Recuperação de Informação

Leonardo Galdino (lcgm), Lucas Santana (lss5),
Bruno vitorino (bvcl)

Crawler

Objetivo: coletar o máximo possível de páginas relevantes dentro dos domínios

Duas abordagens:

- Baseline - (BFS)
- Heurística

Abordagem com BFS

- Primeira abordagem tentada

- Prós:
  - Simples de implementar
  - Relativamente rápido de obter as mil páginas

- Contras:
  - Apresentou resultados bem ruins

Abordagem com BFS

- Resultados

MLSSoccer	22 de 1000 => 0,022
Eurosport	0 de 1000 => 0
Soccerway	53 de 1000 => 0,053
FCTable	14 de 1000 => 0,014
FIFA	2 de 1000 => 0,002
Fbref	215 de 1000 => 0,215
UEFA	4 de 1000 => 0,004
Geral	0,044

Abordagem com BFS

- Algumas observações
  - As páginas relevantes dos sites da FIFA e da UEFA ficam em URLs bem profundas, dificultando bastante o trabalho do crawler com BFS
  - O site da Eurosport apresenta informações sobre vários esportes, então também dificulta o caminho do crawler até as páginas relevantes

Abordagem com Heurística

- Baseada em atribuir um peso às URLs encontradas e colocá-las na lista de URLs a serem exploradas dependendo desse peso.
  - Os pesos são dados com base em palavras presentes ou não nos links obtidos
  - Os pesos podem ser de -1 até 2, sendo 2 o melhor (forte evidência de uma página relevante) e -1 o pior (forte evidência de que o link levará a uma página não relevante)
  - URLs com pesos muito baixos nem eram inseridas na fila de URLs a serem exploradas (evita uso de memória e tempo de processamento para caminhos que não levariam a páginas relevantes)
  - Os outros links são inseridos em ordem, mantendo os links de peso 2 na frente da fila, seguidos pelos de peso 1 e por fim os de peso 0
- Prós:
  - Resultados bem melhores em relação à abordagem com BFS
- Contras:
  - Mais complexo de implementar
  - Mais demorado para coletar páginas

Abordagem com Heurística

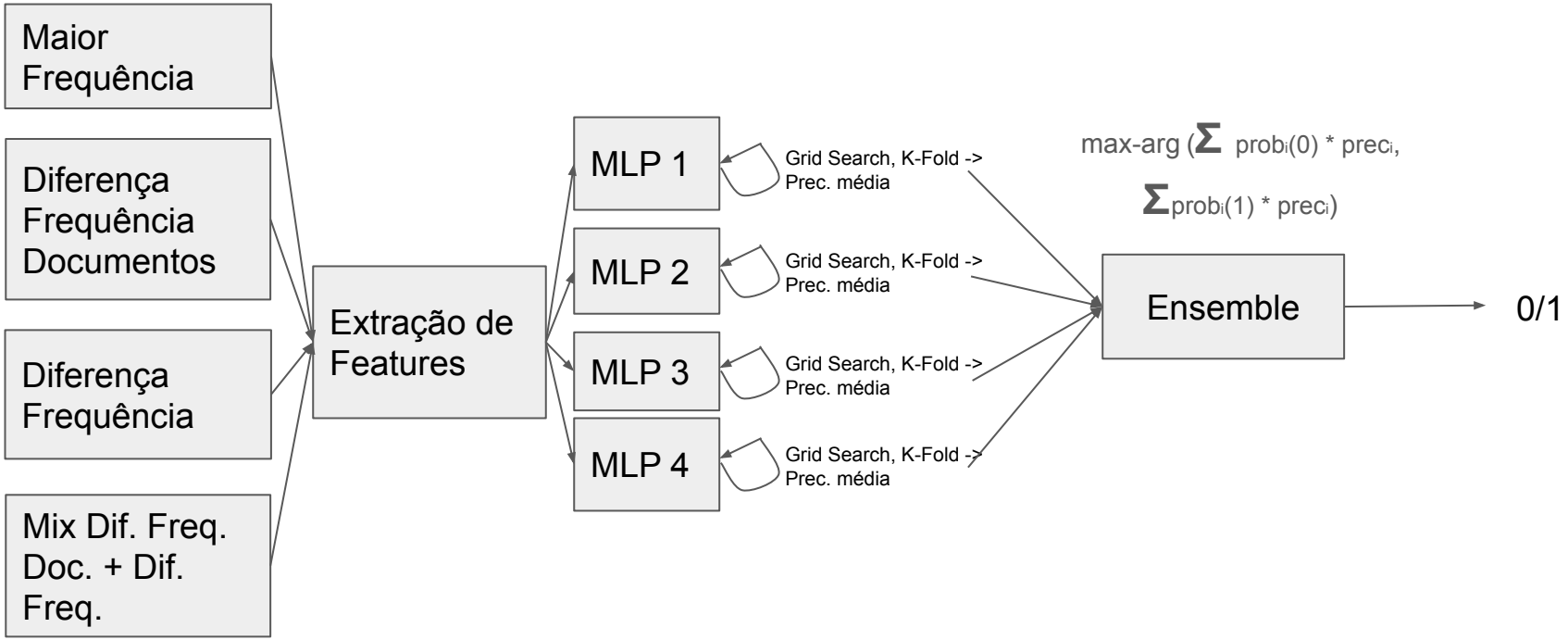
- Resultados

Sites	Heurística	BFS
MLSSoccer	717 de 1000 => 0,717	22 de 1000 => 0,022
Eurosport	1000 de 1000 => 1	0 de 1000 => 0
Soccerway	893 de 1000 => 0,893	53 de 1000 => 0,053
FCTable	999 de 1000 => 0,999	14 de 1000 => 0,014
FIFA	976 de 1000 => 0,976	2 de 1000 => 0,002
Fbref	998 de 1000 => 0,998	215 de 1000 => 0,215
UEFA	1000 de 1000 => 1	4 de 1000 => 0,004
Geral	0,94	0,044

Abordagem com Heurística

- Algumas observações
  - A heurística utilizada conseguiu guiar bem melhor o crawler pelos sites, apresentando um resultado bem satisfatório
  - Em alguns sites o ganho após utilizar a heurística foi muito grande, por exemplo, para o site da Eurosport a BFS conseguiu 0 páginas relevantes enquanto a heurística conseguiu 1000.

Classificador



Classificador - Resultados

```
Metrics:
Confusion matrix:
[22, 0]
[2, 21]
Accuracy: 0.9555555555555556
Precision: 0.9130434782608695
Recall: 1.0
F1-Measure: 0.9545454545454545
Training took 269.181388 seconds
```

```
Metrics:
Confusion matrix:
[22, 0]
[2, 21]
Accuracy: 0.9555555555555556
Precision: 0.9130434782608695
Recall: 1.0
F1-Measure: 0.9545454545454545
Training took 261.539163 seconds
```

```
Metrics:
Confusion matrix:
[23, 0]
[1, 21]
Accuracy: 0.9777777777777777
Precision: 0.9545454545454546
Recall: 1.0
F1-Measure: 0.9767441860465117
Training took 305.662778 seconds
```

Extrator

- Número de Extratores Específicos: 7
  - Eurosport, Fbref, FCTables, FIFA, MLSSoccer, Soccerway, UEFA
- Quantidade total de atributos extraídos: 11
- Gera um dicionário para cada instância
- Distribuição de Presença:

\	name	position	age	birthdate	birthplace	nationality	height	weight	team	number	foot
Eurosport	X	X	X	X	X	X	X	X			
Fbref	X	X		X	X	X	X	X	X		X
FCTables	X	X	X	X		X	X	X	X	X	
FIFA	X	X	X	X		X	X			X	
MLSSoccer	X	X	X	X	X		X	X	X	X	
Soccerway	X	X	X	X	X	X	X	X			X
UEFA	X	X	X	X		X	X	X	X		

Extrator

- Extrator Independente de Domínio:
  - Uso dos metadados principais como semente para busca na DOM Tree. (Posição)
  - Número de passos médio para atingir a menor subárvore que engloba todos os metadados presentes no domínio: 4
    - Outliers: FIFA (6) e UEFA (5) [metadados divididos em duas subárvores]
  - Limite de passos para cima: 6 ou até chegar em body
  - A partir da raíz descer:
    - Até folhas (problemas com tags sem filhos. ex: <img>
    - Até strings (strings que não representam metadados da instância)
  - Solução pensada até o momento:
    - Determinar expressões regulares principais: chaves (ex: ‘[H|h]eight[:]?’) e valores (ex: ‘\d+cm|\d+\W{2}\d+”$’
    - Uma chave indica um valor na vizinhança (irmãos, sobrinhos), um valor é extraído para o dicionário de metadados da instância