



# Recuperação de Informação: Extração de Dados na Web

Prof. Luciano Barbosa

(Parte do material retirado dos slides dos livros adotados)



UNIVERSIDADE  
FEDERAL  
DE PERNAMBUCO



# 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**

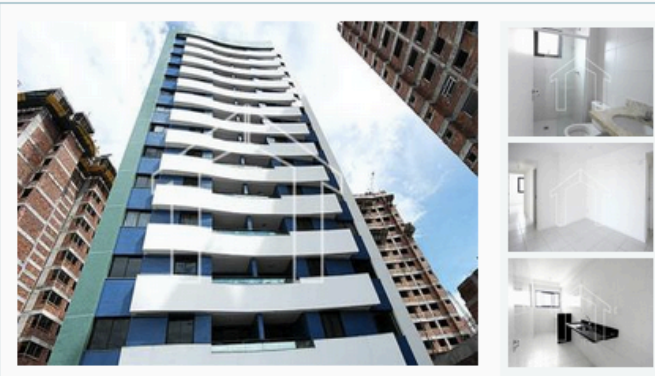
- Agenda
- Cinema
- Eventos
- Promoções

**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.taiguaraimoveis.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



# 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

# 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 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 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 Transactions
  <b>7 Reviews8 </b>9
  <ol>10
  </ol>23
</body>24</html>25
```

(d:  $p_{e4}$ )

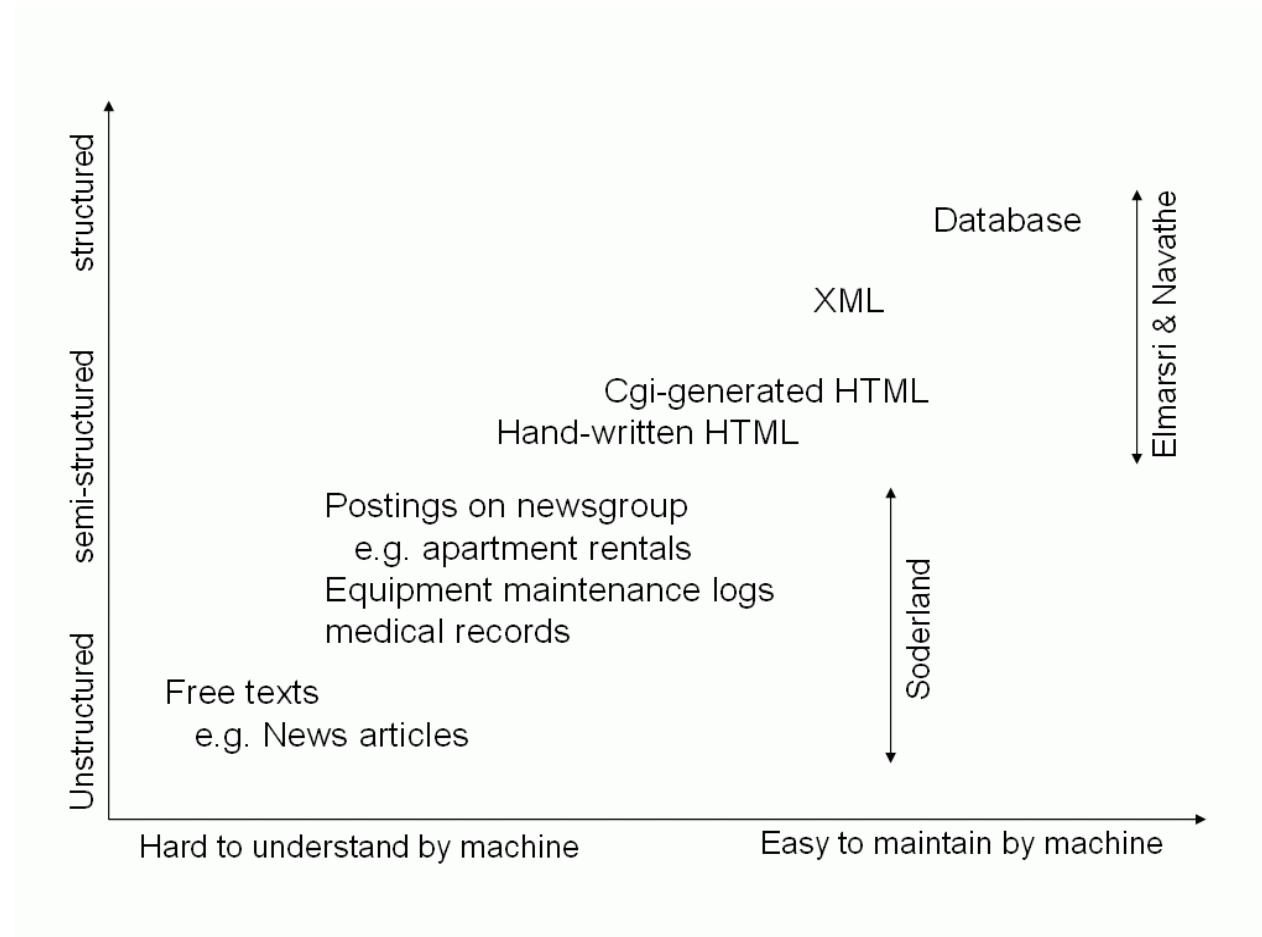


# Extração de Dados na Web

- 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

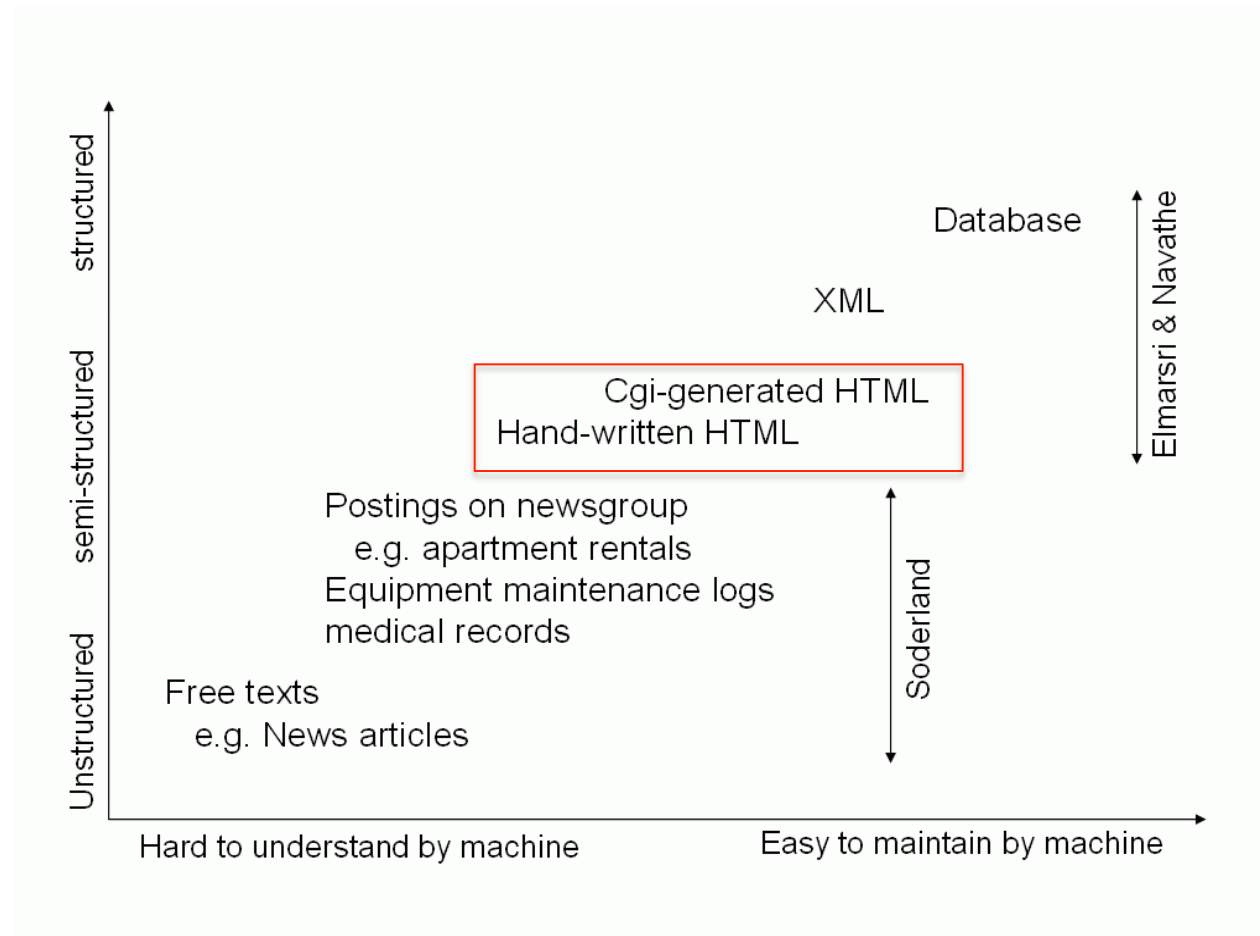


# Extração de Dados





# Extração de Dados

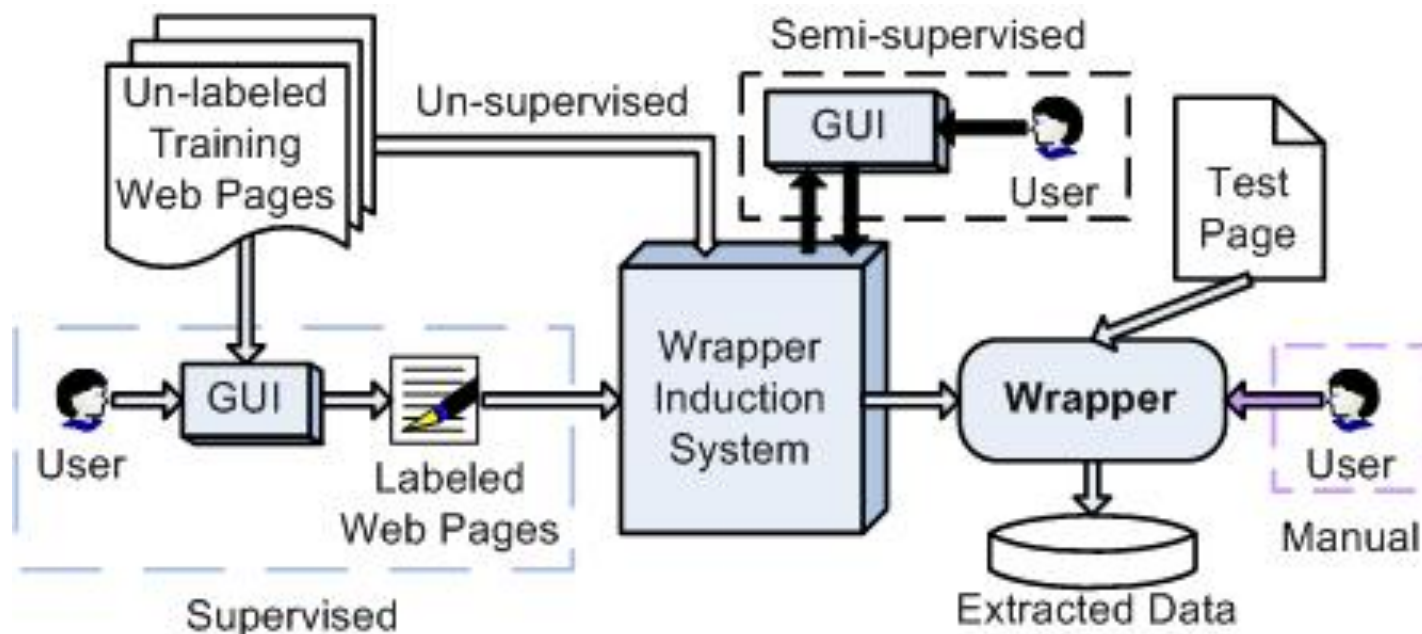






# Formas de Extração

- “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</b><br>

<font face=verdana,arial,helvetica size=-1>

by <a href="/exec/obidos/search-handle-url/index=books&field-author=Kurzweil%2C%20Ray/002-6235079-4593641">

Ray Kurzweil</a><br>

</font>

<br>

<a href="http://images.amazon.com/images/P/0140282025.01.LZZZZZZZ.jpg">

</a>

<font face=verdana,arial,helvetica size=-1>

<span class="small">

<span class="small">

<b>List Price:</b> <span class=listprice>\$14.95</span><br>

<b>Our Price: <font color=#990000>\$11.96</font></b><br>

<b>You Save:</b> <font color=#990000><b>\$2.99</b>

(20%)</font><br>

</span>

<p> <br>



# 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: “`<b>List Price:</b> <span class=listprice>`”
    - Padrão do filtro: “`\$\d+(\.\d{2})?`”
- Pós-filtro para identificar fim do campo
  - Extração do preço do livro da Amazon:
    - Padrão pré-filtro: “`<b>List Price:</b> <span class=listprice>`”
    - Padrão do filtro: “`\$\d+(\.\d{2})?`”
    - Padrão pós-filtro: “`</span>`”

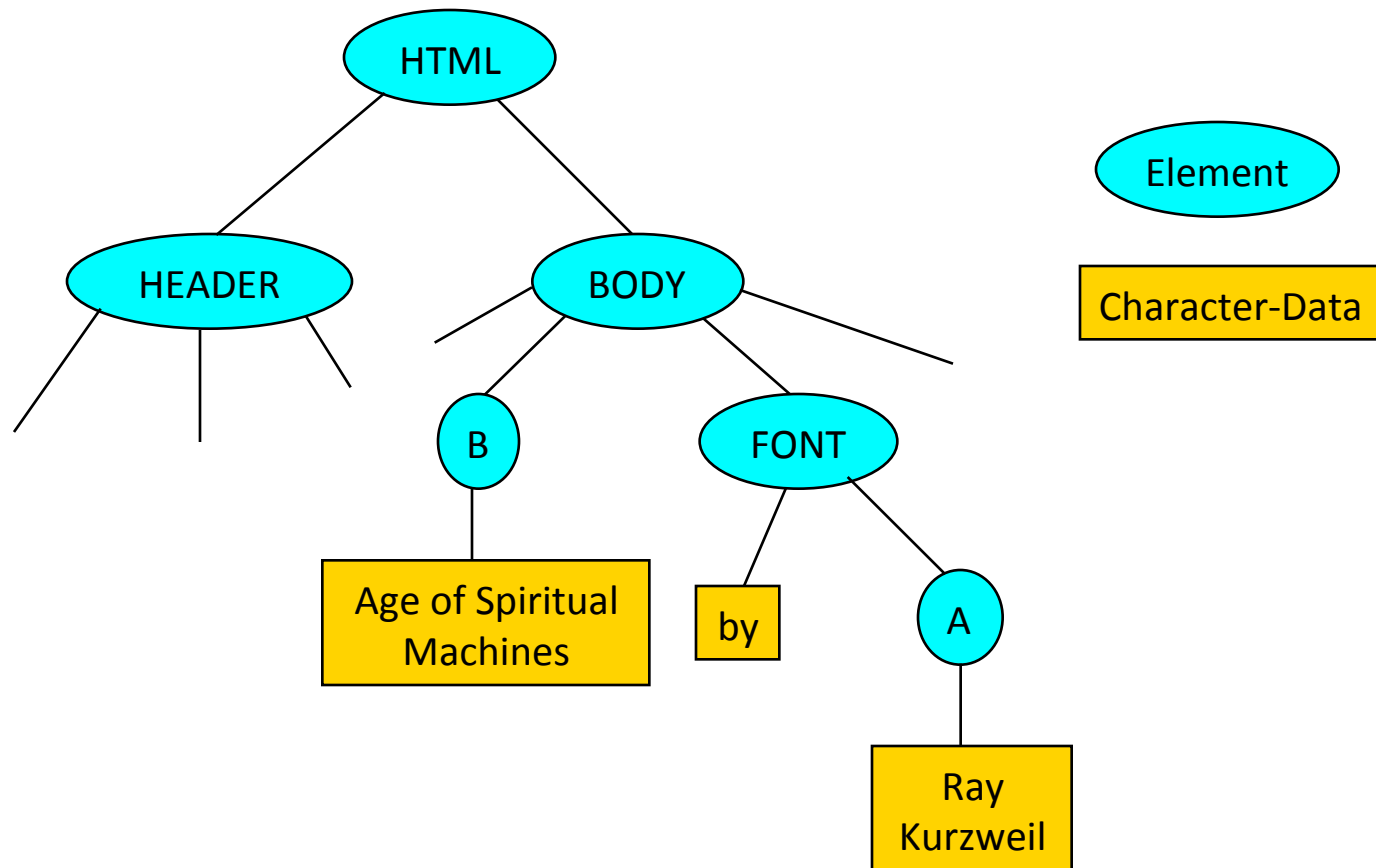


# Extração baseada usando DOM Trees

- 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



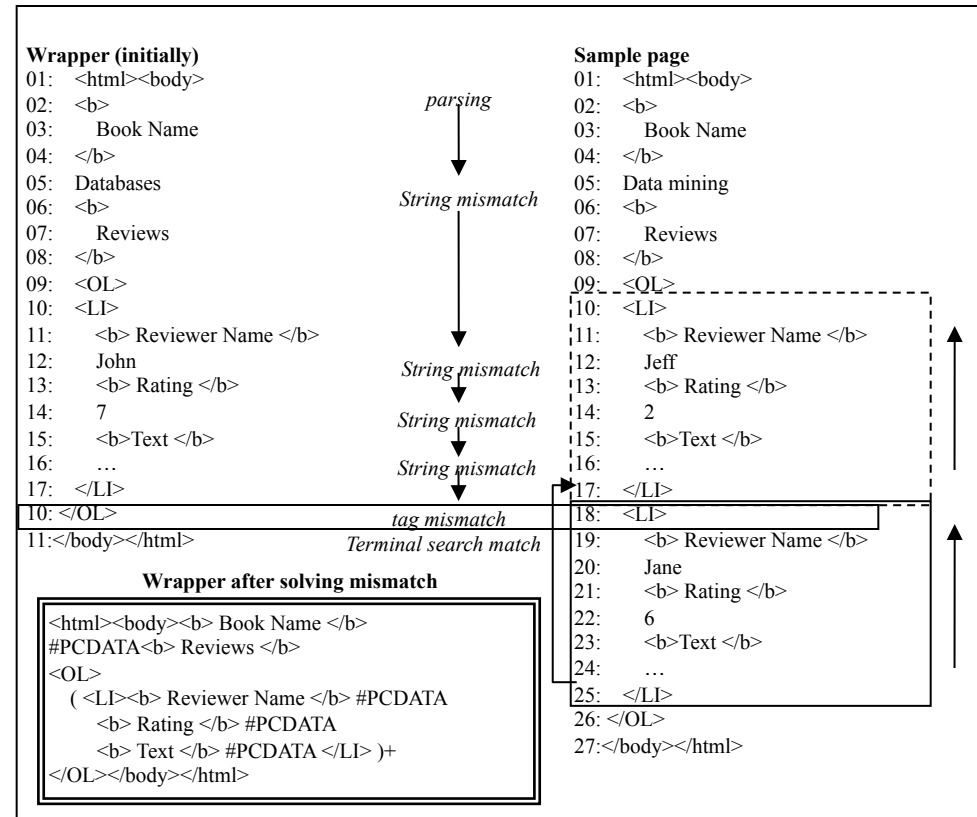
**Title:** HTML → BODY → B → CharacterData

**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

The screenshot shows a TripAdvisor forum thread titled "Andorra in Late May" dated April 10, 2014, 12:50 PM. The main post by user "clinton Boston..." asks for advice on what to do in Andorra in late May, mentioning they are traveling with an 18-month-old daughter and looking for a mid-range hotel. Three posts are highlighted with blue arrows pointing to the word "Registros" on the left:

- Post 1:** "Andorra in Late May" by "clinton Boston..." (Apr 10, 2014, 12:50 PM). Content: "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." (13 posts, 1 review).
- Post 2:** "1. Re: Andorra in Late May" by "williams305 Liverpool" (Apr 11, 2014, 1:02 PM). Content: "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..." (302 posts, 1 review).
- Post 3:** "2. Re: Andorra in Late May" by "YukeTheTraveler Finland" (Apr 11, 2014, 2:57 PM). Content: "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" (883 posts, 10 reviews).

The right sidebar contains various travel-related links and resources, including "Beyond destination forums", "Explore the world! TripAdvisor has reviews and information on over 400,000 locations, including:", "Popular cities", and "Explore other Andorra resources:".



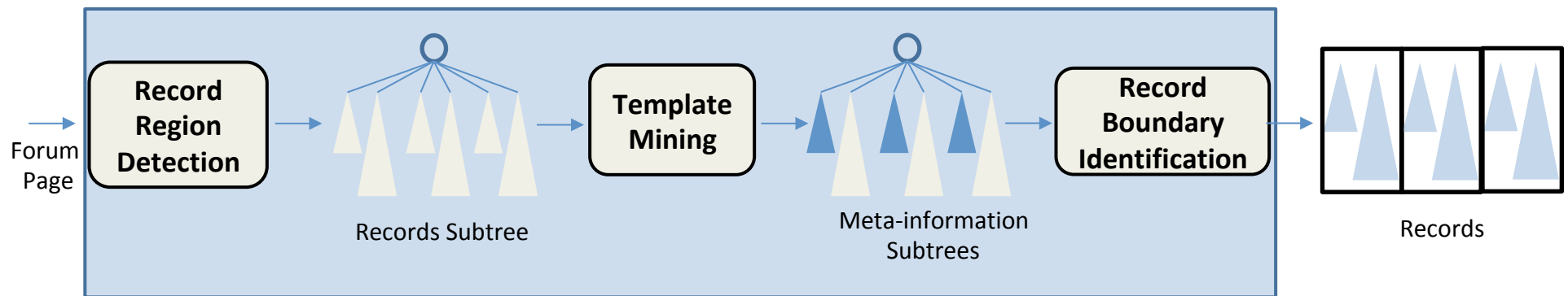


# Extração de Registros de Fóruns

- 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





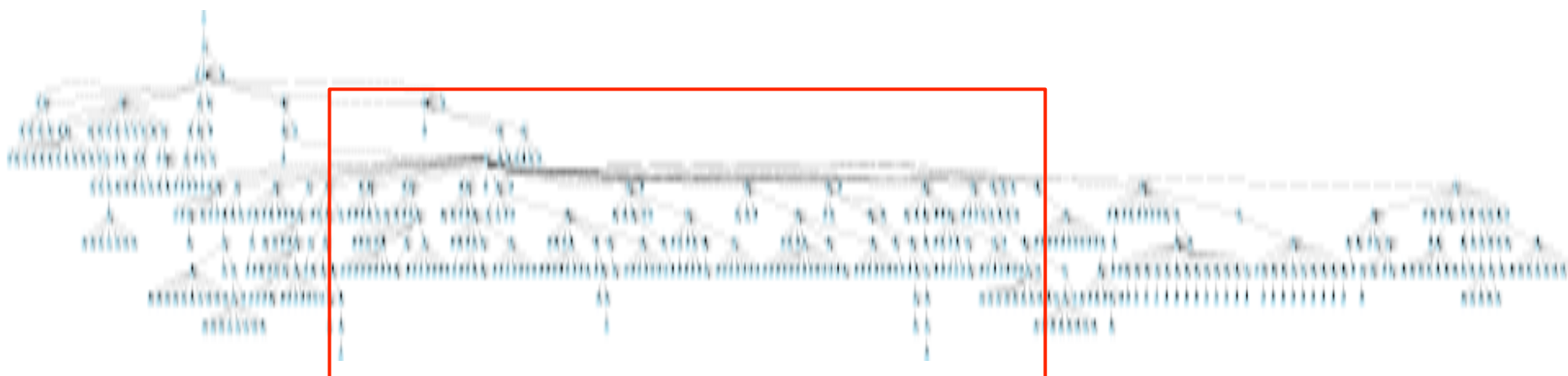
# Detecção da Região do Registro

- 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 "cinfin" on April 10, 2014. The post asks for advice on what to do in Andorra in late May, mentioning a 18-month-old daughter and a search for a mid-range hotel. Below the main post, there are two replies. The first reply, by "williams2005", 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 information, including "Beyond destination forums", "Explore the world!", "Travel Destinations", and "Popular cities".



# 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

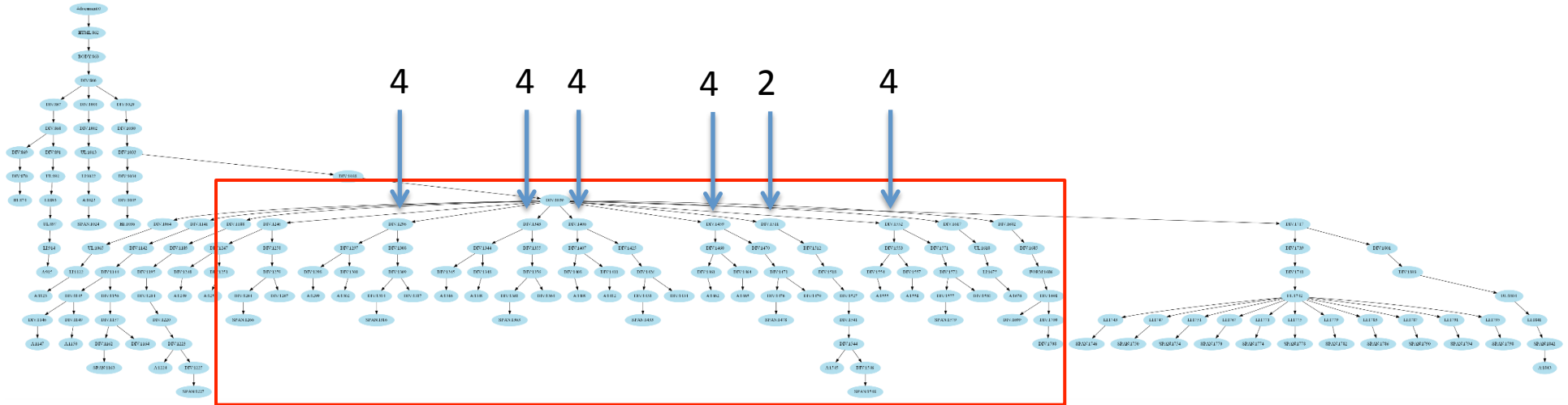
## Meta-informação

- Meta-informação (ou template): detectores de usuário, data/tempo e título

The screenshot shows a forum thread on TripAdvisor. The main post is titled "Andorra in Late May" and was posted by user "cristofon" on April 10, 2014, at 12:50 PM. The post content discusses a family trip to Andorra in May, seeking advice on where to stay and what to do. Below the main post, there are two replies. The first reply, by user "williams3305", is dated April 11, 2014, at 1:02 PM and discusses the difficulty of finding a good place to stay in Andorra. The second reply, by user "YuleTheTraveler", is dated April 11, 2014, at 2:57 PM and provides a link to a website with more information about Andorra. The thread also includes a section for "Travelers interested in this topic also viewed..." and a list of "Beyond destination forums".

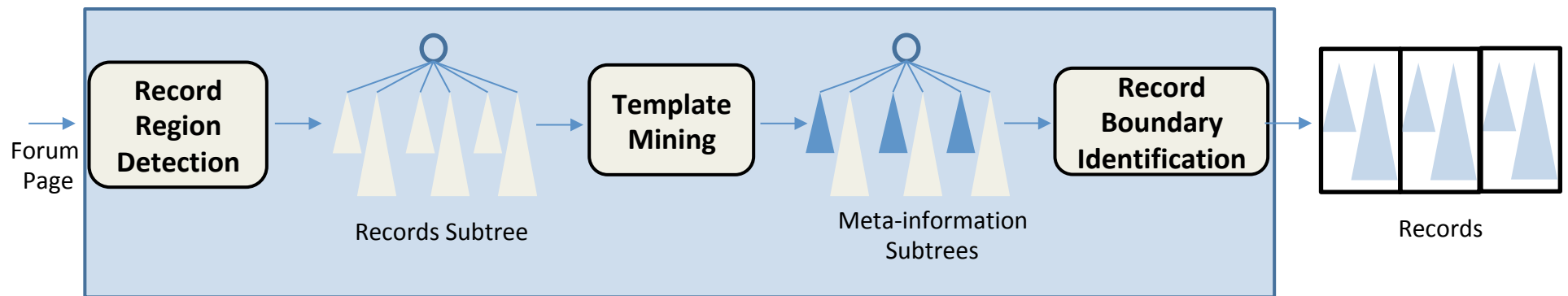


# Nós-Folha Detectados e seus Ascendentes





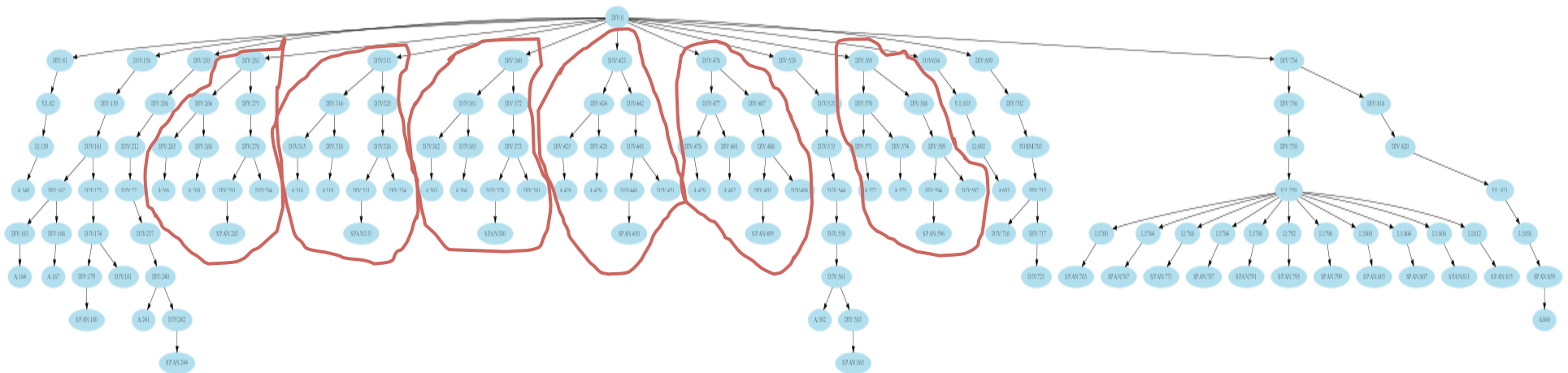
# Mineração de Templates





# Mineração de Templates


- Suposição: árvores de meta-informação têm estruturas similares







## Registros



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

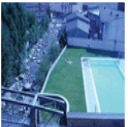
We will be passing through Andorra in late May. We're looking for advice/suggestions on what to do. We are traveling with our 18 month old daughter. We aren't really interested in the shopping, possibly a good area for nature, hiking (nothing too difficult) Any recommendations on a place to stay? We're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




**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](http://hola-andorra.com/animals/...pardinesgb.html)

posts: 383  
reviews: 10

[Reply](#)

[Report inappropriate content](#)

## Anúncio

- 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
- Artrium 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, Rullu

### 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](#)
- [Soleu Hotels](#)

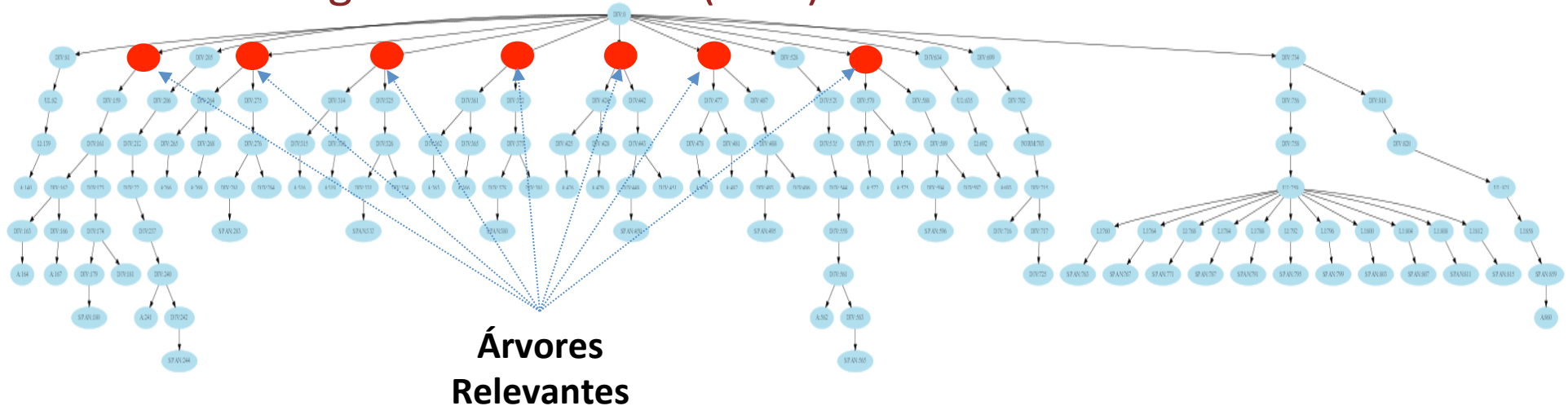


1·DIV·DIV·A·SPAN·I·I·I·DIV·DIV·DIV·



# Mineração de Templates

- Clustering das assinaturas (HAC)



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:

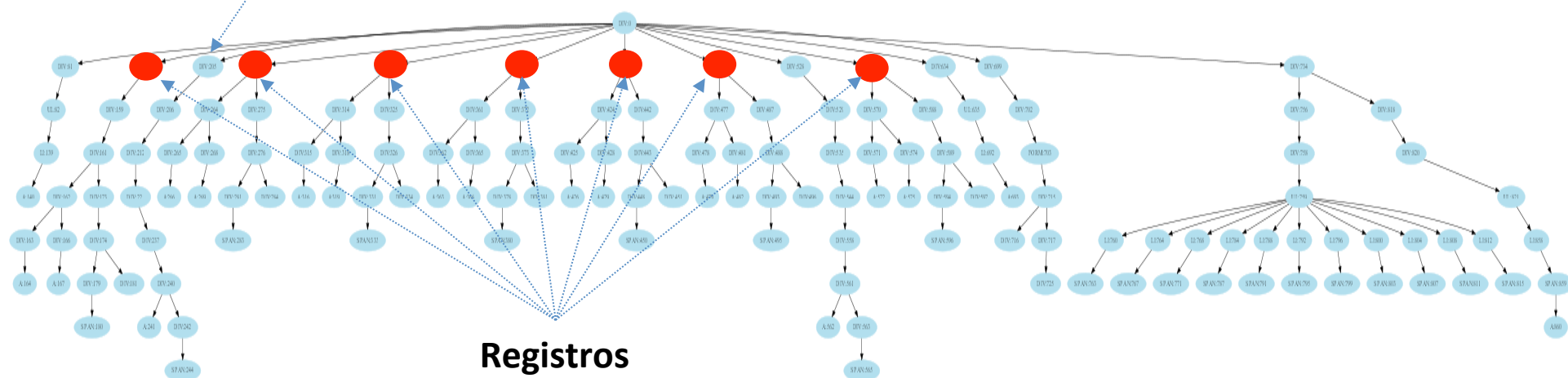
## Maior cluster (Árvores relevantes)

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**

**Registros**





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 ...



# Supervisionado: Extração de Especificações de Produtos

Features	
Color	Blue, Gold, White
Metal	White Gold
Stone	Diamond, Gemstone
Diamond Color	White H-I



Color	Blue, Gold, White
Metal	White Gold
Stone	Diamond, Gemstone
Diamond Color	White H-I



# Extração de Especificações de Produtos

- Dois passos
  - Detecção de tabelas/listas com especificações
  - Extração dos atributos e valores dessas regiões



# Detecção de Especificação

OVERVIEW

SPECS

REVIEWS 121

Q&A 91

ACCESSORIES

table.specTable | 716 × 959

Imaging	
Lens Mount	Sony E-Mount
Camera Format	Full-Frame
Pixels	Actual: 12.4 Megapixel Effective: 12.2 Megapixel

Elements Console Sources Network Timeline Profiles Application Security Audits

<div data-selenium="specWrapper" class="specWrapper">

<table class="specTable" data-selenium="specTable"> == \$0

<tbody data-selenium="specBody">

<tr>...</tr>

<tr>...</tr>

<tr>...</tr>

<tr>...</tr>

<tr>...</tr>

<tr>...</tr>

<tr>...</tr>

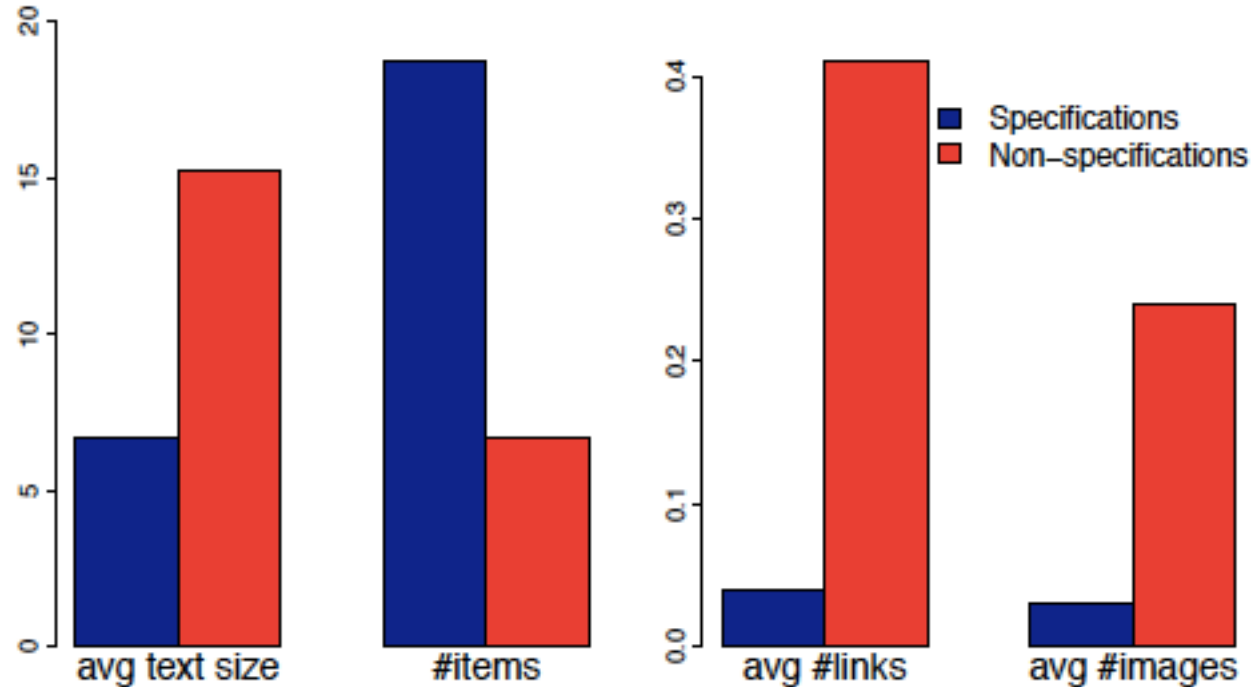
<tr>...</tr>

<tr>...</tr>





# Tabelas/Listas com ou sem Especificações





# Tabelas/Listas com e sem Especificações

Features	
Color	Blue, Gold, White
Metal	White Gold
Stone	Diamond, Gemstone
Diamond Color	White H-I

Com Especificação

		price
0509703PART/R3	<a href="#">Zipp 900 Clincher Rear Wheel</a>	\$1,849.99
0508068PART/R3	<a href="#">Easton EC90 TT 90mm Tubular Wheelset</a>	\$1,799.99
0508654/R3	<a href="#">Shimano Dura-Ace WH-7900-C50-TU Tubular Wheelset</a>	\$2,799.99
1599957/R3C	<a href="#">Wipperman 10S0 10/speed Chain (Shimano Compatible)</a>	\$29.99
GISEC0PART/R3C	<a href="#">Giro Section Helmet '10</a>	\$39.99
1729849/R3C	<a href="#">Kryptonite R4 Retractable Combo Cable Lock</a>	\$16.99
SHM161GPART/R3C	<a href="#">Shimano SH-M161G MTB Cycling Shoe</a>	\$109.99

Sem Especificação



# Extração de Atributos e Valores



## Sony Alpha a7S II Mirrorless Digital Camera (Body Only)

OVERVIEW

SPECS  
▼

REVIEWS 121

Q&A 91

ACCESSORIES

Lens Mount

716 x 63

Sony E-Mount

Camera Format

Full-Frame

Pixels

Actual: 12.4 Megapixel  
Effective: 12.2 Megapixel

```
Elements Console Sources Network Timeline Profiles Application Security Audits
▶<tr>...</tr>
▶<tr>...</tr>
▼<tr>
  <td class="specTopic fs18" data-selenium="specTopic">
    Camera Format
  </td>
  <td class="specDetail fs18" data-selenium="specDetail">
    "
    Full-Frame
    "
  </td>
</tr>
```



# Em Resumo

- Etapa 1: busca pelo nó pai que contém a informação estruturada
- Etapa 2: extração da informação dos filhos



# Avaliando Acurácia da Extração

- Rotular um conjunto de dados extraídos
- Medir
  - Total de extrações possíveis:  $N$
  - Total de pares extraídos pelo sistema:  $E$
  - Total de pares extraídos corretamente:  $C$
- Computar
  - $\text{Recall} = C/N$
  - $\text{Precision} = C/E$
  - $\text{F-Measure} = 2 * (\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision})$



# Chamada

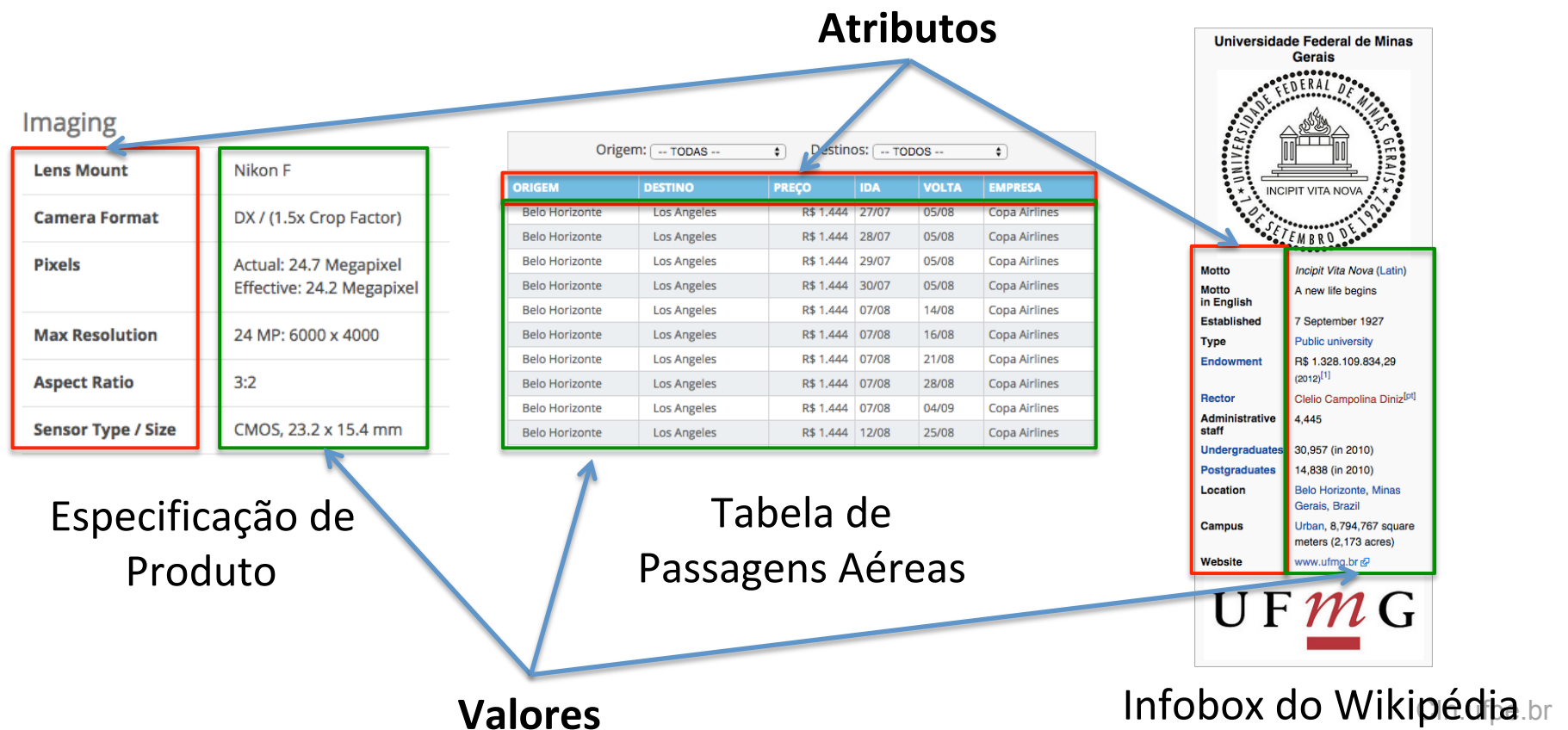


# **Projeto: Coleta e Busca de Entidades Estruturadas em um Domínio**



# Entidade Estruturada

- Def: objeto com atributos e valores associados
- Exemplos:









# Benefícios: Busca Estruturada

O QUE VOCÊ PRECISA?

QUAL TIPO?

ONDE? [Ver todas as localidades](#)

Comprar

Alugar

Lançamentos

Todos os imóveis

BELO HORIZONTE - MG

+

FAIXA DE PREÇO?

QUARTOS

SUÍTES

VAGAS

ÁREA (m²)

35.429 anúncios encontrados

Busca avançada

BUSCAR

Imóveis à venda em Belo Horizonte

Exibir como

Lista

Galeria

Pronto para Morar

A partir de **R\$ 805.246**

**SION**

Rua Patagônia  
Belo Horizonte - Mg

Apartamento Pronto Para Morar  
3 quartos | 1 suíte | 2 vagas |  
102 a 189m<sup>2</sup>

[?]

Atualizado há 14 dias

Em Obras

A partir de **R\$ 329.000**

**JARAGUA**

Rua Professor Jerson Martins  
Belo Horizonte - Mg

Apartamento em Obras  
2 a 3 quartos | 1 suíte |  
1 a 3 vagas | 64 a 176m<sup>2</sup>

[?]

Atualizado há 14 dias

LED & LCD TVs

CURRENT OFFERS

☐ On Sale (131)

☐ Free Shipping Eligible (362)

☐ Special Offers (228)

☐ Outlet Items (85)

TV TYPE

Clear

☒ LED (405)

☐ Smart (218)

☐ 4K UHD (79)

☐ Curved (17)

☐ OLED (3)

☐ 3D (71)

☐ Outdoor (47)

☒ LED Flat-Panel (15)

See More

TV SCREEN SIZE

Find Out Which TV Is Best for You

How do LED, plasma, OLED and LCD TVs compare? What resolution and refresh rate do you need? Simplify your TV search by understanding these and other features to consider. [Learn more in the TV Buying Guide](#)

All Items (423)

Best Buy Items (233)

Marketplace Seller Items (190)

Sort by: Best Selling

Items per page: 15

< 1 2 3 ... 29 >

Filters: LED X LED Flat-Panel X LCD X LCD Flat-Panel X Clear All

**Insignia™ - 32" Class (31-1/2" Diag.) - LED - 720p - HDTV - Black**

Model: NS-32D312NA15 | SKU: 6080010

• 720p resolution

• 60Hz refresh rate

• ENERGY STAR Certified

★★★★★ 4.5 (1,529 Reviews)

PRICE MATCH GUARANTEE

**\$159.99**

ON SALE

SAVE \$20 (Reg. \$179.99)

Add to Cart

Atributos

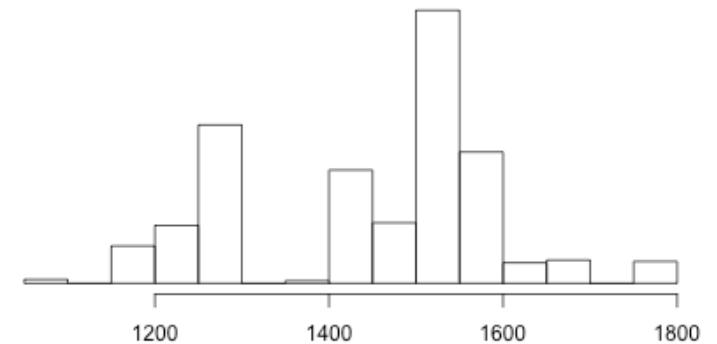


# Benefícios: Análise Estatística

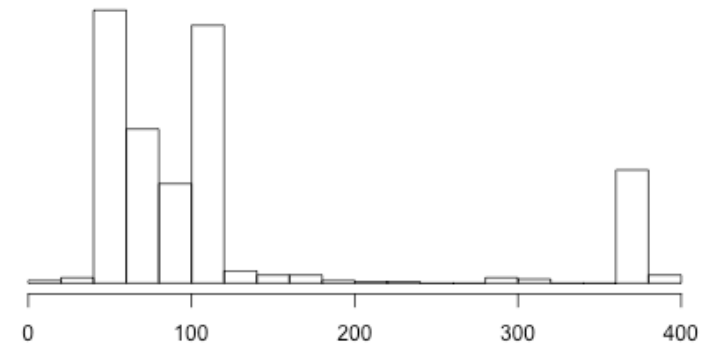
Origem: -- TODAS --		Destinos: -- TODOS --			
ORIGEM	DESTINO	PREÇO	IDA	VOLTA	EMPRESA
Belo Horizonte	Los Angeles	R\$ 1.444	27/07	05/08	Copa Airlines
Belo Horizonte	Los Angeles	R\$ 1.444	28/07	05/08	Copa Airlines
Belo Horizonte	Los Angeles	R\$ 1.444	29/07	05/08	Copa Airlines
Belo Horizonte	Los Angeles	R\$ 1.444	30/07	05/08	Copa Airlines
Belo Horizonte	Los Angeles	R\$ 1.444	07/08	14/08	Copa Airlines
Belo Horizonte	Los Angeles	R\$ 1.444	07/08	16/08	Copa Airlines
Belo Horizonte	Los Angeles	R\$ 1.444	07/08	21/08	Copa Airlines
Belo Horizonte	Los Angeles	R\$ 1.444	07/08	28/08	Copa Airlines
Belo Horizonte	Los Angeles	R\$ 1.444	07/08	04/09	Copa Airlines
Belo Horizonte	Los Angeles	R\$ 1.444	12/08	25/08	Copa Airlines



Ticket Price (Median = R\$ 1528 , Min = R\$ 1078 )



Days prior to the trip (Median= 83 days)





# Benefícios: Mercado de Dados



## RESTAURANTS

43 restaurant specific attributes for restaurants of every type in the US, UK, France, Germany, and Australia.

[LEARN MORE](#)



## DOCTORS

Database of over 1 million physician, dentist, and healthcare provider listings with the key data you need to make informed decisions.

[LEARN MORE](#)



## HOTELS

Database of 140,000 hotel listings with over 35 attributes covering everything you need to know about a hotel.

[LEARN MORE](#)



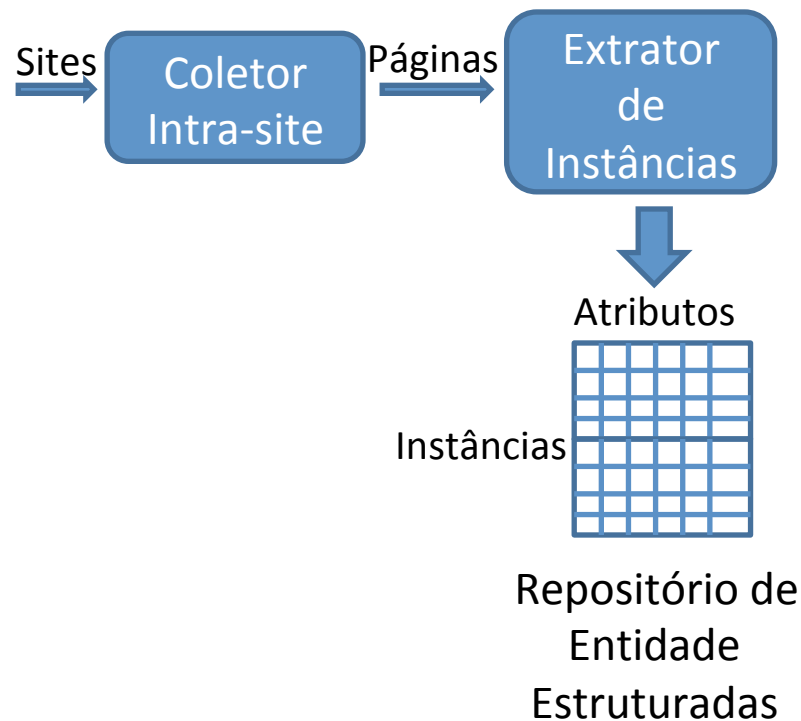
# Grande Interesse da Indústria



infochimps



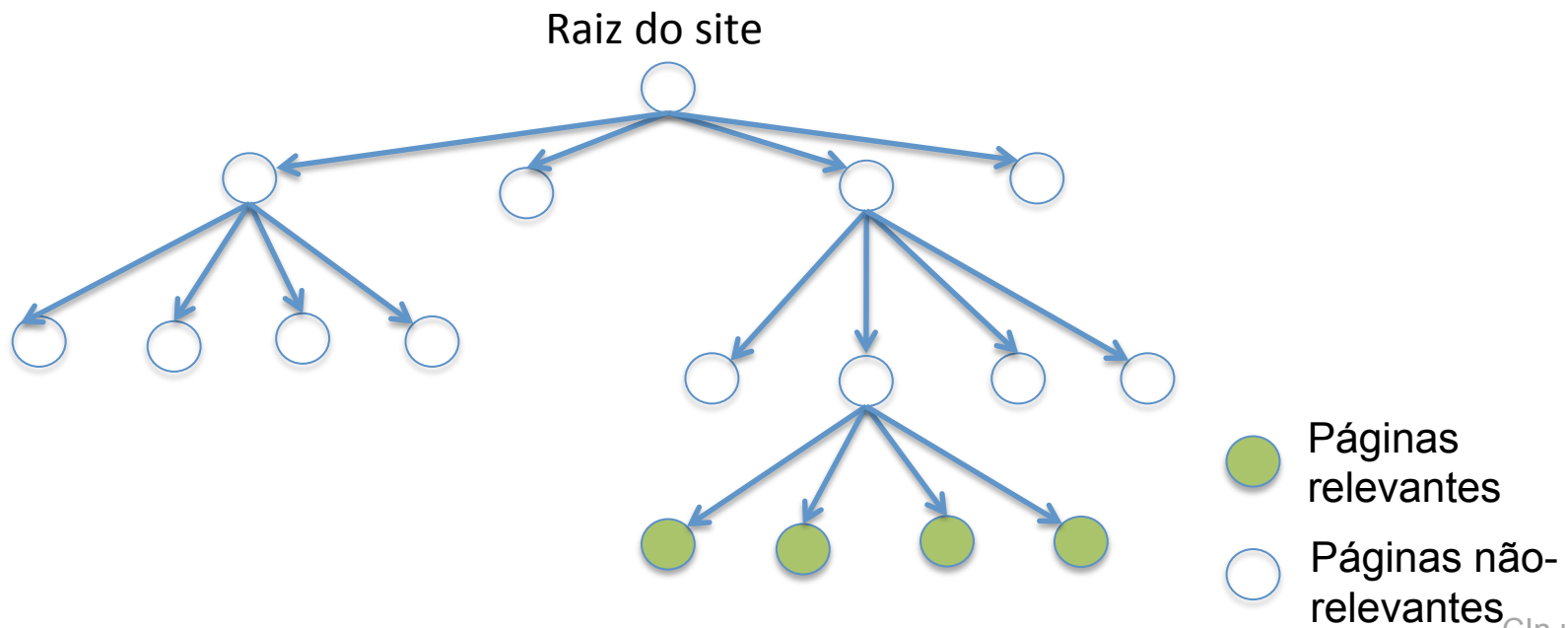
# Coletor Focado em Entidades





# Coletor Intra-Site

- 2 tarefas:
  1. Localizar páginas relevantes (Seletor de Links)
  2. Detectar páginas relevantes (Seletor de Páginas)





# Tarefa 1: Localizar Páginas Relevantes

- Desafio: evitar regiões não produtivas do site

OLX

Buscar Ajuda Meus Anúncios Lojas Minha conta Inserir anúncio GRATIS

Buscar por palavra-chave

☐ Procurar pelo título do anúncio

Busca por categorias

- Animais e acessórios
- Bebês e crianças
- Música e hobbies
- Moda e beleza
- Para a sua casa
- Esportes
- Eletrônicos e
- Imóveis**
- Empregos e
- Veículos e barcos

★ Salvar busca

"OLX bom demais, eu recomendo e garanto!"  
jonathan, 22/06/2015 [Veja mais](#)

Brasil > RJ

DDD 21 - Rio de Janeiro e região, 1.065.367 DDD 24 - Serra, Angra dos Reis e região, 54.220  
DDD 22 - Norte do Estado e Região dos Lagos, 148.310

Bicicleta KHS 27.5 Top R\$ 4.300

Maraville Pronto para Morar R\$ 1.300

Lançamento em Jacarepaguá Up Barra Mais R\$ 319.000

Sobre Galeria

Todos 1.267.897 Particular 750.309 Profissional 517.588 Ordenar por preço

Minerais e Tijolos Tudo Direto Hoje 18:10  
Rio de Janeiro, Campo Grande - DDD 21  
Jardinagem e construção

Cornetão de fibra roscável para Drive, marca Selenium modelo HC52-26 Hoje 18:10  
Rio de Janeiro, Penha Circular - DDD 21  
Áudio, TV, vídeo e fotografia

Branco pro driver invicta R\$ 599 Hoje 18:10  
Rio de Janeiro, Campo Grande - DDD 21  
Bijouterias, relógios e acessórios





# Tarefa 1: Localizar Páginas Relevantes

1. Encontrar manualmente 10 sites no domínio
2. Implementar 2 estratégias (1000 páginas visitadas por site):
  - Baseline: busca em largura
  - Heurística (usar âncora)
  - **Extra:** implementar um classificador de links
3. Comparar estatísticas:
  - Harvest ratio: (número de páginas relevantes coletadas)/(total de páginas visitadas)
  - **Mostrar tabela com resultados**
- Importante:
  - Evitar sobrecarregar o site
  - Respeitar o robots.txt
  - Detectar o conteúdo da página com o campo Content-Type



# Tarefa 2: Detectar Páginas com Instâncias

**OLX** Buscar Ajuda Meus Anúncios Lojas Minha conta Inserir anúncio

RJ > Rio de Janeiro e região > Venda > Apartamentos > Zona Oeste > Vila Valqueire

**Apartamento - Condomínio Nova Valqueire**  
Inserido em: 27 Junho 18:18.

**R\$750.000**

carlos  
(21) 9814 ... ver número

Seu nome  
Seu e-mail  
Seu telefone (Opcional)  
Mensagem

☐ Envie-me uma cópia

**Enviar mensagem**

**Dicas de Segurança**

- Evite pagar adiantado
- Desconfie de anúncios não realistas

**Favoritos** **Denunciar** **Compartilhar**

**Editar** **Excluir** **Topo**

**Preço: R\$750.000** [Simular financiamento](#)

Apartamento no Condomínio Nova Valqueire, com 8 anos de idade, 140 m², prédio com 4 andares e 12 unidades, móveis planejados da FAVO nos quartos e cozinha, dependências de empregada, porcelanato em todos os ambientes, banheira de hidromassagem, toldos nas varandas, fogão embutido e depurador, aparelhos de ar condicionado na sala e quartos, 2 vagas cobertas de garagem, portas com puxadores e fechaduras biométricas/senhas.  
Preço diretamente com o proprietário. R\$750.000,00  
Preço com corretores R\$800.000,00  
Estudo Proposta

**Características:** Ar condicionado, armários embutidos, varanda, área de serviço, quarto de empregada

**Detalhes do imóvel**

- Tipo: Venda - apartamento padrão
- Condomínio: R\$ 500
- IPPU: R\$ 1381
- Área útil: 140 m²
- Quartos: 3
- Vagas na garagem: 2



## Classificados



buscar

### 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

Infonet → Classificados → Imóveis → Apartamentos para vender

### Belíssimo Condomínio Soberano Jardins.



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

[Ver os anúncios deste anunciante](#)

Marcar esse anúncio como: Categoria errada Anúncio proibido

[Copiar URL](#)

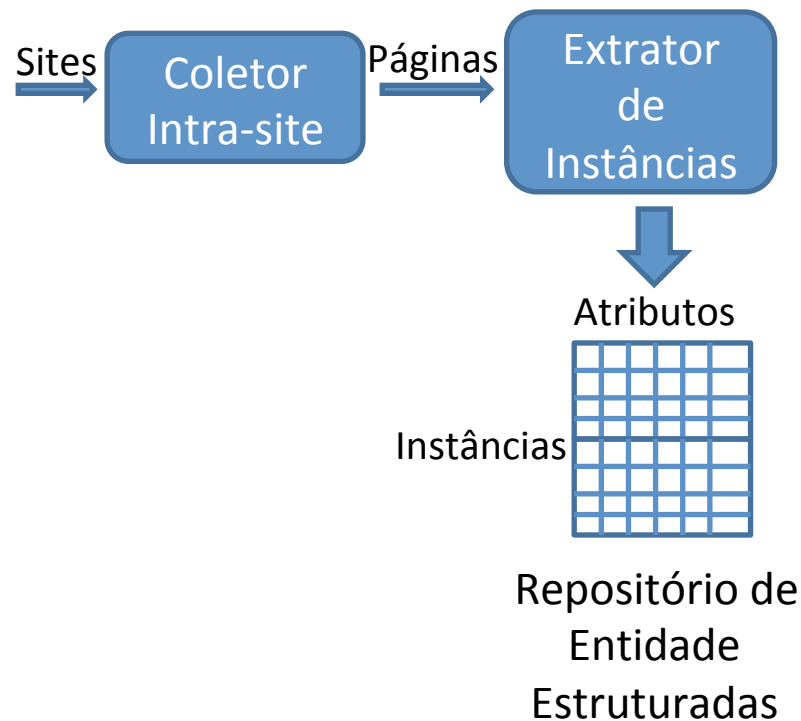


## Tarefa 2: Detectar Páginas com Instâncias

1. Rotular exemplos positivos e negativos (10 positivos e 10 negativos por site)
2. Criar o conjunto de features (ex.: bag of words) usando feature selection (ex. frequência ou information gain)
3. Treinar o classificador com uma ferramenta de ML (ex.: scikit-learn, weka etc)
  - Métodos: Naïve bayes, Decision tree (J48), SVM (SMO), Logistic regression (logistic), Multilayer perceptron
  - Extra: otimizar hiper-parâmetros e diagnosticar modelos
4. Comparar estratégias:
  - Accuracy, precision e recall
  - Tempo de treinamento
  - **Mostrar tabela com os resultados**



# Coletor Focado em Entidades





# Tarefa 3: Extrair Instâncias com seus Valores e Atributos

**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**

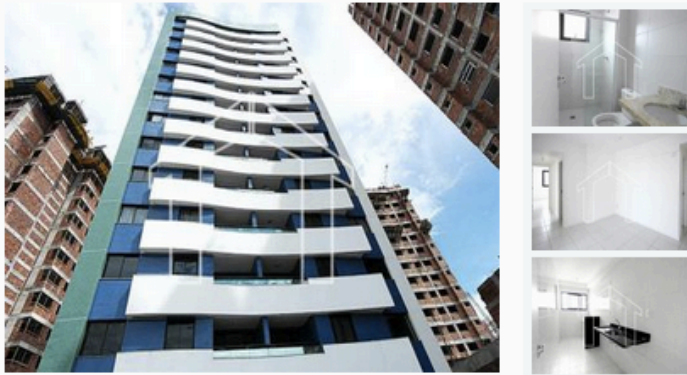
- Agenda
- Cinema
- Eventos
- Promoções

**Especiais**

- Imposto de Renda

**Serviços**

**Belíssimo Condomínio Soberano Jardins.**



Belíssimo apto c/ 3/4, suite, 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.taiguaraimeveis.com.br](http://www.taiguaraimeveis.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	1.400.000
Contato	(79)9828-1120
Telefone	(79)9828-1120



# Tarefa 3: Extrair Instâncias com seus Valores e Atributos

1. Criar um wrapper para cada site
  - Criação do conjunto rotulado
2. Implementar uma solução que funcione para todos os sites
  - Ex: detectores de tipos do domínio
  - **Extra:** implementar mais uma solução e fazer análise de erros
3. Comparar estratégias:
  - Accuracy, precision e recall
  - **Mostrar tabela com os resultados**

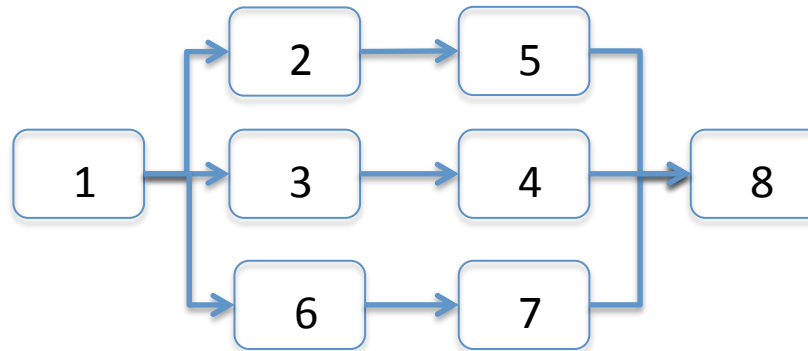


- Criar projeto no github
- Adicionar como colaborador: ProfLuciano
- Colocar código, dados (conteúdo das páginas e informação extraída) e apresentação



# Tarefas

1. Encontrar 10 sites no domínio
2. Crawling: implementar busca em largura
3. Classificação: rotular exemplos positivos e negativos
  - Positivo: página com entidade estruturada
4. Classificação: criar classificador de páginas
5. Crawling: implementar heurística
6. Extração: criar wrappers para cada site
7. Extração: implementar único wrapper que funcione para todos os sites do domínio
8. Integrar crawler e classificador (medir harvest ratio)







# Pontos Importantes

- Saída do sistema
  - Páginas coletadas
  - Resultado da extração das páginas
- Para a apresentação do projeto, devem ser preparados slides
- Apresentar possíveis problemas encontrados e discutir os resultados
- Presença nas aulas de acompanhamento obrigatória para todos os integrantes do grupo
  - Pontualidade é importante
  - Conta para a nota de participação
- O integrante que não fizer sua parte no projeto recebe 0