



Coleta e Extração de Dados na Web

Luciano Barbosa



UNIVERSIDADE
FEDERAL
DE PERNAMBUCO



Roteiro da Aula

- Dados existentes na Web
- Coleta de dados
- Extração de dados estruturados



Dados na Web

- APIs (<https://www.pythonforbeginners.com/api/list-of-python-apis>)
- Dados abertos
 - <http://dados.recife.pe.gov.br/>
 - <https://opendata.cityofnewyork.us/>
- Listas de datasets:
 - <https://github.com/awesomedata/awesome-public-datasets>
 - <https://www.forbes.com/sites/bernardmarr/2016/02/12/big-data-35-brilliant-and-free-data-sources-for-2016/#525c0c58b54d>



Coletando Dados de Páginas HTML

- Originalmente criada para engenhos de busca
- Aumentar cobertura da base
- Manter cópias sincronizadas com as online

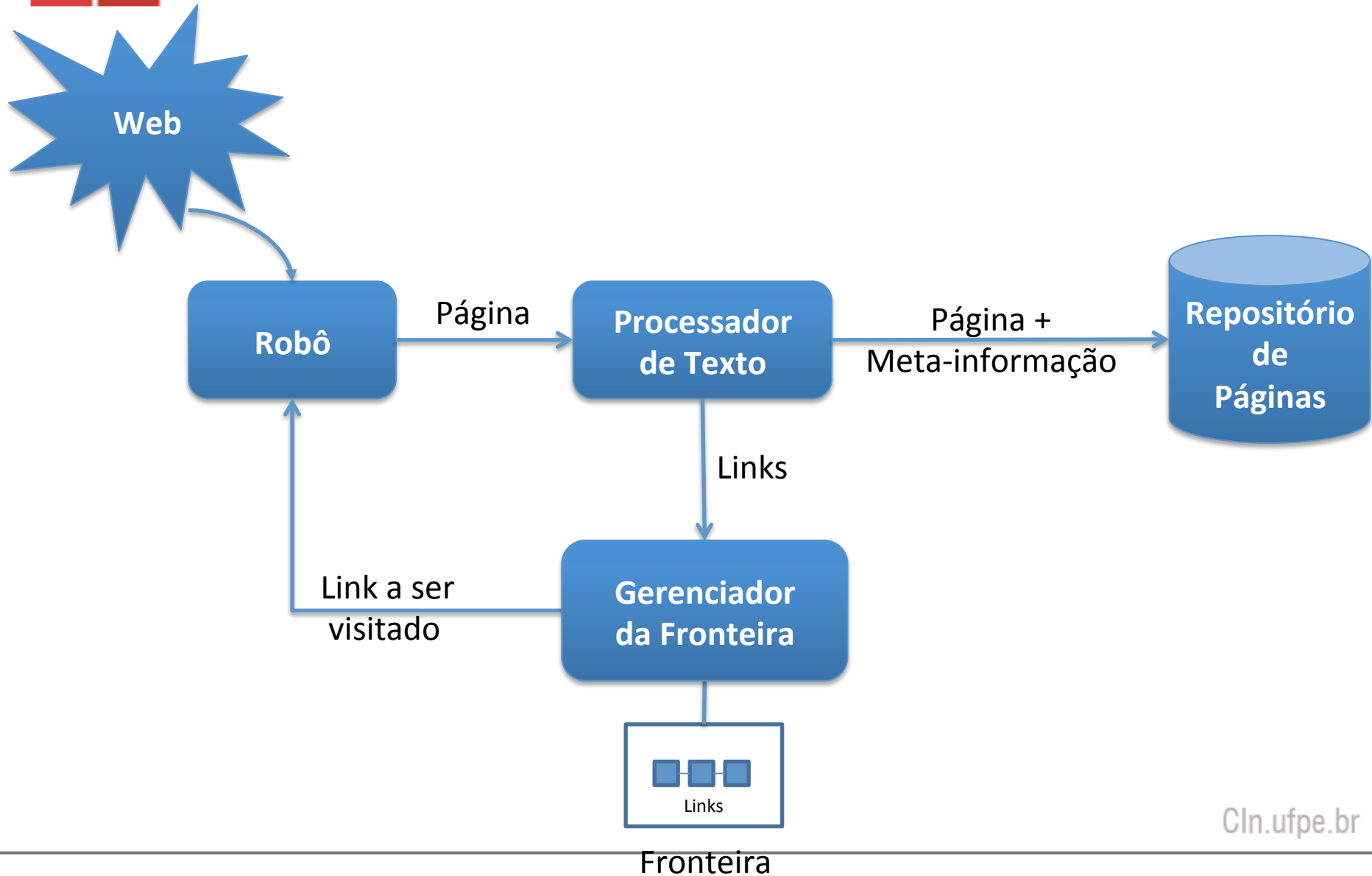


Breve História

- Primeiro crawler criado em 1993 por um estudante do MIT
 - Projeto WWW (World Wide Web Wanderer)
 - Usado para computar estatísticas sobre a Web
- Em junho de 1994, um estudante da UofW cria o WebCrawler
 - Cria índice pra consulta com cerca de 4 mil sites
 - Sucesso comercial
- Outros engenhos de busca baseados em coletores: Lycos (1994), Excite (1995), Altavista (1995) e HotBot (1996)
- Atualmente usado por todos grandes engenhos de busca



Funcionamento de um Coletor Geral





Aspectos a Considerar na Construção de um Coletor

Escalabilidade

**Seleção de
Conteúdo**

**Sobrecarga
em Sites**

**Conteúdo
Adversário**



Aspectos a Considerar na Construção de um Coletor

Escalabilidade

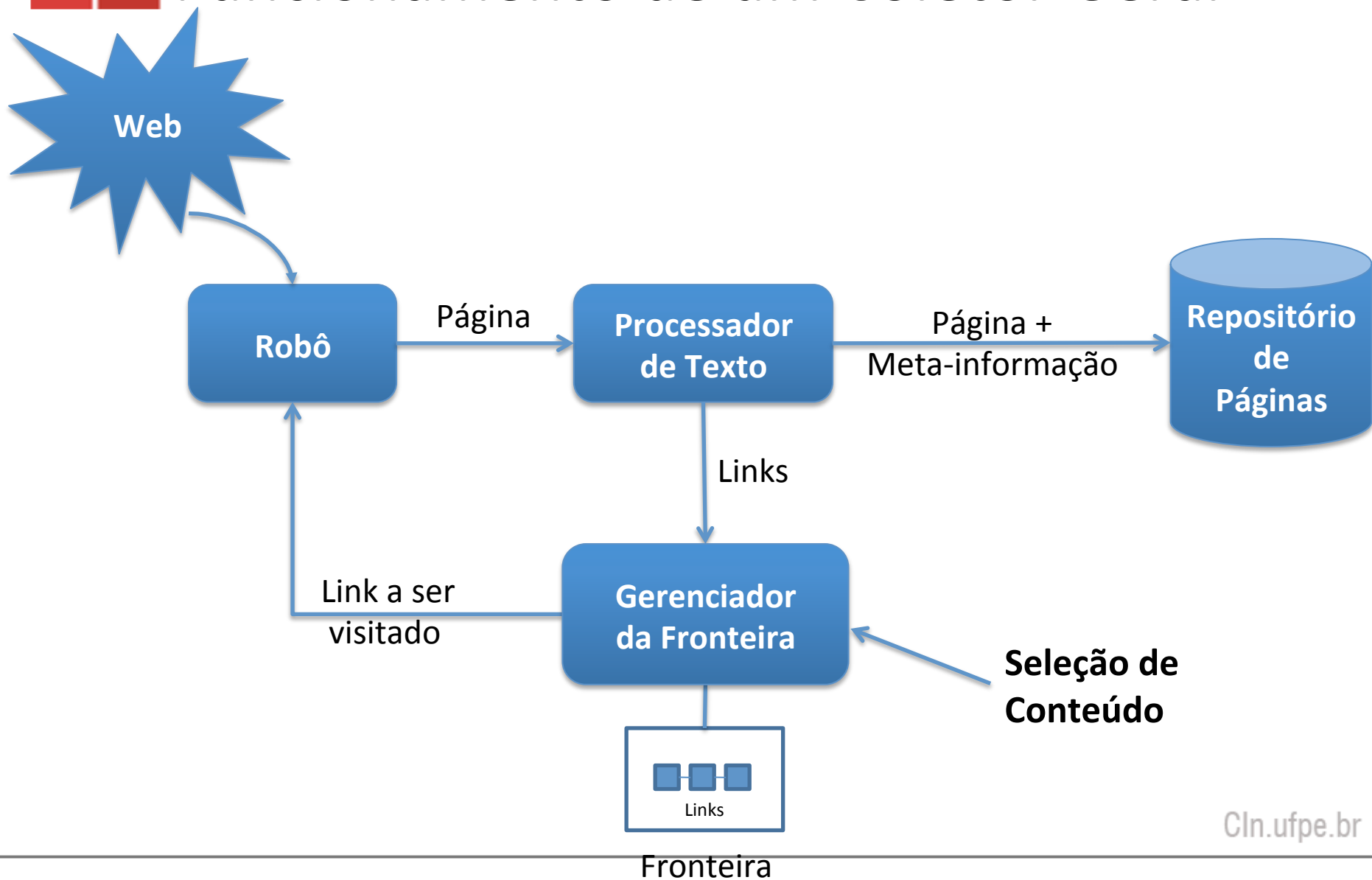
**Seleção de
Conteúdo**

**Sobrecarga
em Sites**

**Conteúdo
Adversário**



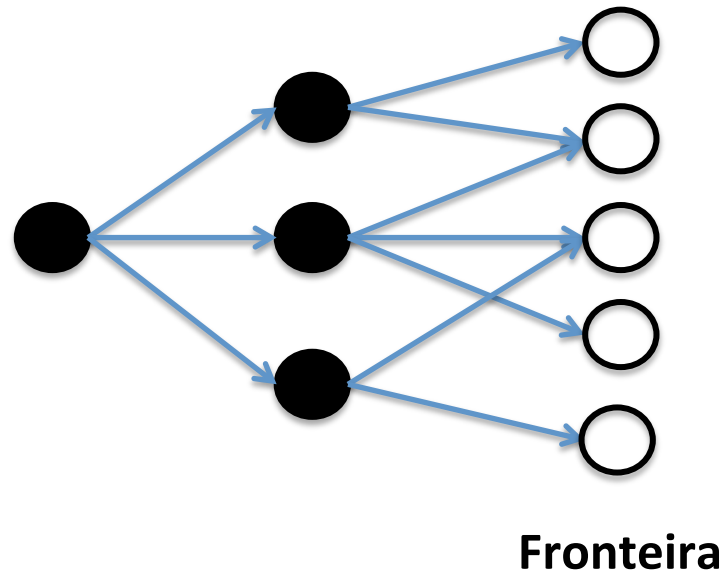
Funcionamento de um Coletor Geral





Seleção de Conteúdo

- Motivação: recursos limitados
- Prorizar links a serem visitados (fronteira)





Coletor Focado

- Objetivo: coletar dados em um dado tópico ou domínio
- Exemplo de aplicações: busca vertical jurídica

The screenshot shows the JusBrasil website interface. At the top, there is a search bar with the text 'eleição' and a magnifying glass icon. To the right of the search bar is a 'Menu' button. Below the search bar, there is a navigation bar with links: 'Tudo', 'Notícias', 'Artigos', 'Jurisprudência', 'Diários', 'Legislação', and 'Modelos e peças'. Below the navigation bar, it says 'Página 1 de 1.771.005 resultados para "eleição"'. The main content area shows a definition of 'Eleições' (Sinônimo de Eleição) with a small icon of a ballot box. Below this, there are two article snippets. The first snippet is titled '2014 - Eleições e Copa do Mundo.' and discusses the general elections held after the World Cup. The second snippet is titled 'Voto nulo e novas eleições' and discusses the need for new elections if a large number of votes are null.

JusBrasil eleição

Tudo Notícias Artigos Jurisprudência Diários Legislação Modelos e peças

Página 1 de 1.771.005 resultados para "eleição"

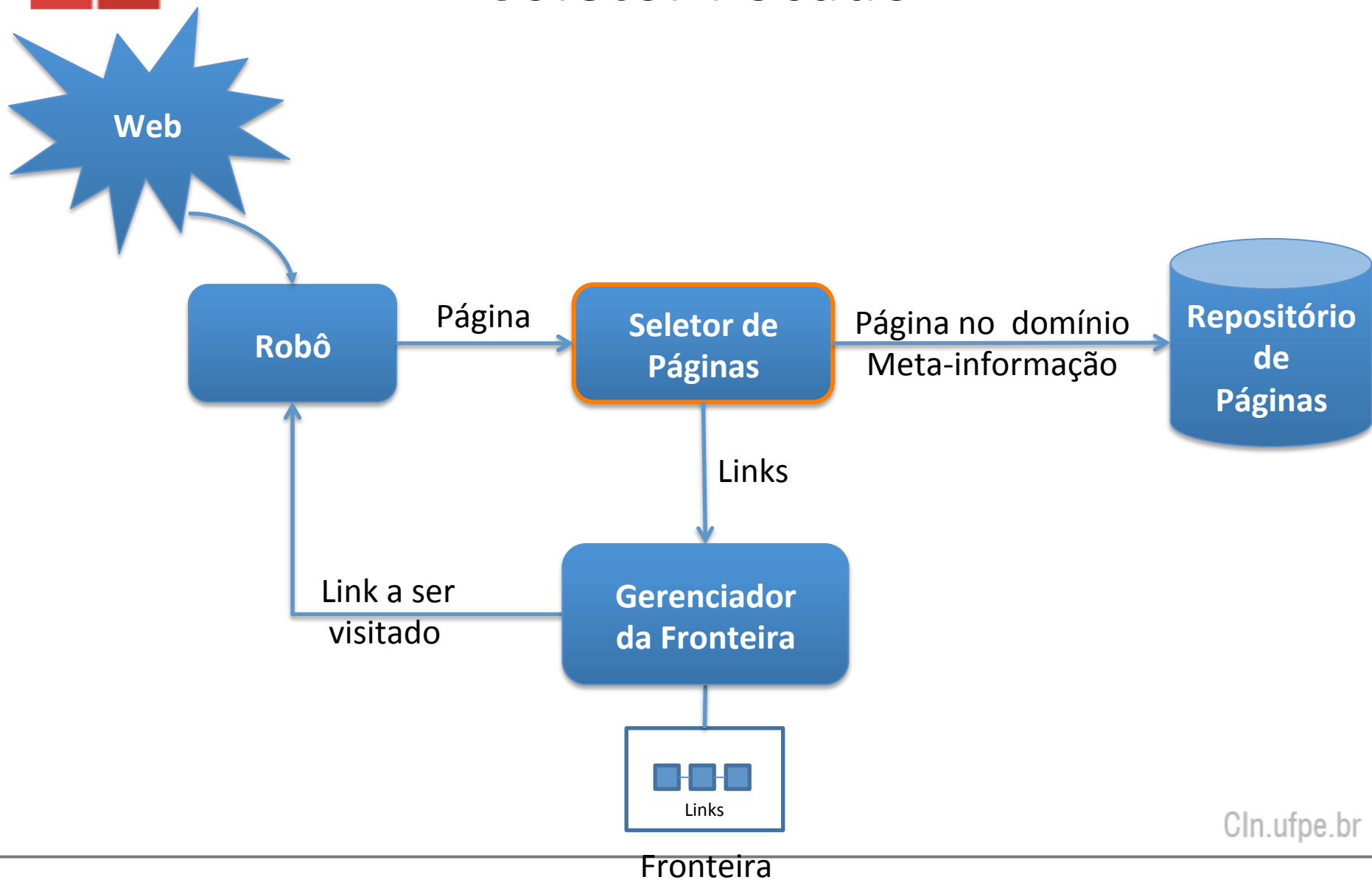
Eleições (Sinônimo de Eleição)
Ato pelo qual o povo escolhe mediante sufrágio ou aclamação uma ou mais pessoas que recebem a delegação de representá-lo, exercendo determinada função.
Tópico • 31 seguidores

2014 - Eleições e Copa do Mundo.
que as eleições gerais, que será realizada logo após a Copa do Mundo, sejam decididas não pelos quase 142 milhões... para a eleição presidencial do país e de alguns governos estaduais, a exemplo de Mato Grosso. Nós
Artigo • Antonio • 04/06/2014

Voto nulo e novas eleições
De dois em dois anos, em eleições municipais ou regionais, sempre surge alguém para hastear a necessidade de marcação de nova eleição se a nulidade atingir mais de metade dos votos do país... decorre da constatação de fraude nas ...
Artigo • Danielli • 11/09/2014

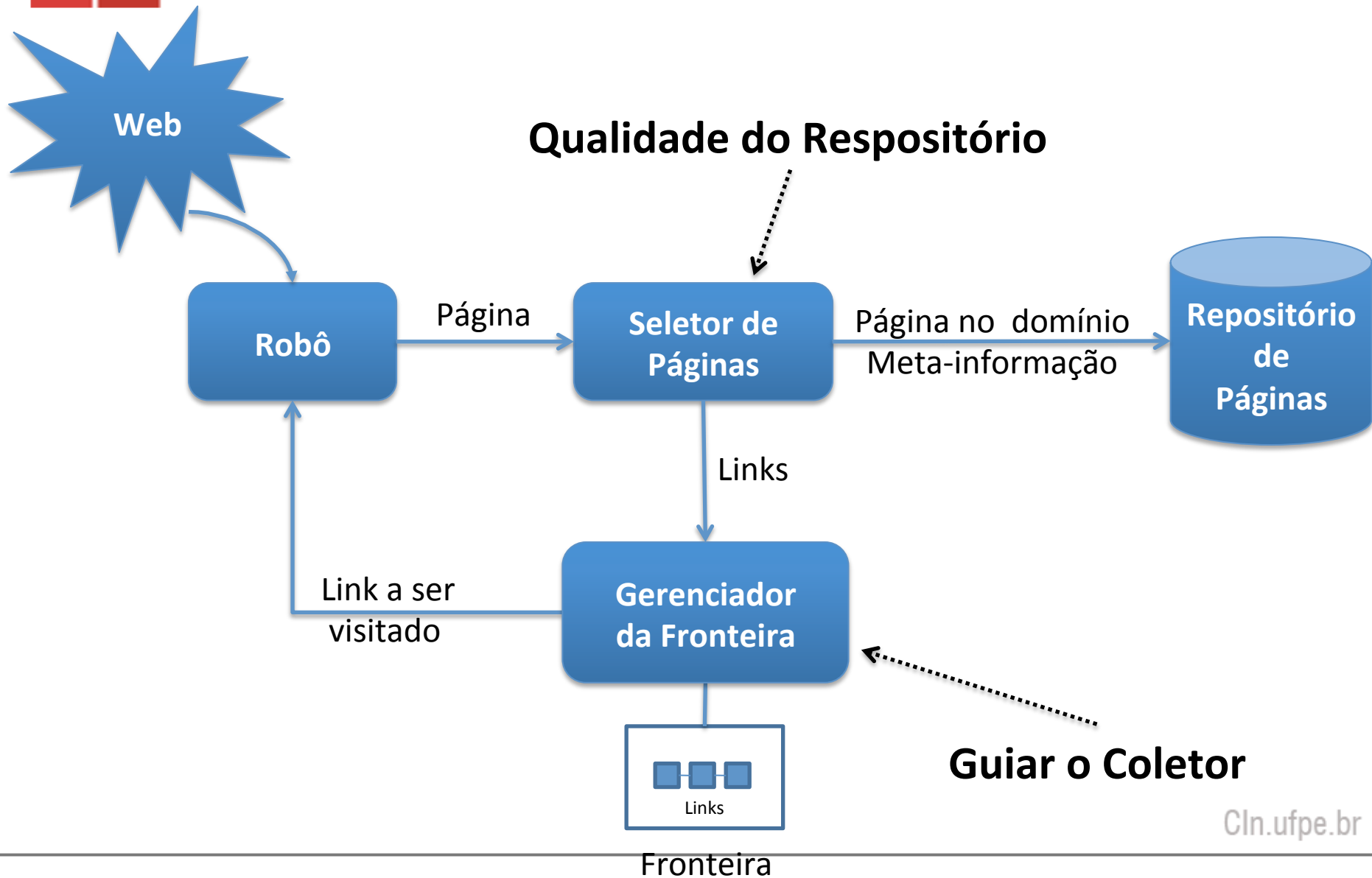


Coletor Focado



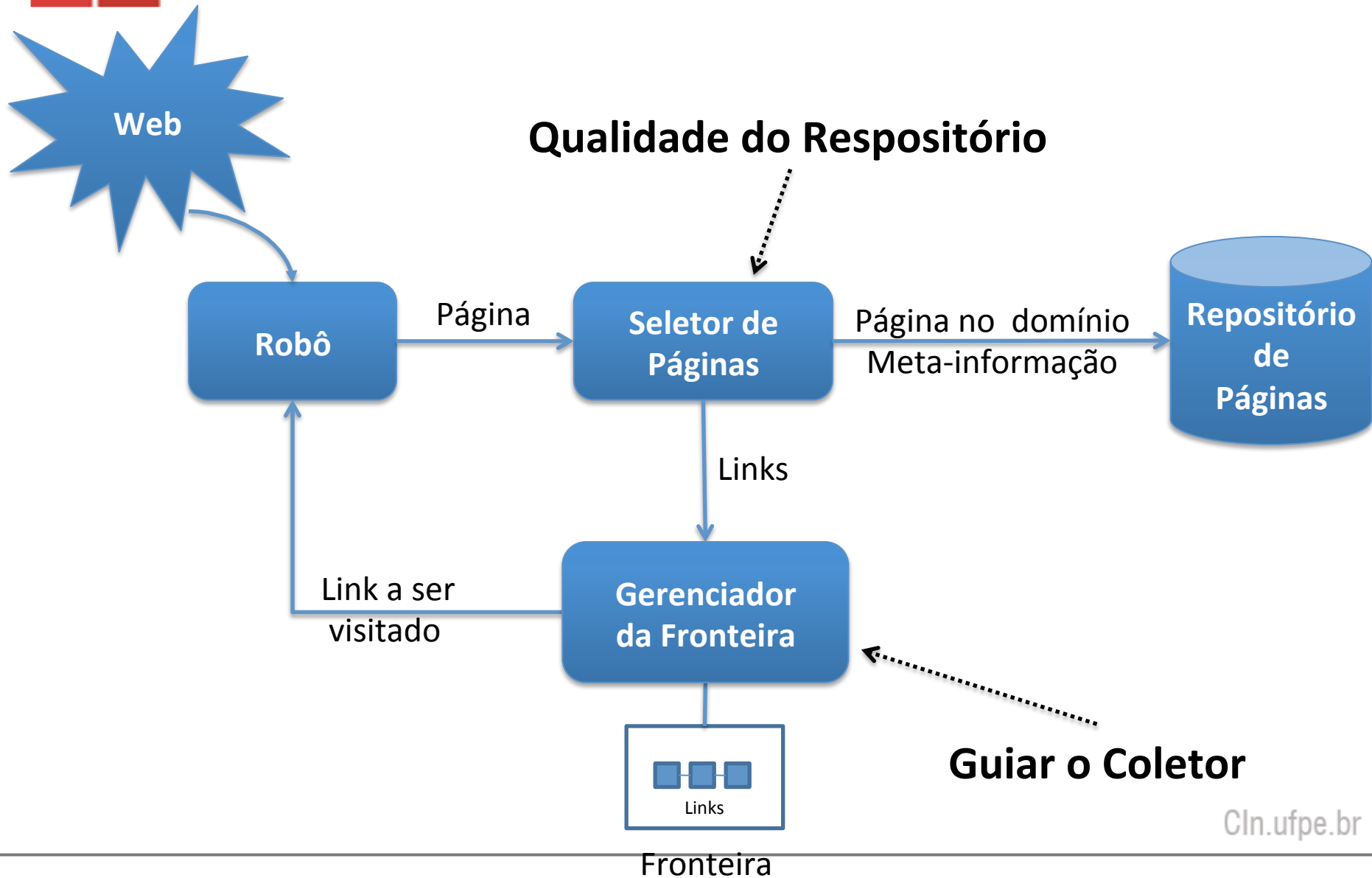


Funcionamento de um Coletor Focado





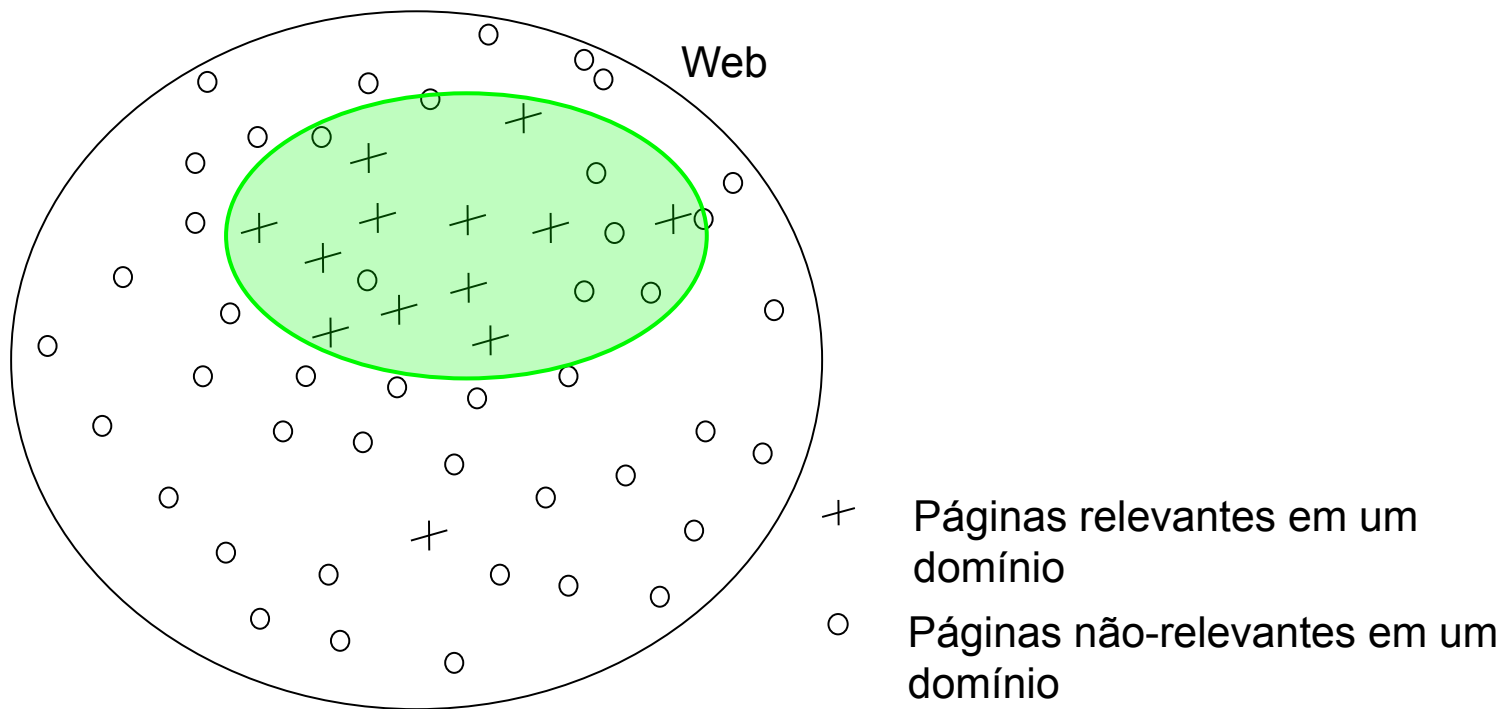
Funcionamento de um Coletor Focado





Seleção de Conteúdo em Coletores Focados

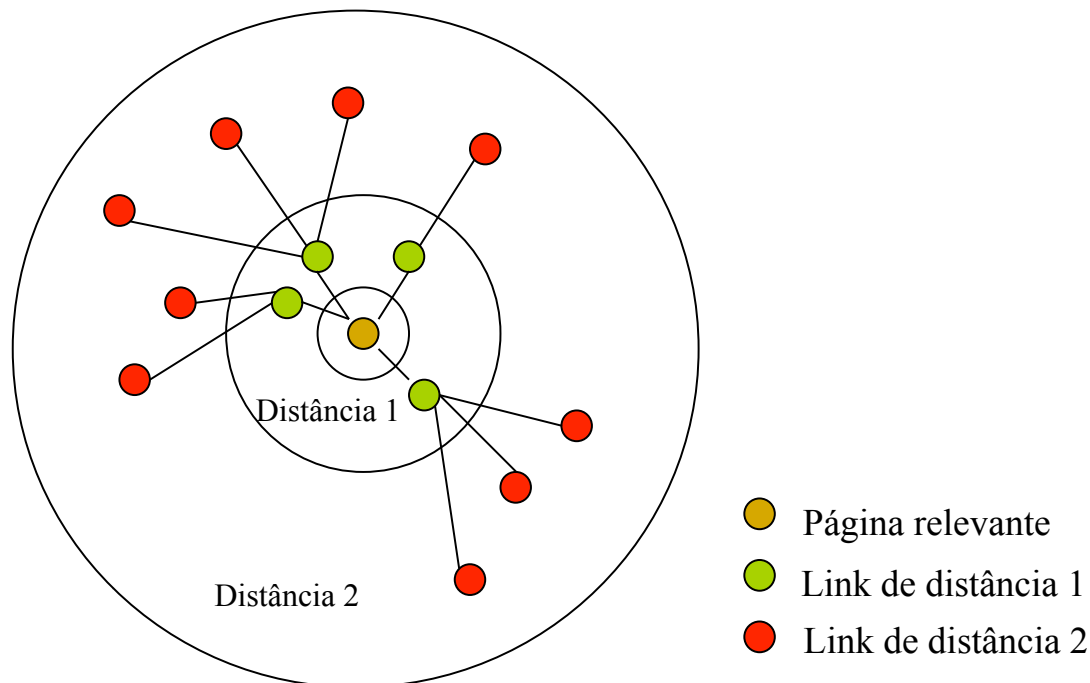
- Objetivo: seguir links promissores para encontrar páginas relevantes





Classificador de Links

- Estima a distância de uma URL a ser visitada para páginas relevantes
- Contexto: tokens na URL, âncora e ao redor da URL
- Baseado no grafo de contexto

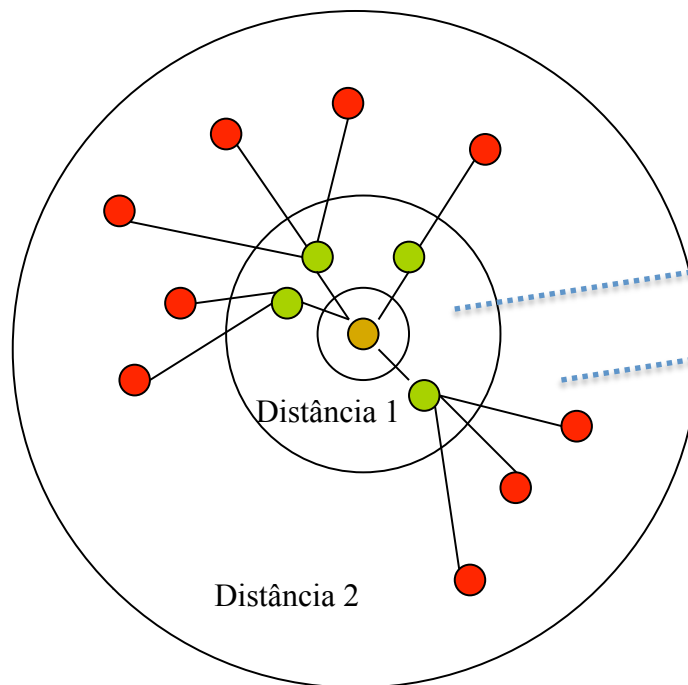


Grafo de Contexto



Classificador de Links

- Estima a distância de uma URL a ser visitada para páginas relevantes
- Contexto: tokens na URL, âncora e ao redor da URL
- Baseado no grafo de contexto



Exemplo: Formulários para
Busca por Empregos

Distância	URL	Âncora	Ao Redor
1	job search career	job advanced search career	job work search career
2	job	job career	job career

- Página relevante
- Link de distância 1
- Link de distância 2



Aspectos a Considerar na Construção de um Coletor

Escalabilidade

**Seleção de
Conteúdo**

**Sobrecarga
em Sites**

**Conteúdo
Adversário**



Conteúdo Adversário

- Não traz informação útil
- Tipos
 - Documentos duplicados
 - Crawler traps
 - Web Spam



Crawler Traps

- Geração automática de URLs para páginas de pouca relevância ou repetidas (ex.: calendários)
- Levam a um grande número de requisições desnecessárias
- Não necessariamente mal intencionadas
- Ex: <http://www.collinaimoveis.com.br/imoveis/>
- Usadas heurísticas (ex.: limitar o tamanho da URL ou número de páginas de um site)



Aspectos a Considerar na Construção de um Coletor

Escalabilidade

**Seleção de
Conteúdo**

**Sobrecarga
em Sites**

**Conteúdo
Adversário**



Robots.txt



www.zapimoveis.com.br/robots.txt

```
User-agent: *  
Disallow: /Erros/  
Disallow: /Erros/NaoEncontrado  
Disallow: /FichaImpressao/  
Disallow: /tr/  
Disallow: /95377733/  
Disallow: /anunciante/  
Disallow: /ie8/  
Disallow: /FichaImovel/Enviar  
Disallow: /Busca/Enviar  
Disallow: /FichaCampanha/Enviar
```



SiteMaps

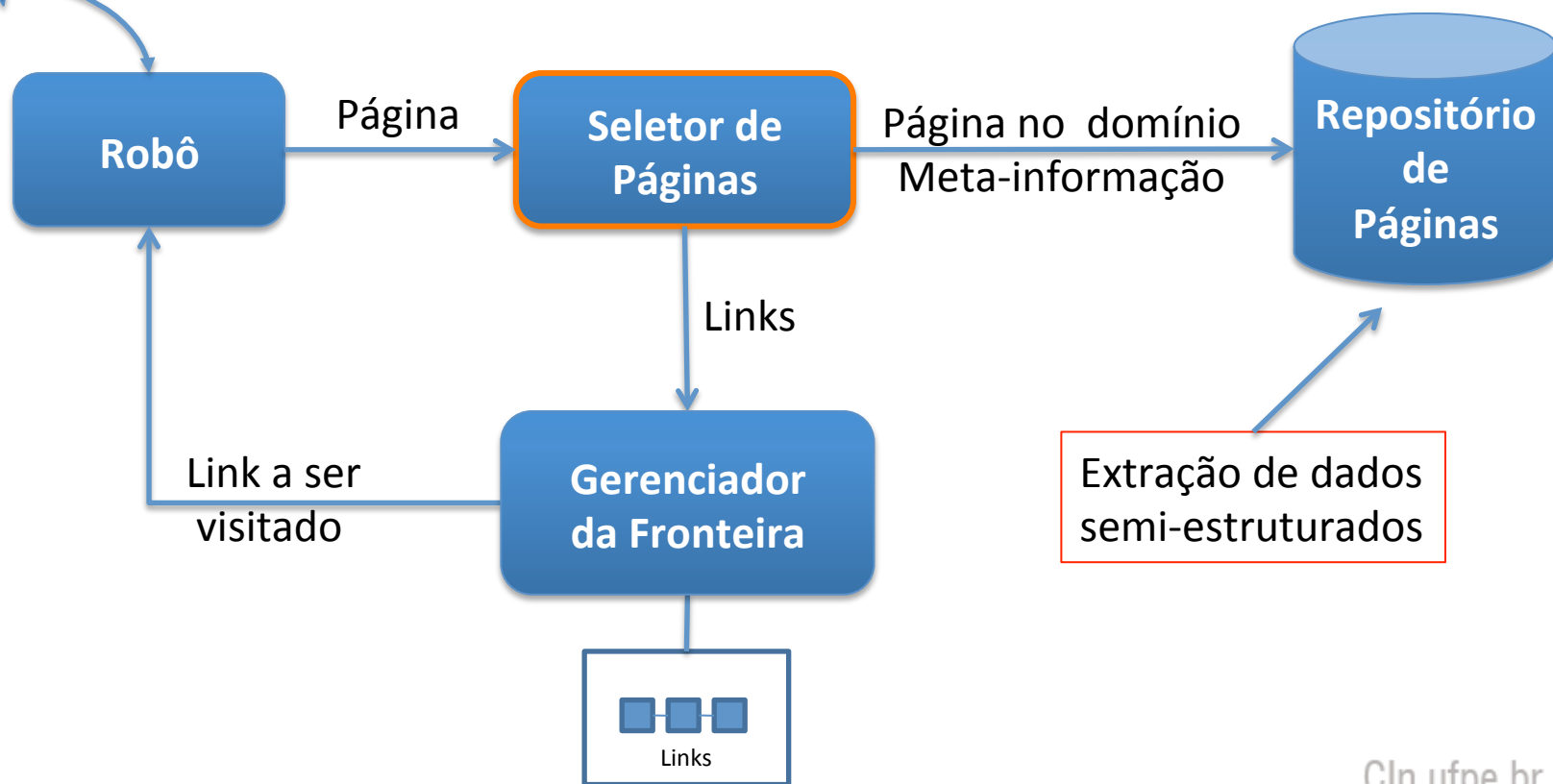
- Lista de URLs que podem ser coletadas, data da última modificação e taxa de atualização
- Criados pelos administradores

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  <url>
    <loc>http://www.company.com/</loc>
    <lastmod>2008-01-15</lastmod>
    <changefreq>monthly</changefreq>
    <priority>0.7</priority>
  </url>
  <url>
    <loc>http://www.company.com/items?item=truck</loc>
    <changefreq>weekly</changefreq>
  </url>
  <url>
    <loc>http://www.company.com/items?item=bicycle</loc>
    <changefreq>daily</changefreq>
  </url>
</urlset>
```



Coletor Focado e Extração de Dados

Web



Extração de dados
semi-estruturados

Fronteira



Extração de Dados Semi-Estruturados

- Objetivo: extrair informação relevante de páginas HTML -> transformar informação semi-estruturada em páginas Web para uma “base de dados” estruturada
- Entrada: página HTML
- Saída: Estrutura
 - Ex.: autor, preço, isbn para livros



Extração de Dados Semi-Estruturados

INFONET Classificados

Infonet → Classificados → Imóveis → Apartamentos para vender

Classificados

- Criar anúncio
- Lote de anúncios
- Anúncios salvos
- Dúvidas
- Fale conosco

Notícias

- Cidade
- Cultura
- Economia
- Educação
- Esporte
- Política
- Saúde

Diversão





- Agenda
- Cinema
- Eventos
- Promoções

Especiais

- Imposto de Renda
- Vídeos

Serviços

Belíssimo Condomínio Soberano Jardins.



Belíssimo apto c/ 3/4, suíte, 2 Wc, sala, cozinha, varanda 1 vaga de garagem, área de lazer completa em uma excelente localidade no bairro: Luzia. Ref.: 01674 Tel.: (79) 3302-6824/(79) 9828-1120 Cr 211PJ www.taiguaraimeiovels.com.br

Bairro: Luzia
Número de quartos: 3
Área: 78
Preço: 1.400,00
Contato: (79) 9828-1120
Telefone: (79) 9828-1120

Anúncio criado em 8 de Junho de 2015 15:08:04
1593 visitas desde a criação.

Marcar esse anúncio como: ☐ Categoria errada ☐ Anúncio proibido



Bairro	Luzia
Número de Quartos	3
Área	75
Preço	1.400,00
Contato	(79)9828-1120
Telefone	(79)9828-1120

Template preenchido

Página HTML

CIn.ufpe.br



Exemplo de Páginas de um mesmo Site

```
<html>1<body>2
  <b>3 Book4 Name5 </b>6 Databases
  <b>7 Reviews8 </b>9
  <ol>10
    <li>11
      <b>12 Reviewer13 Name14 </b>15 John
      <b>16 Rating17 </b>18 7
      <b>19 Text20 </b>21 ...
    </li>22
  </ol>23
</body>24</html>25
```

(a: p_{e1})

```
<html>1<body>2
  <b>3 Book4 Name5 </b>6 Query Opt.
  <b>7 Reviews8 </b>9
  <ol>10
    <li>11
      <b>12 Reviewer13 Name14 </b>15 John
      <b>16 Rating17 </b>18 8
      <b>19 Text20 </b>21 ...
    </li>22
  </ol>23
</body>24</html>25
```

(c: p_{e3})

```
<html>1<body>2
  <b>3 Book4 Name5 </b>6 Data Mining
  <b>7 Reviews8 </b>9
  <ol>10
    <li>11
      <b>12 Reviewer13 Name14 </b>15 Jeff
      <b>16 Rating17 </b>18 2
      <b>19 Text20 </b>21 ...
    </li>22
    <li>11
      <b>12 Reviewer13 Name14 </b>15 Jane
      <b>16 Rating17 </b>18 6
      <b>19 Text20 </b>21 ...
    </li>22
  </ol>23
</body>24</html>25
```

(b: p_{e2})

```
<html>1<body>2
  <b>3 Book4 Name5 </b>6 Transactions
  <b>7 Reviews8 </b>9
  <ol>10
    </ol>23
</body>24</html>25
```

(d: p_{e4})



Extração de Dados na Web

- Pros:
 - Muitas páginas geradas automaticamente a partir de um banco de dados
 - Estrutura HTML das páginas em um mesmo site (ou parte dele) é específica e regular

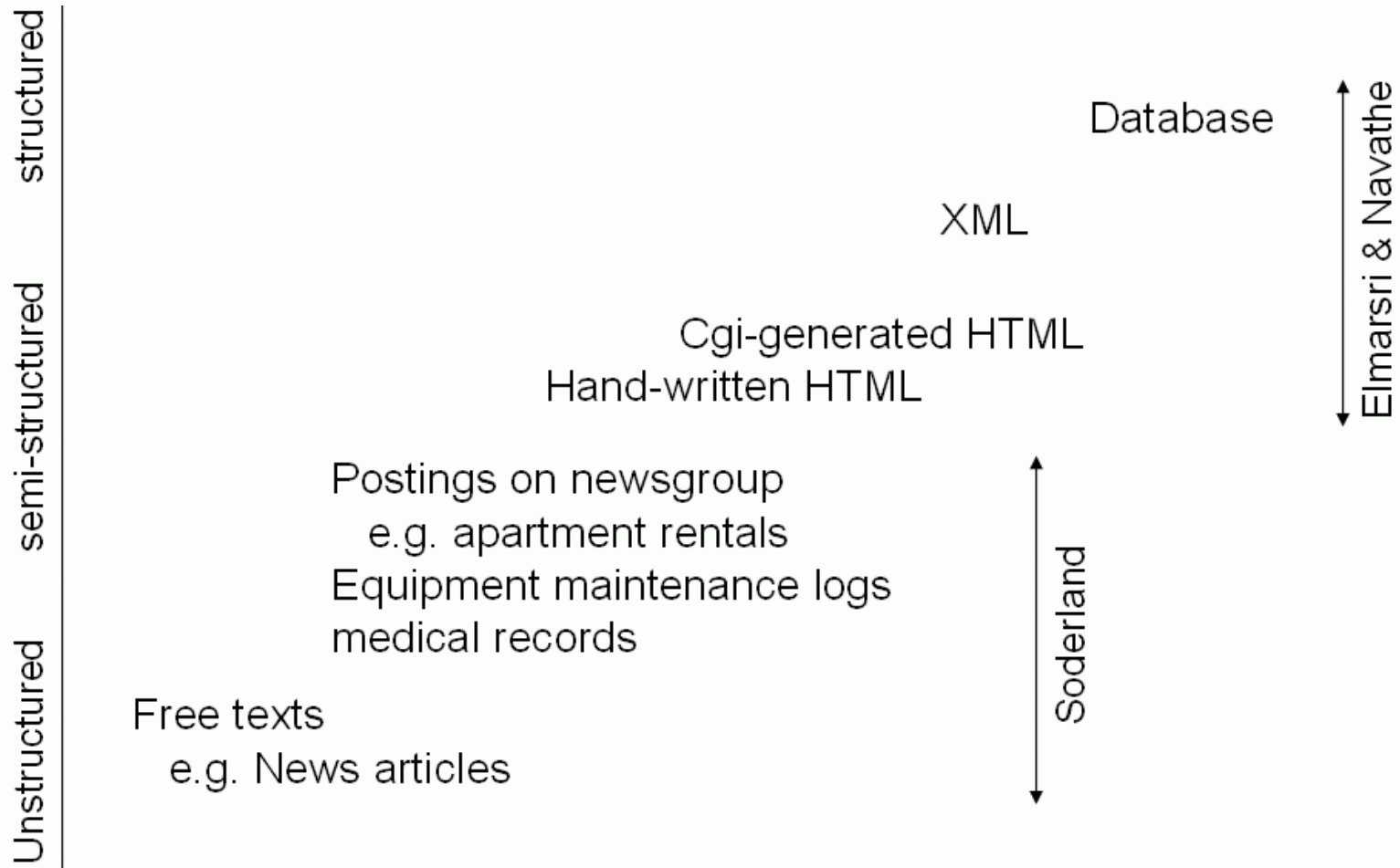


Extração de Dados na Web

- Cons:
 - Páginas criadas para o consumo humano e não de programas
 - Não identifica explicitamente os campos em tags
 - Estrutura varia entre sites
- Extração permite que o website seja visto como uma base de dados estruturados
- Extrator também chamado de wrapper
- Desafio: construir um extrator único para sites diferentes

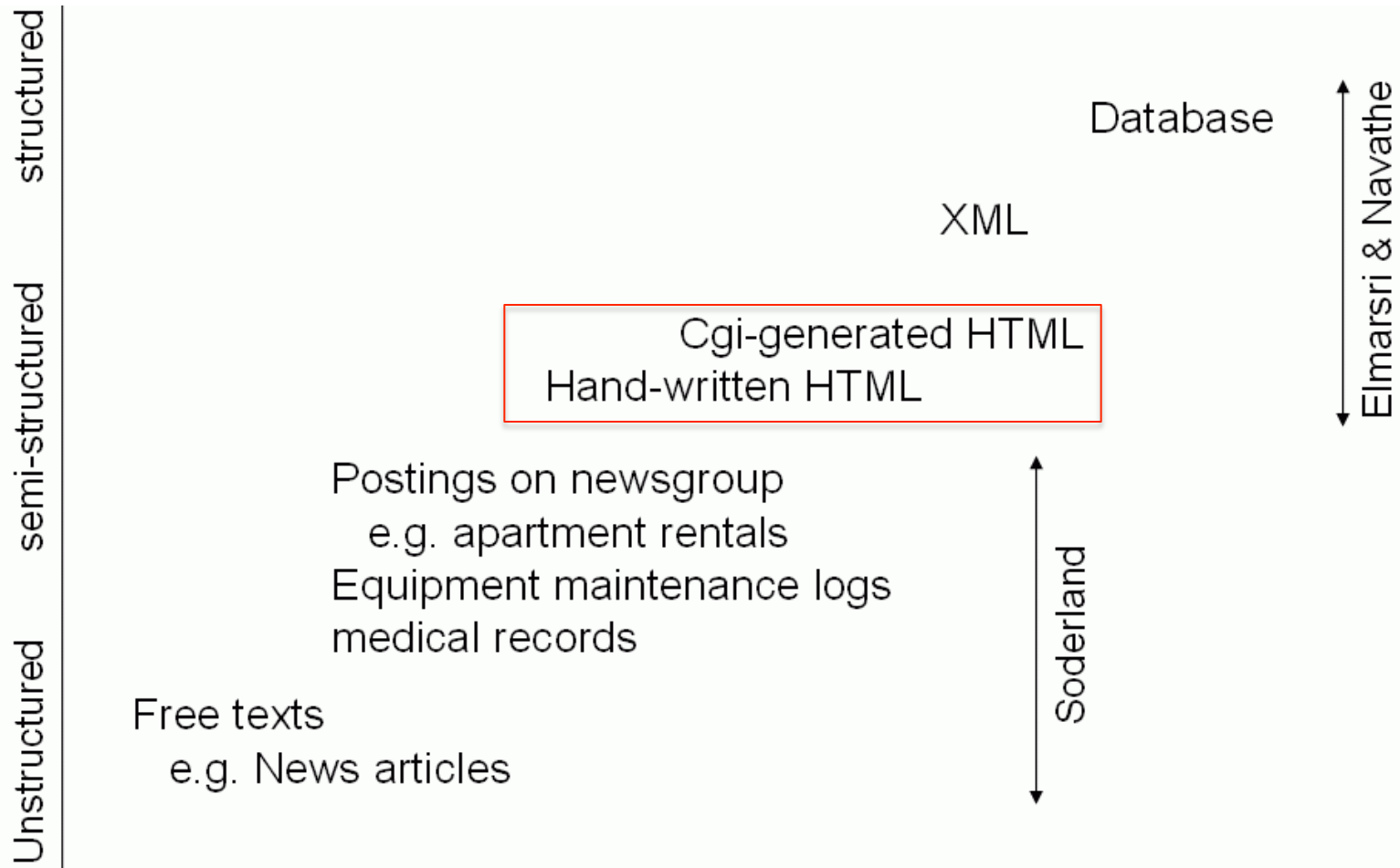


Extração de Dados





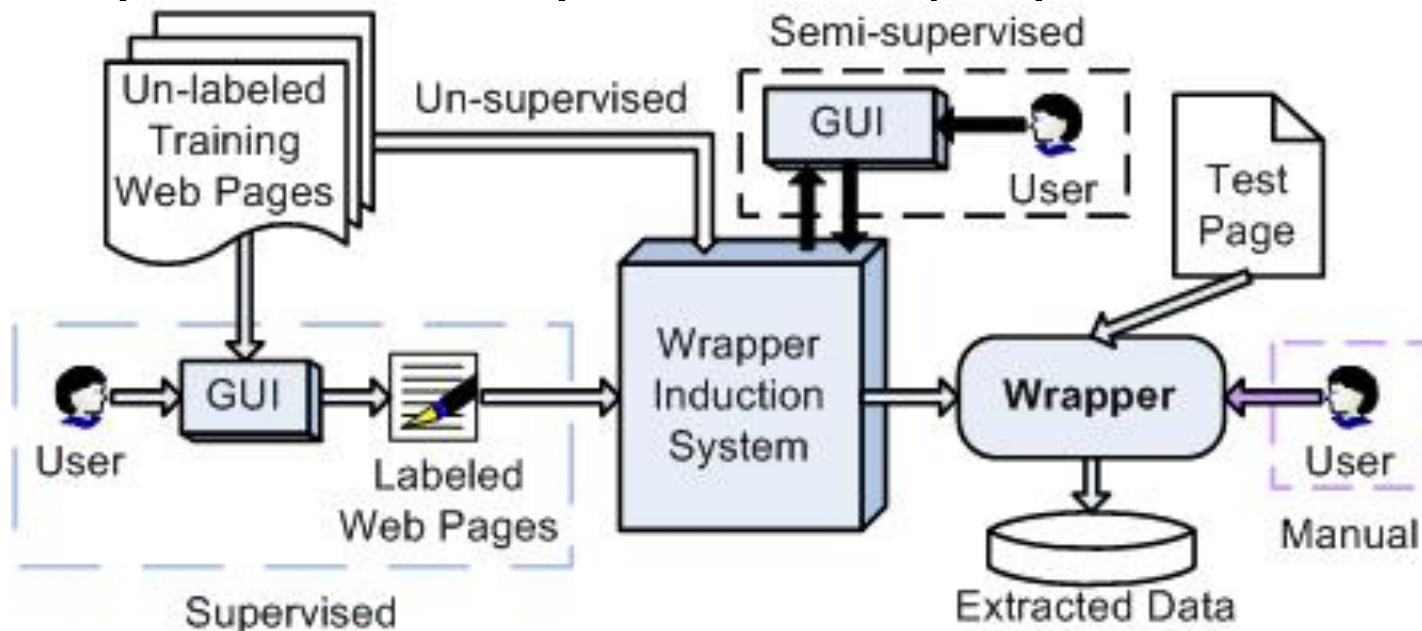
Extração de Dados





Formas de Extração

- “Manual”
- Supervisionada (baseado em anotação)
- Semi-supervisionada
- Não supervisionada (sem anotação)





Extração “Manual”

....

</td></tr>

</table>

<b class="sans">The Age of Spiritual Machines : When Computers Exceed Human Intelligence

by

Ray Kurzweil

List Price: \$14.95

Our Price: \$11.96

You Save: \$2.99

(20%)

<p>



Resultado da Extração

Title: **The Age of Spiritual Machines :**
When Computers Exceed Human Intelligence

Author: **Ray Kurzweil**

List-Price: **\$14.95**

Price: **\$11.96**

:

:



Exemplos de Padrões de Extração

- Expressão regular para extração
 - Preço: “`\$\d+(\.\d{2})?`”
- Pré-filtro para identificar contexto
 - Extração do preço do livro da Amazon:
 - Padrão pré-filtro: “`List Price: `”
 - Padrão do filtro: “`\$\d+(\.\d{2})?`”
- Pós-filtro para identificar fim do campo
 - Extração do preço do livro da Amazon:
 - Padrão pré-filtro: “`List Price: `”
 - Padrão do filtro: “`\$\d+(\.\d{2})?`”
 - Padrão pós-filtro: “``”

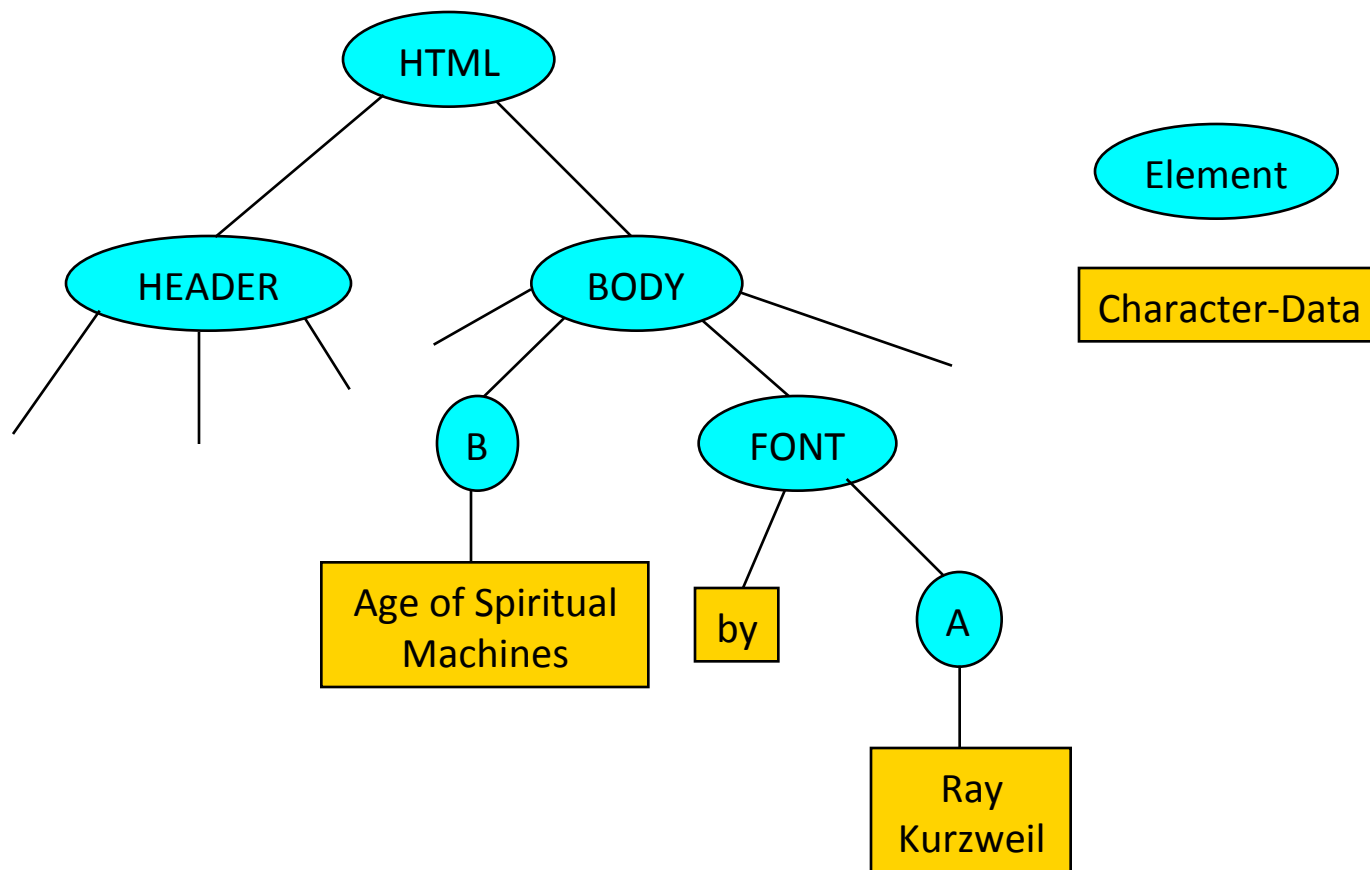


Extração baseada usando DOM Trees

- Abordagem simples
 - Parsing de DOM trees
 - Padrões de extração: caminhos da raiz da DOM Tree ao nó contendo o texto
 - Padrões de expressões regulares para identificar os nós de dados



Extração baseada usando DOM Trees



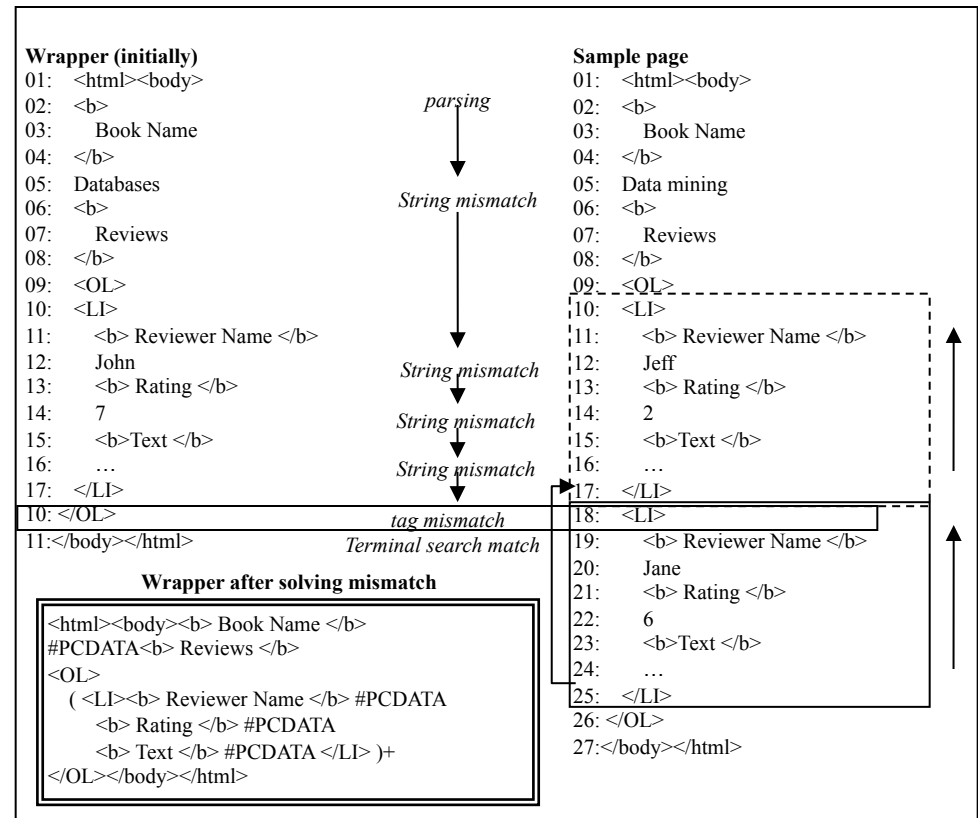
Title: HTML → BODY → B → CharacterData

Author: HTML → BODY → FONT → A → CharacterData



Não Supervisionado

- RoadRunner
 - Entrada: páginas com mesmo template
 - Faz o alinhamento de duas páginas de entrada ao mesmo tempo
 - Muito custoso computacionalmente





Semi-Supervisionado: Extração de Registros de Fóruns

Registros

Andorra in Late May
Apr 10, 2014, 12:50 PM

We will be passing through Andorra in late May. We're looking for advice/suggestions on what to do. We are traveling with our 18 month old daughter. We aren't really interested in the shopping, possibly a good area for nature, hiking (nothing too difficult) Any recommendations on a place to stay? We're looking for a mid range hotel, nothing too fancy, but not in the middle of the city.

posts: 13
reviews: 1

[Reply](#) [Report inappropriate content](#)

Travelers interested in this topic also viewed... [Hide x](#)

La Mola ★★★★★
#20 of 50 hotels in Encamp Parish
9 reviews
"Simple and friendly"
Marianne_D123 March 20, 2014

[Show Prices](#)

[See all 50 hotels in Encamp Parish](#)

6 replies to this topic

1-4 of 6 replies sorted by [Oldest first](#)

1. Re: Andorra in Late May
Apr 11, 2014, 1:02 PM

To be honest, there's not a lot to do in Andorra with a toddler if you don't like shopping, because you aren't going to be skiing, serious hiking, mountain-biking etc. And a lot of places are either on the one main through road, or ski resorts (deserted in summer), or both. However . . .

We were there last June for a walking holiday, but we couldn't do a lot of the walks we wanted to because there was still a lot of snow at high level. We stayed at the Hotel Coma in Ordino, which might fit the bill for you. Ordino is not exactly a hive of activity, but it's a nice little village. And we just about managed to fill the week with lower level walks, and we enjoyed it enough that we are going back in September, and intend to stay at the Hotel Coma again.

[Reply](#) [Report inappropriate content](#)

2. Re: Andorra in Late May
Apr 11, 2014, 2:57 PM

There are easy nature trails at Lake Engolasters that are good with a toddler. Here is a link to one of them hola-andorra.com/animals/...pardinesgb.html

posts: 883
reviews: 10

[Reply](#) [Report inappropriate content](#)

- Best resort for snowboarders?
- Best town in Andorra to stay in?
- Which airport should we use for Andorra, and how to get there?
- Car hire from Barcelona?
- Driving directions from Barcelona.
- Prettiest areas or villages for walking or sightseeing holiday in Andorra?
- Is there a nanny or creche service in any resorts in Andorra?

Beyond destination forums

- Air Travel
 - Business Travel
 - Timeshares / Holiday Rentals
- [See all »](#)

Explore the world! TripAdvisor has reviews and information on over 400,000 locations, including:

Hotels
Gui Panajon
Atrium Tropical Exclusive Club & Spa In Ko Samui
Okemo Inn
Hotel Wing International Shin-Osaka
Iberostar Rose Hall Suites in Rose Hall
Travel Destinations
Garden Grove Hotels
New Orleans
Sightseeing
Border Trade Street, Rull

Explore other Andorra resources:

[Andorra Bed and Breakfast](#)

Popular cities

[Andorra la Vella Hotels](#)
[Arius Hotels](#)
[Canillo Hotels](#)
[El Tarter Hotels](#)
[Encamp Hotels](#)
[Les Escaldes Hotels](#)
[Ordino Hotels](#)
[Pas de la Casa Hotels](#)
[Soldeu Hotels](#)

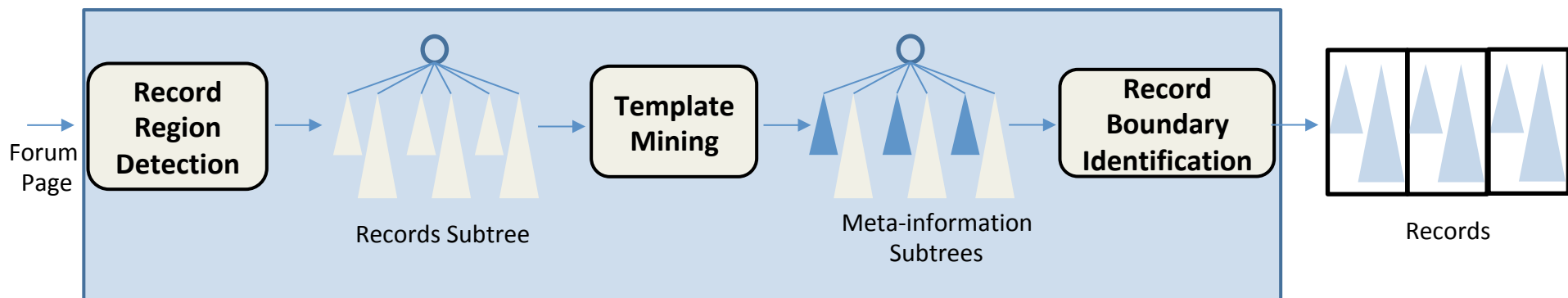


Extração de Registros de Fóruns

- Objetivo: extrair registros de fóruns
- Abordagem simples: construir um wrapper para cada site
- Solução:
 - Requisitos:
 - Independente de site e tópico
 - Pouca supervisão
 - Desafio: estrutura das árvores varia bastante entre sites



Extração de Registros de Fóruns





Detecção da Região do Registro

- Suposição: registros são irmãos do mesmo nó-pai na DOM tree

The screenshot shows a forum thread on TripAdvisor. The main post is titled "Andorra in Late May" and was posted by user "cinthia" on April 10, 2014. The post asks for advice on what to do in Andorra, mentioning a 18-month-old daughter and a mid-range hotel. Below the main post, there are two replies. The first reply, by "williams3205", discusses the weather and activities in Andorra. The second reply, by "YukeTheTraveler", mentions nature trails at Lake Engolasters. To the right of the forum thread, there is a sidebar with various links and recommendations, including "Beyond destination forums", "Explore the world!", "Travel Destinations", and "Popular cities".

Andorra in Late May
Apr 10, 2014, 12:50 PM

We will be passing through Andorra in late May. We're looking for advice/suggestions on what to do. We are traveling with our 18 month old daughter. We aren't really interested in the shopping, possibly a good area for nature, hiking (nothing too difficult) Any recommendations on a place to stay? We're looking for a mid range hotel, nothing too fancy, but not in the middle of the city.

[Reply](#) [Report inappropriate content](#)

Travelers interested in this topic also viewed...

La Mola ★★★★★
#20 of 50 hotels in Encamp Parish
3 reviews
"Simple and friendly"
Marlene_D103 March 20, 2014

[Show Prices](#)

[See all 50 hotels in Encamp Parish](#)

6 replies to this topic

1-6 of 6 replies Sorted by [Oldest first](#)

1. Re: Andorra in Late May
Apr 11, 2014, 1:02 PM

To be honest, there's not a lot to do in Andorra with a toddler if you don't like shopping, because you aren't going to be skiing, serious hiking, mountain-biking etc. And a lot of places are either on the one main through road, or ski resorts (deserted in summer), or both. However...

We were there last June for a walking holiday, but we couldn't do a lot of the walks we wanted to because there was still a lot of snow at high level. We stayed at the Hotel Coma in Ordino, which might fit the bill for you. Ordino is not exactly a hive of activity, but it's a nice little village. And we just about managed to fill the week with lower level walks, and we enjoyed it enough that we are going back in September, and intend to stay at the Hotel Coma again.

[Reply](#) [Report inappropriate content](#)

2. Re: Andorra in Late May
Apr 11, 2014, 2:57 PM

There are easy nature trails at Lake Engolasters that are good with a toddler. Here is a link to one of them hola-andorra.com/animas/pardineagb.html

[Reply](#) [Report inappropriate content](#)

Beyond destination forums

- Air Travel
- Business Travel
- Timeshares / Holiday Rentals

[See all »](#)

Explore the world! TripAdvisor has reviews and information on over 400,000 locations, including:

Hotels

- Gui Panajon
- Amrum Tropical Exclusive Club & Spa in Ko Samui
- Olema Inn
- Hotel Wing International Shinjuku
- Derostar Rose Hall Suites in Rose Hall

Travel Destinations

- Garden Grove Hotels
- New Orleans
- Sightseeing
- Border Trade Street, Ruit

Explore other Andorra resources:

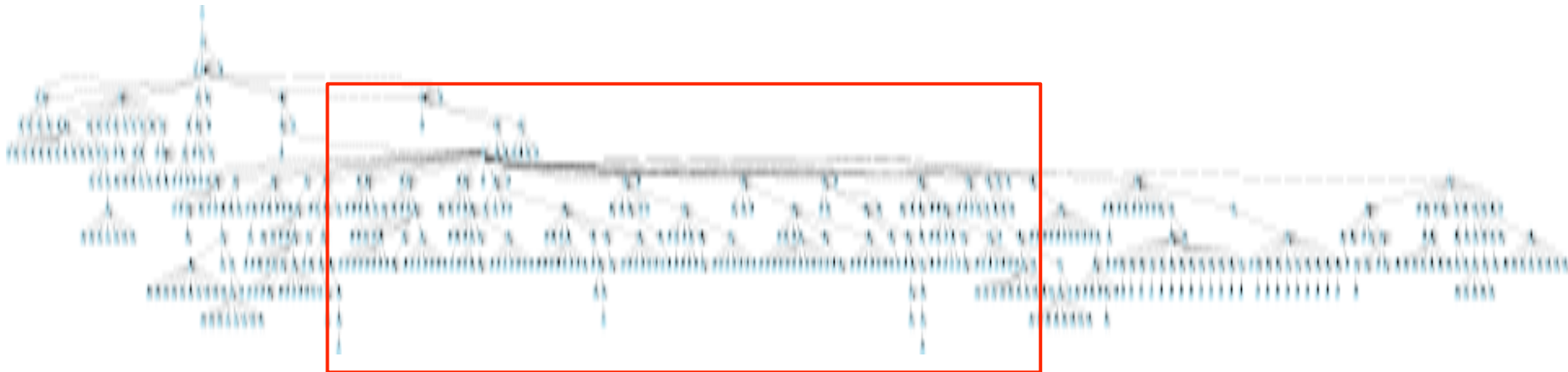
- Andorra Bed and Breakfast

Popular cities

- Andorra la Vella Hotels
- Arniol Hotels
- Canillo Hotels
- El Tarter Hotels
- Encamp Hotels
- Les Escaldes Hotels
- Ordino Hotels
- Pas de la Casa Hotels
- Soldeu Hotels



Dom Tree



Região dos Registros



Detecção da Região do Registros

- Suposições:
 - Registros são irmãos do mesmo nó pai na DOM tree
 - Árvore do nó pai é “balanceada” com relação à distribuição de nós-folha de meta-informação
- Meta-informação (ou template): detectores de usuário, data/tempo e título

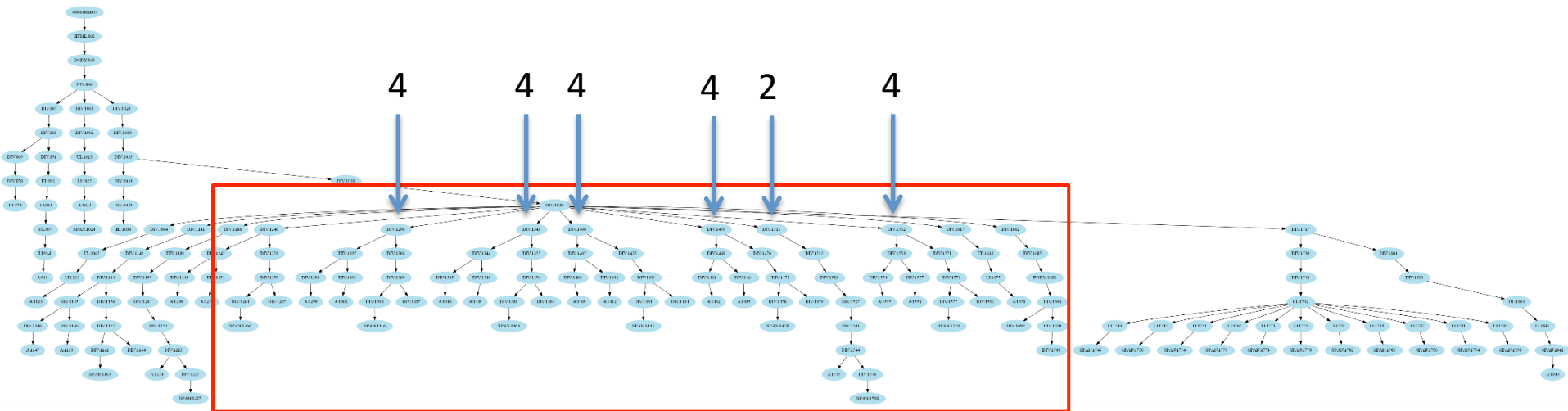
The screenshot displays a forum thread on TripAdvisor. Annotations with blue dashed arrows point to specific elements in the thread, illustrating the detection of meta-information:

- Title:** "Andorra in Late May" (highlighted with a blue box).
- User:** "cristofon" (highlighted with a blue box).
- Date:** "Apr 10, 2014, 12:50 PM" (highlighted with a blue box).
- Post Count:** "posts: 13" (highlighted with a blue box).
- Reply Count:** "replies: 1" (highlighted with a blue box).
- Content:** The main body of the post, starting with "We will be passing through Andorra in late May. We're looking for advice/suggestions on what to do..." (highlighted with a blue box).
- Reply:** "1. Re: Andorra in Late May" (highlighted with a blue box).
- User:** "williams3305" (highlighted with a blue box).
- Date:** "Apr 11, 2014, 1:02 PM" (highlighted with a blue box).
- Content:** The body of the reply, starting with "To be honest, there's not a lot to do in Andorra with a toddler if you don't like shopping..." (highlighted with a blue box).
- Reply:** "2. Re: Andorra in Late May" (highlighted with a blue box).
- User:** "YuleTheTraveler" (highlighted with a blue box).
- Date:** "Apr 11, 2014, 2:57 PM" (highlighted with a blue box).
- Content:** The body of the reply, starting with "There are easy nature trails at Lake Engolasters that are good with a toddler..." (highlighted with a blue box).

Other visible elements include a list of hotels in Encamp Parish, a list of travel destinations, and a list of popular cities in Andorra.

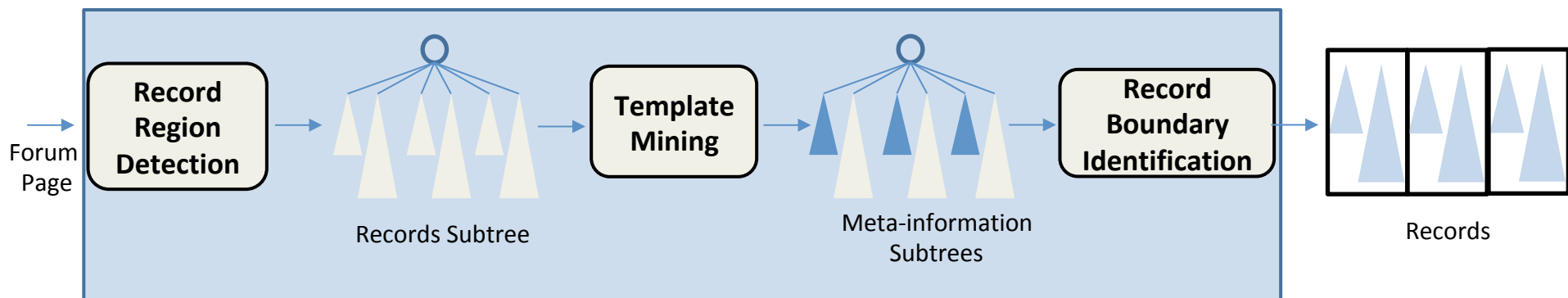


Nós-Folha Detectados e seus Ascendentes





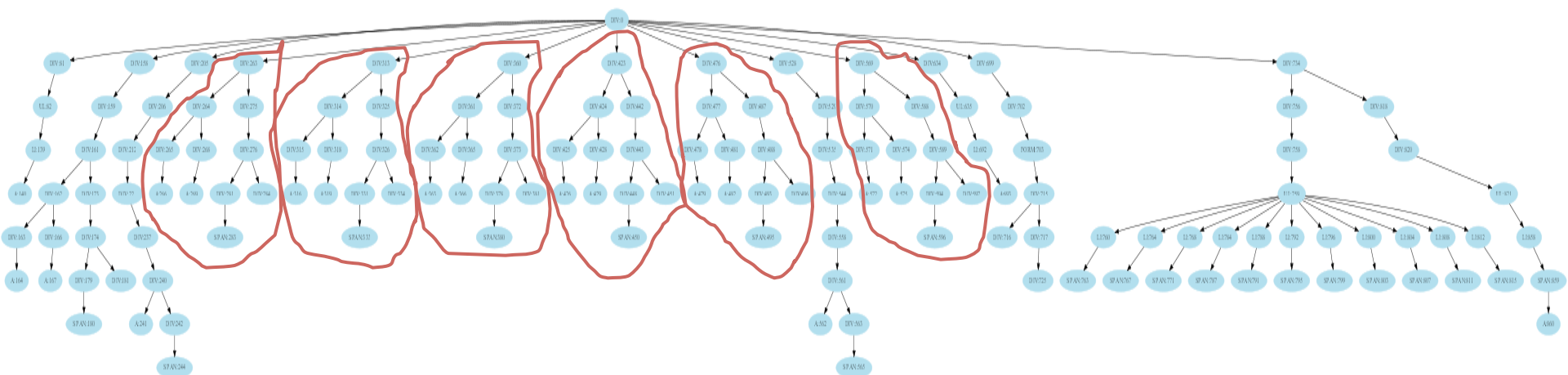
Mineração de Templates






Mineração de Templates

- Suposição: árvores de meta-informação têm estruturas similares





Registros



clinton
Boston...

posts: 13
reviews: 1

Andorra in Late May

Apr 10, 2014, 12:50 PM

We will be passing through Andorra in late May. We're looking for advice/suggestions on what to do. We are traveling with our 18 month old daughter. We aren't really interested in the shopping, possibly a good area for nature, hiking (nothing too difficult) Any recommendations on a place to stay? We're looking for a mid range hotel, nothing too fancy, but not in the middle of the city.

[Reply](#)

[Report inappropriate content](#)

Travelers interested in this topic also viewed... [Hide x](#)




La Mola ★★★★★
#20 of 50 hotels in Encamp Parish
9 reviews
"Simple and friendly"
Marianne_D123 March 20, 2014

[Show Prices](#)

[See all 50 hotels in Encamp Parish](#)

6 replies to this topic

1-4 of 6 replies sorted by [Oldest first](#)



williams305
Liverpool

posts: 302
reviews: 1

1. Re: Andorra in Late May


Apr 11, 2014, 1:02 PM

To be honest, there's not a lot to do in Andorra with a toddler if you don't like shopping, because you aren't going to be skiing, serious hiking, mountain-biking etc. And a lot of places are either on the one main through road, or ski resorts (deserted in summer), or both. However . . .

We were there last June for a walking holiday, but we couldn't do a lot of the walks we wanted to because there was still a lot of snow at high level. We stayed at the Hotel Coma in Ordino, which might fit the bill for you. Ordino is not exactly a hive of activity, but it's a nice little village. And we just about managed to fill the week with lower level walks, and we enjoyed it enough that we are going back in September, and intend to stay at the Hotel Coma again.

[Reply](#)

[Report inappropriate content](#)



YukeTheTraveler
Finland

posts: 883
reviews: 10

2. Re: Andorra in Late May

Apr 11, 2014, 2:57 PM

There are easy nature trails at Lake Engolasters that are good with a toddler. Here is a link to one of them hola-andorra.com/animals/...pardinesgb.html

[Reply](#)

[Report inappropriate content](#)

- Best resort for snowboarders?
- Best town in Andorra to stay in?
- Which airport should we use for Andorra, and how to get there?
- Car hire from Barcelona?
- Driving directions from Barcelona.
- Prettiest areas or villages for walking or sightseeing holiday in Andorra?
- Is there a nanny or creche service in any resorts in Andorra?

Beyond destination forums

- Air Travel
 - Business Travel
 - Timeshares / Holiday Rentals
- [See all »](#)

Explore the world! TripAdvisor has reviews and information on over 400,000 locations, including:

- Hotels
- Gul Panajon
- Atrium Tropical Exclusive Club & Spa In Ko Samui
- Olema Inn
- Hotel Wing International Shin-Osaka
- Iberostar Rose Hall Suites in Rose Hall
- Travel Destinations
- Garden Grove Hotels
- New Orleans
- Sightseeing
- Border Trade Street, Rull

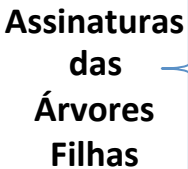
Explore other Andorra resources:

[Andorra Bed and Breakfast](#)

Popular cities

- [Andorra la Vella Hotels](#)
- [Arius Hotels](#)
- [Canillo Hotels](#)
- [El Tarter Hotels](#)
- [Encamp Hotels](#)
- [Les Escaldes Hotels](#)
- [Ordino Hotels](#)
- [Pas de la Casa Hotels](#)
- [Soldeu Hotels](#)

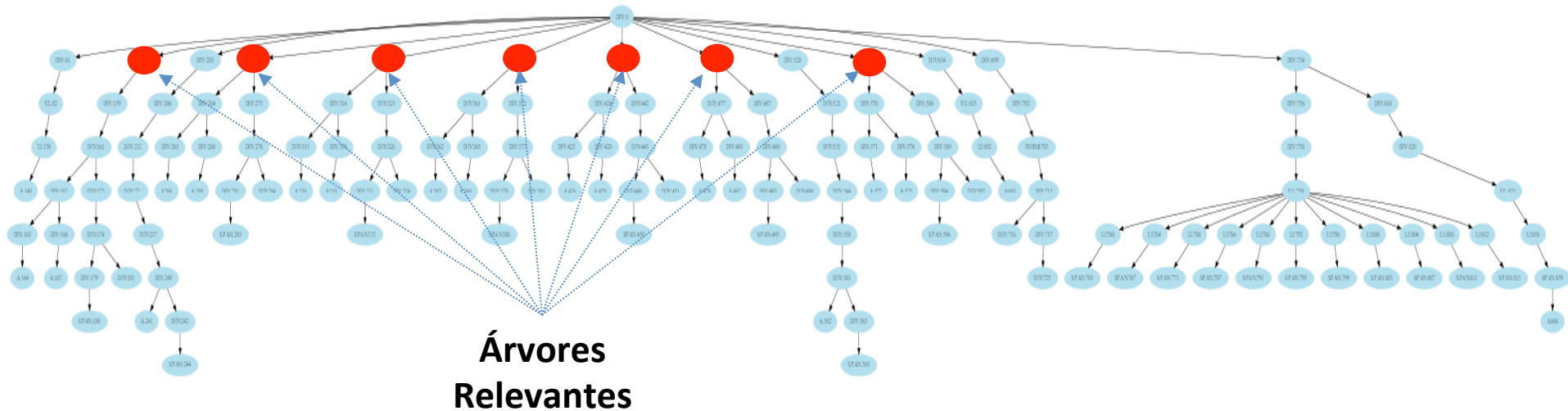
Anúncio



```
L:DIV:DIV::A:SPAN:LI:UL:DIV:DIV:DIV:
```



Mineração de Templates



Árvores Relevantes

**Maior cluster
(Árvores relevantes)**

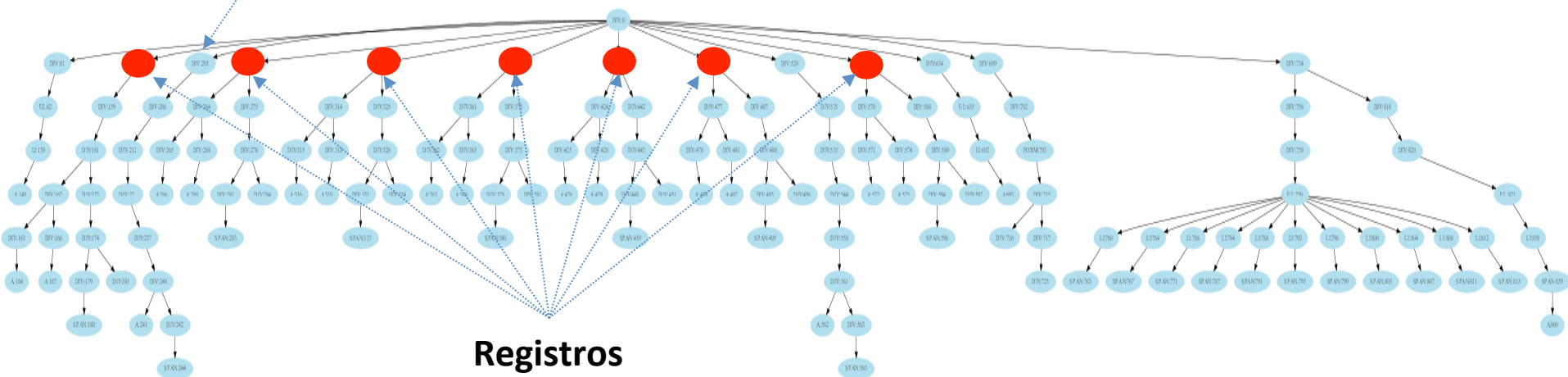
A:LI:UL:DIV:
A:DIV:A:DIV:DIV::SPAN:DIV::DIV:DIV:DIV:DIV:DIV:
A::SPAN:DIV:DIV:DIV:DIV:DIV:DIV:DIV:
A:DIV:A:DIV:DIV::SPAN:DIV::DIV:DIV:DIV:DIV:
A:DIV:A:DIV:DIV::SPAN:DIV::DIV:DIV:DIV:DIV:
A:DIV:A:DIV:DIV::SPAN:DIV::DIV:DIV:DIV:DIV:
A:DIV:A:DIV:DIV::SPAN:DIV::DIV:DIV:DIV:DIV:
A::SPAN:DIV:DIV:DIV:DIV:DIV:DIV:DIV:
A:DIV:A:DIV:DIV::SPAN:DIV::DIV:DIV:DIV:DIV:
A:LI:UL:DIV:
DIV::DIV:DIV:DIV:FORM:DIV:DIV:
SPAN:LI::SPAN:LI::SPAN:LI::SPAN:LI::SPAN:LI::SPAN:LI::SPAN:LI::SPAN:LI::SPAN:LI:U
L:DIV:DIV::A:SPAN:LI:UL:DIV:DIV:DIV:



A:DIV:A:DIV:DIV::SPAN:DIV::DIV:DIV:DIV:DIV:DIV:DIV:
A:DIV:A:DIV:DIV::SPAN:DIV::DIV:DIV:DIV:DIV:DIV:
A:DIV:A:DIV:DIV::SPAN:DIV::DIV:DIV:DIV:DIV:DIV:
A:DIV:A:DIV:DIV::SPAN:DIV::DIV:DIV:DIV:DIV:DIV:
A:DIV:A:DIV:DIV::SPAN:DIV::DIV:DIV:DIV:DIV:DIV:
A:DIV:A:DIV:DIV::SPAN:DIV::DIV:DIV:DIV:DIV:DIV:
A:DIV:A:DIV:DIV::SPAN:DIV::DIV:DIV:DIV:DIV:DIV:



Anúncio





c0ffeeCat Manila posts: 5 reviews: 2 Save this Post 3. Re: Andorra in Late May Sep 30, 2014, 5:15 AM Hi, I created a forum ...



Crawlers e Parsers Disponíveis

- Crawlers
 - Scrapy: <https://doc.scrapy.org/en/latest/>
 - Pyspider: <https://github.com/binux/pyspider>
 - MechanicalSoup: <https://mechanicalsoup.readthedocs.io/en/stable/>
- Parsers
 - BeautifulSoup: <https://pypi.org/project/beautifulsoup4/>
 - HTMLParser: <https://docs.python.org/2/library/htmlparser.html>
- Headless browsers
 - SeleniumHQ: <https://www.seleniumhq.org/>
 - PhantomJS: <http://phantomjs.org/>