storage Mode -> Memory Card Type (Cameras) SD/SDHC/SDXC
name -> Nikon COOLPIX P530 Refurbished
Zoom -> ()42x
```

Sigma Photo

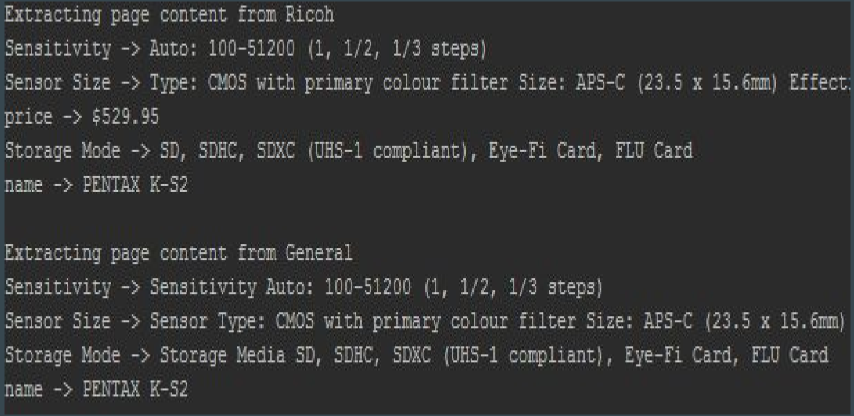
- 7 de 8 atributos extraídos
    - Não possui informações sobre o zoom
- Métricas
    - Específico:
        - Precision: 1
        - Recall: 0.638888888888888
        - F-measure: 0.7796610169491525
    - Geral
        - Precision: 1
        - Recall: 0.638888888888888
        - F-measure: 0.7796610169491525



```
Extracting page content from SigmaPhoto
Sensitivity -> ISO 100-6400
Sensor Size -> 23.4×15.5mm (0.9in. ×0.6in.)
Megapixels -> 29.5()5,440×3,616/()2,720×1,808/()2,720×1,80833.2 megapixels
price -> $799.00
Shutter Speed -> 1/4000 - 30 sec., Bulb (With Extended Mode : Max. 2 min.)
Storage Mode -> SD Card, SDHC Card, SDXC Card, Eye-Fi Card
name -> sd Quattro Camera

Extracting page content from General
Sensitivity -> ISO Sensitivity ISO 100-6400
Sensor Size -> Image Sensor Size 23.4×15.5mm (0.9in. ×0.6in.)
Megapixels -> 29.5()5,440×3,616/()2,720×1,808/()2,720×1,80833.2 megapixels
price -> 799
Shutter Speed -> Shutter Speed 1/4000 - 30 sec., Bulb (With Extended Mode : Max. 2 min.)
Storage Mode -> Storage Media SD Card, SDHC Card, SDXC Card, Eye-Fi Card
name -> Sigma sd Quattro Mirrorless Camera
```

Ricoh

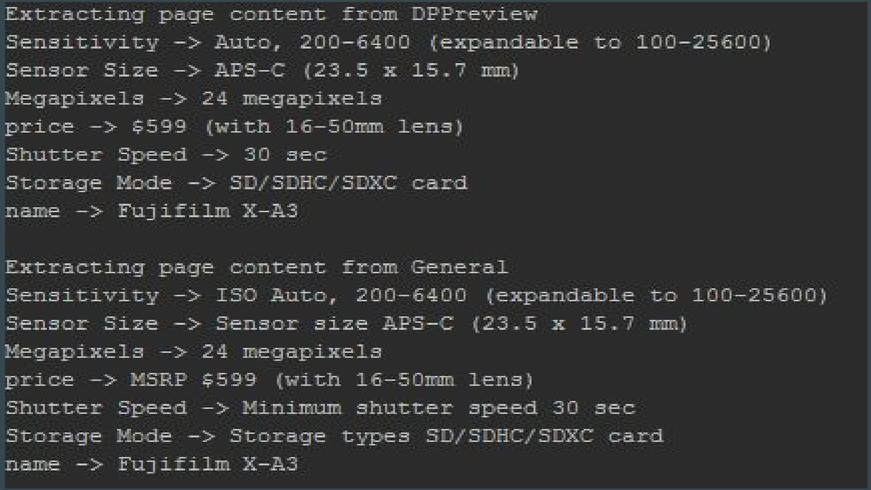
- 6 de 8 atributos extraídos
  - Não possui informações sobre o tamanho do sensor (sensor size) e zoom
- Métricas
  - Específico:
    - Precision: 1
    - Recall: 0.6477272727272727
    - F-measure: 0.7862068965517242
  - Geral
    - Precision: 1
    - Recall: 0.625
    - F-measure: 0.7692307692307693



```
Extracting page content from Ricoh
Sensitivity -> Auto: 100-51200 (1, 1/2, 1/3 steps)
Sensor Size -> Type: CMOS with primary colour filter Size: APS-C (23.5 x 15.6mm) Effect
price -> $529.95
Storage Mode -> SD, SDHC, SDXC (UHS-1 compliant), Eye-Fi Card, FLU Card
name -> PENTAX K-S2

Extracting page content from General
Sensitivity -> Sensitivity Auto: 100-51200 (1, 1/2, 1/3 steps)
Sensor Size -> Sensor Type: CMOS with primary colour filter Size: APS-C (23.5 x 15.6mm)
Storage Mode -> Storage Media SD, SDHC, SDXC (UHS-1 compliant), Eye-Fi Card, FLU Card
name -> PENTAX K-S2
```

DP Preview

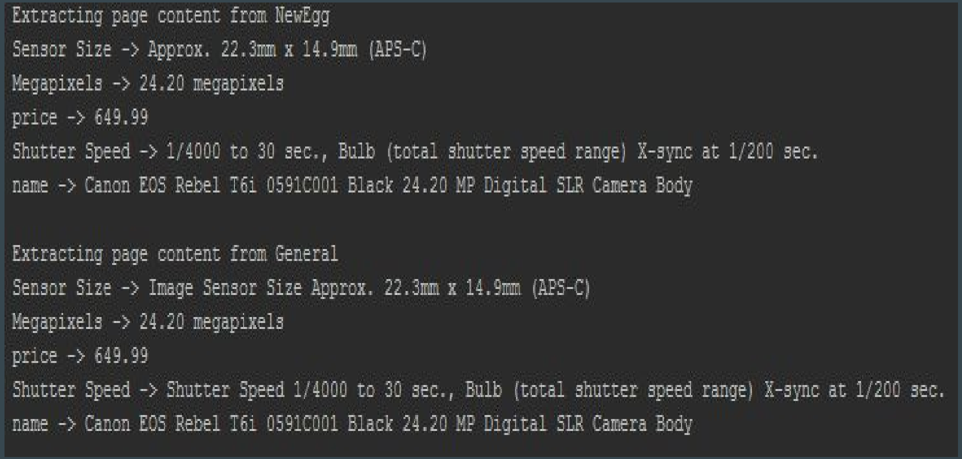
- 7 de 8 atributos extraídos
  - Não possui informações sobre o zoom
- Métricas
  - Específico:
    - Precision: 1
    - Recall: 0.895
    - F-measure: 0.9445910290237467
  - Geral
    - Precision: 1
    - Recall: 0.895
    - F-measure: 0.9445910290237467



```
Extracting page content from DPPreview
Sensitivity -> Auto, 200-6400 (expandable to 100-25600)
Sensor Size -> APS-C (23.5 x 15.7 mm)
Megapixels -> 24 megapixels
price -> $599 (with 16-50mm lens)
Shutter Speed -> 30 sec
Storage Mode -> SD/SDHC/SDXC card
name -> Fujifilm X-A3

Extracting page content from General
Sensitivity -> ISO Auto, 200-6400 (expandable to 100-25600)
Sensor Size -> Sensor size APS-C (23.5 x 15.7 mm)
Megapixels -> 24 megapixels
price -> MSRP $599 (with 16-50mm lens)
Shutter Speed -> Minimum shutter speed 30 sec
Storage Mode -> Storage types SD/SDHC/SDXC card
name -> Fujifilm X-A3
```

New Egg

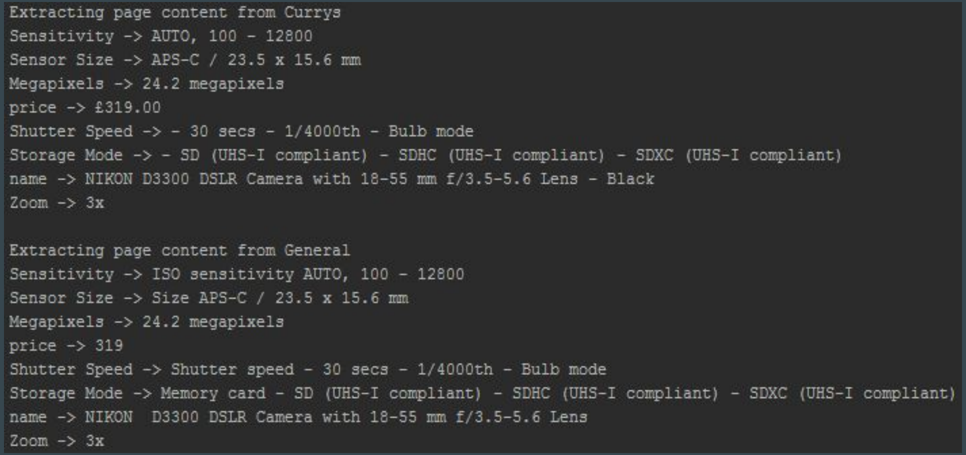
- 6 de 8 atributos extraídos
  - Não possui informações sobre o tamanho do sensor (sensor size) e zoom
- Métricas
  - Específico:
    - Precision: 1
    - Recall : 0.8545454545
    - F-Measure: 0.9215686275
  - Geral
    - Precision: 1
    - Recall: 0.7159090909090909
    - F-Measure: 0.8344370860927153



```
Extracting page content from NewEgg
Sensor Size -> Approx. 22.3mm x 14.9mm (APS-C)
Megapixels -> 24.20 megapixels
price -> 649.99
Shutter Speed -> 1/4000 to 30 sec., Bulb (total shutter speed range) X-sync at 1/200 sec.
name -> Canon EOS Rebel T6i 0591C001 Black 24.20 MP Digital SLR Camera Body

Extracting page content from General
Sensor Size -> Image Sensor Size Approx. 22.3mm x 14.9mm (APS-C)
Megapixels -> 24.20 megapixels
price -> 649.99
Shutter Speed -> Shutter Speed 1/4000 to 30 sec., Bulb (total shutter speed range) X-sync at 1/200 sec.
name -> Canon EOS Rebel T6i 0591C001 Black 24.20 MP Digital SLR Camera Body
```

Currys

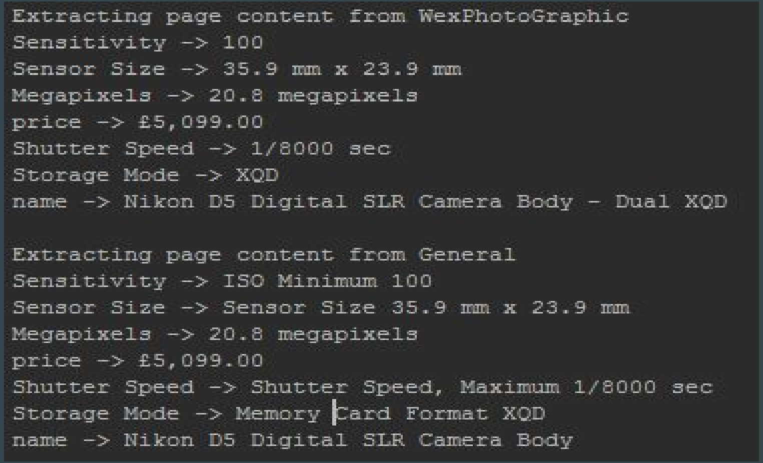
- 8 de 8 atributos extraídos
- Métricas
  - Específico:
    - Precision: 1
    - Recall: 0.9296875
    - F-measure: 0.9635627530364
  - Geral
    - Precision: 1
    - Recall: 0.9296875
    - F-measure: 0.9635627530364



```
Extracting page content from Currys
Sensitivity -> AUTO, 100 - 12800
Sensor Size -> APS-C / 23.5 x 15.6 mm
Megapixels -> 24.2 megapixels
price -> £319.00
Shutter Speed -> - 30 secs - 1/4000th - Bulb mode
Storage Mode -> - SD (UHS-I compliant) - SDHC (UHS-I compliant) - SDXC (UHS-I compliant)
name -> NIKON D3300 DSLR Camera with 18-55 mm f/3.5-5.6 Lens - Black
Zoom -> 3x

Extracting page content from General
Sensitivity -> ISO sensitivity AUTO, 100 - 12800
Sensor Size -> Size APS-C / 23.5 x 15.6 mm
Megapixels -> 24.2 megapixels
price -> 319
Shutter Speed -> Shutter speed - 30 secs - 1/4000th - Bulb mode
Storage Mode -> Memory card - SD (UHS-I compliant) - SDHC (UHS-I compliant) - SDXC (UHS-I compliant)
name -> NIKON  D3300 DSLR Camera with 18-55 mm f/3.5-5.6 Lens
Zoom -> 3x
```

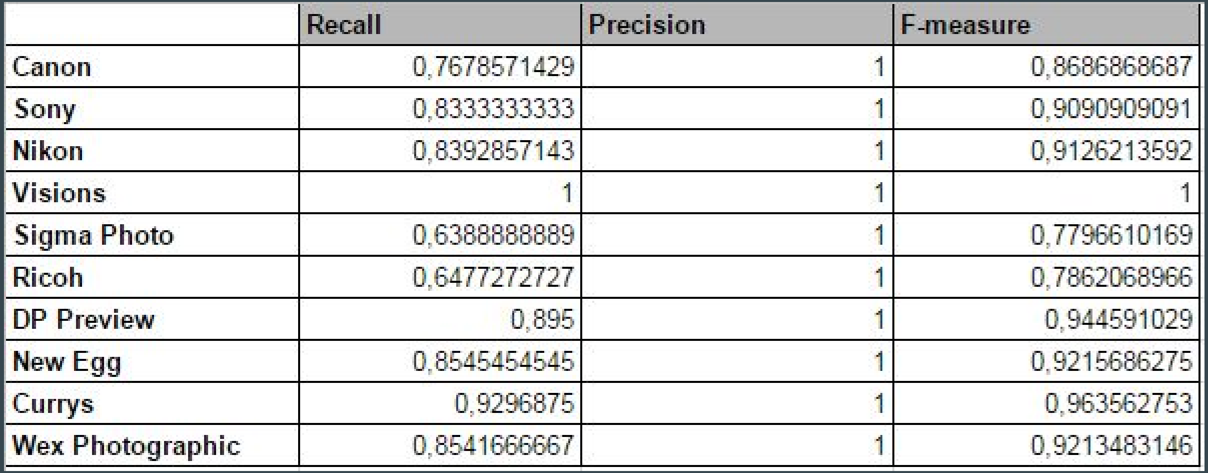
Wex Photographic

- 7 de 8 atributos extraídos
  - Não possui informações sobre zoom
- Métricas
  - Específico:
    - Precision: 1
    - Recall: 0.8541666666666666
    - F-measure: 0.9213483146067416
  - Geral
    - Precision: 1
    - Recall: 0.875
    - F-measure: 0.9333333333333333



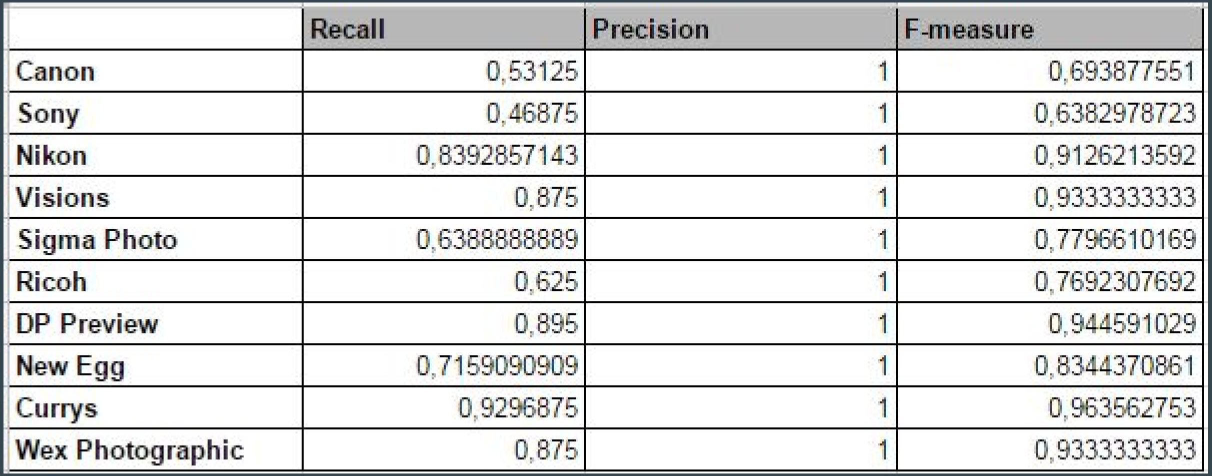
```
Extracting page content from WexPhotoGraphic
Sensitivity -> 100
Sensor Size -> 35.9 mm x 23.9 mm
Megapixels -> 20.8 megapixels
price -> £5,099.00
Shutter Speed -> 1/8000 sec
Storage Mode -> XQD
name -> Nikon D5 Digital SLR Camera Body - Dual XQD

Extracting page content from General
Sensitivity -> ISO Minimum 100
Sensor Size -> Sensor Size 35.9 mm x 23.9 mm
Megapixels -> 20.8 megapixels
price -> £5,099.00
Shutter Speed -> Shutter Speed, Maximum 1/8000 sec
Storage Mode -> Memory Card Format XQD
name -> Nikon D5 Digital SLR Camera Body
```

Extratores específicos - Resumo



	Recall	Precision	F-measure
Canon	0,7678571429	1	0,8686868687
Sony	0,8333333333	1	0,9090909091
Nikon	0,8392857143	1	0,9126213592
Visions	1	1	1
Sigma Photo	0,6388888889	1	0,7796610169
Ricoh	0,6477272727	1	0,7862068966
DP Preview	0,895	1	0,944591029
New Egg	0,8545454545	1	0,9215686275
Currys	0,9296875	1	0,963562753
Wex Photographic	0,8541666667	1	0,9213483146

Extrator Geral - Resumo



	Recall	Precision	F-measure
Canon	0,53125	1	0,693877551
Sony	0,46875	1	0,6382978723
Nikon	0,8392857143	1	0,9126213592
Visions	0,875	1	0,9333333333
Sigma Photo	0,6388888889	1	0,7796610169
Ricoh	0,625	1	0,7692307692
DP Preview	0,895	1	0,944591029
New Egg	0,7159090909	1	0,8344370861
Currys	0,9296875	1	0,963562753
Wex Photographic	0,875	1	0,9333333333