



# Indexação de Texto

Prof. Luciano Barbosa

(Parte do material retirado dos slides dos livros adotados)



UNIVERSIDADE  
FEDERAL  
DE PERNAMBUCO

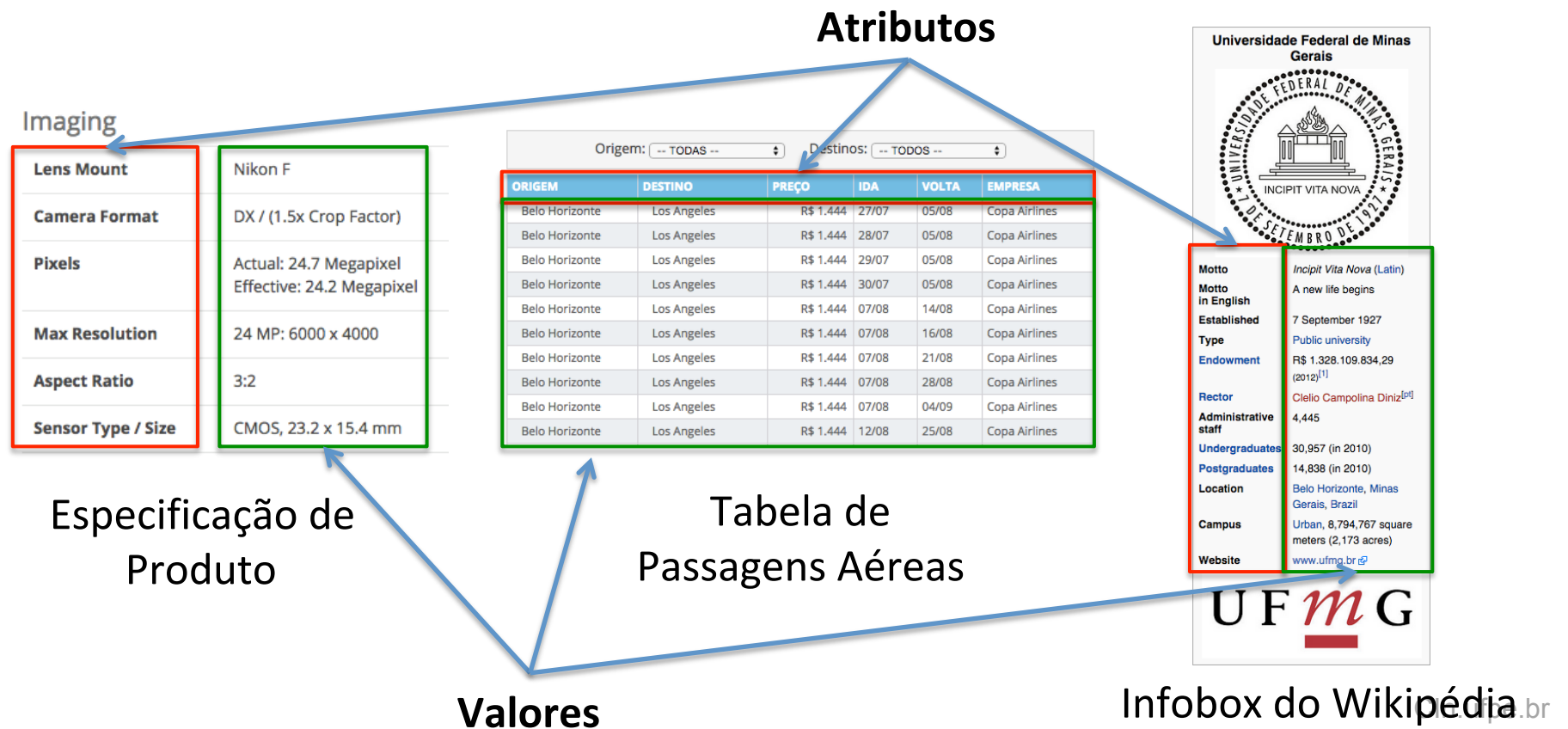


# Projeto: Coleta e Busca de Entidades Estruturadas em um Domínio



# Entidade Estruturada

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







# Benefícios: Busca Estruturada

O QUE VOCÊ PRECISA?

Comprar Alugar Lançamentos Todos os imóveis

QUAL TIPO?

ONDE? Ver todas as localidades

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

Filters: LED LED Flat-Panel LCD LCD Flat-Panel [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](#)

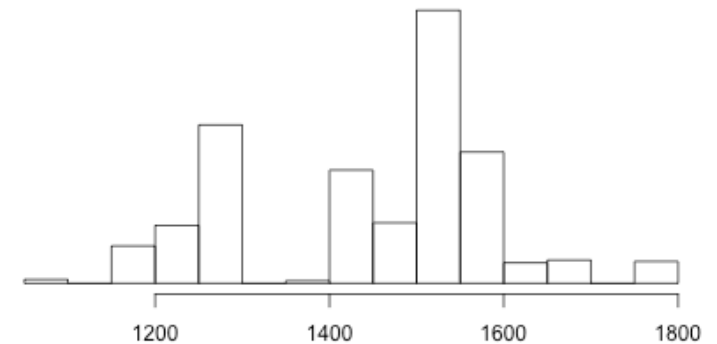


# Benefícios: Análise Estatística

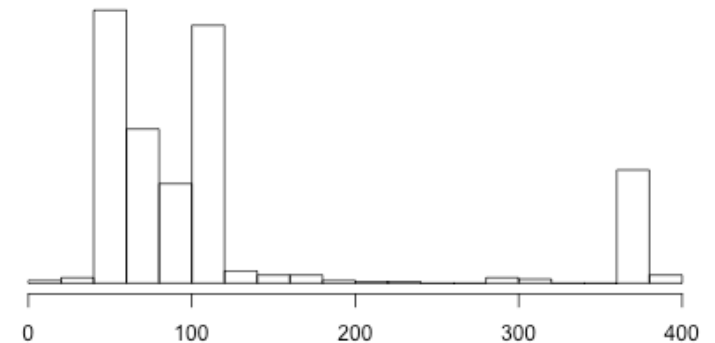
Origem: -- TODAS --		Destinos: -- TODOS --			
ORIGEM	DESTINO	PREÇO	IDA	VOLTA	EMPRESA
Belo Horizonte	Los Angeles	R\$ 1.444	27/07	05/08	Copa Airlines
Belo Horizonte	Los Angeles	R\$ 1.444	28/07	05/08	Copa Airlines
Belo Horizonte	Los Angeles	R\$ 1.444	29/07	05/08	Copa Airlines
Belo Horizonte	Los Angeles	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

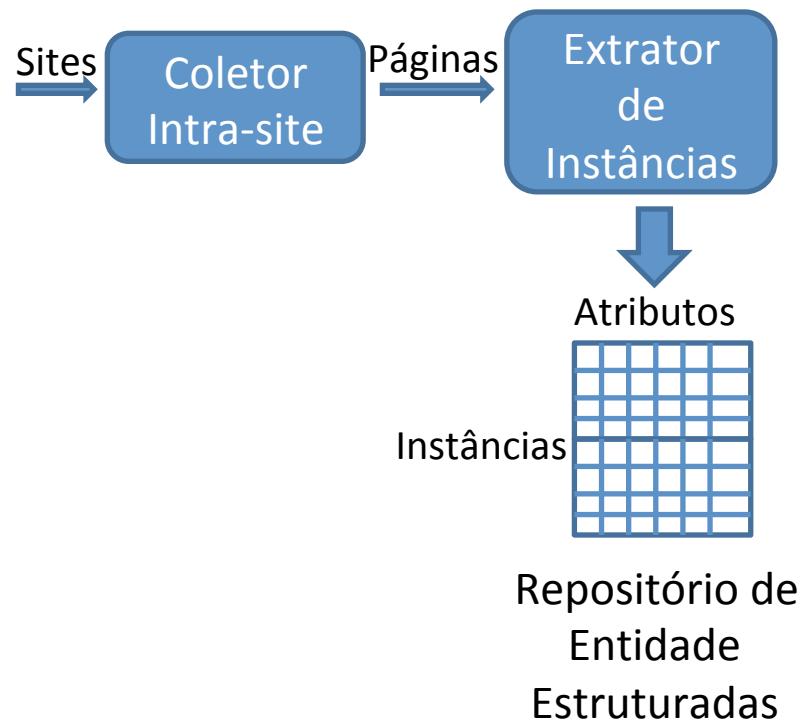


infochimps





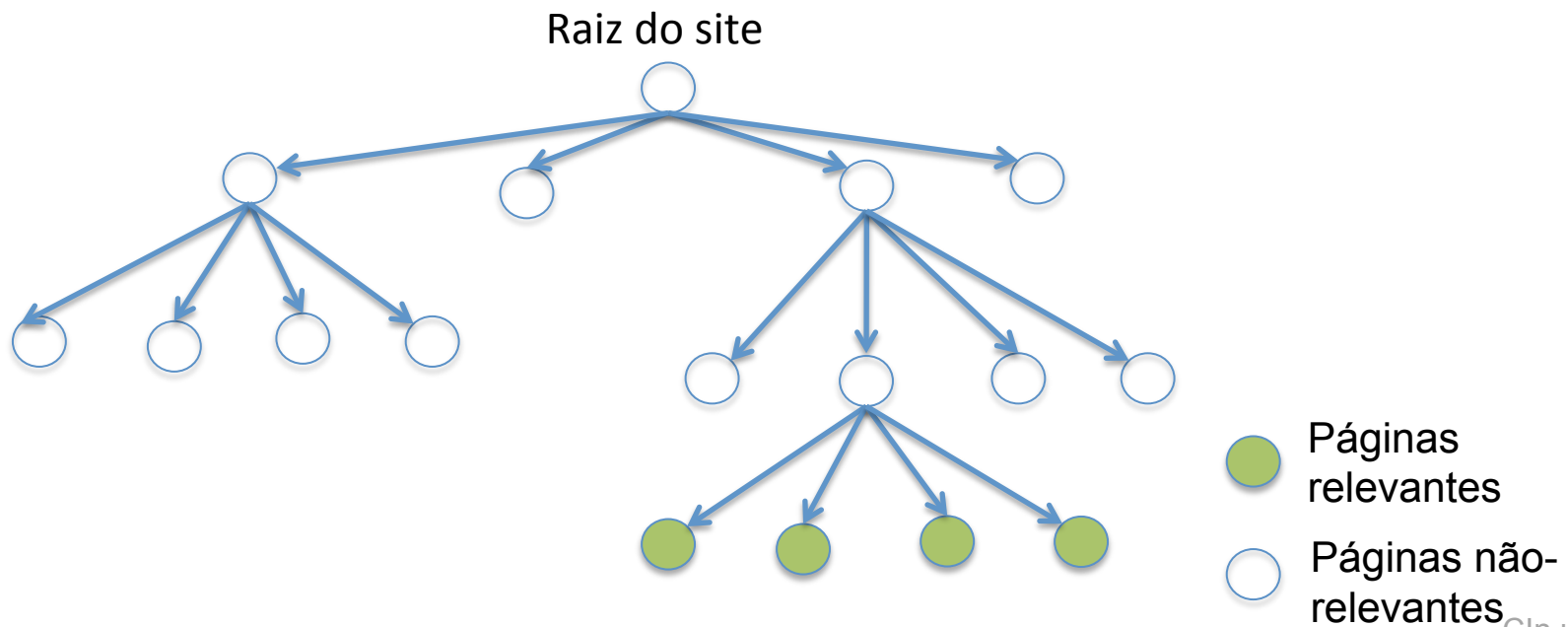
# Coletor Focado em Entidades





# 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

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)
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](#)

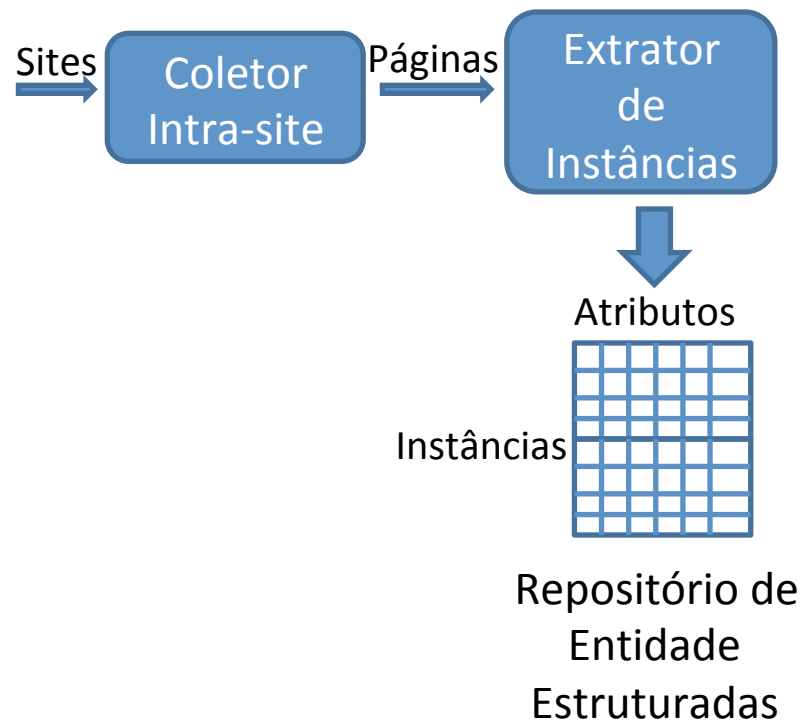


## Tarefa 2: Detectar Páginas com Instâncias

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
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**

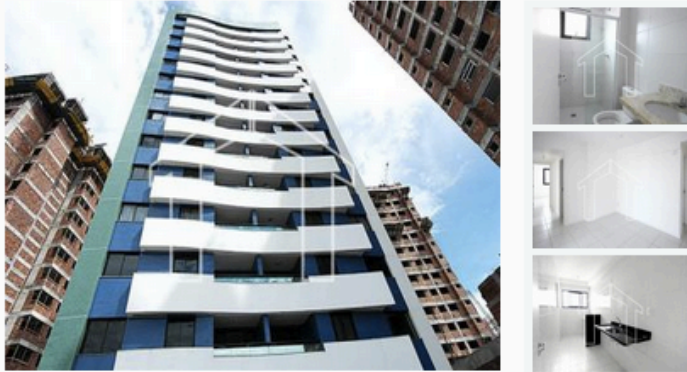
- Agenda
- Cinema
- Eventos
- Promoções

**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

**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
3. Comparar estratégias:
  - Accuracy, precision e recall
  - **Mostrar tabela com os resultados**

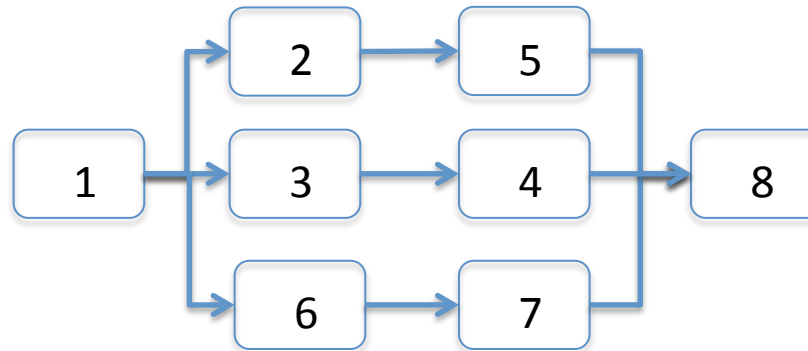


- Criar conta no github: <https://education.github.com/pack>
- Adicionar como colaborador: ProfLuciano
- Colocar código e dados (conteúdo das páginas e informação extraída)



# 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)



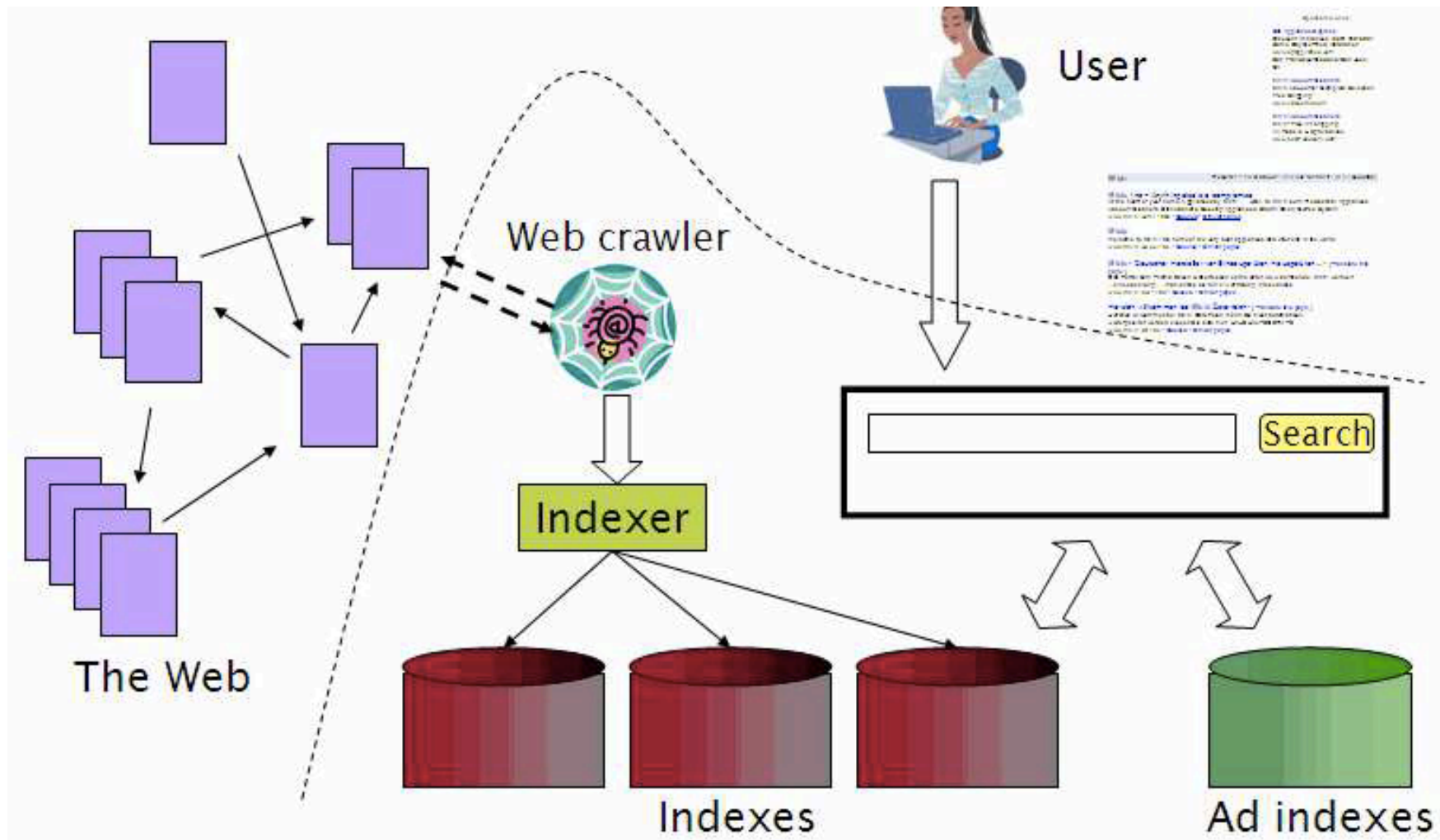


# Pontos Importantes

- Saída do sistema
  - Páginas coletadas
  - Resultado da extração das páginas
- Para a apresentação do projeto, devem ser preparados slides
- 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



# Visão Geral de um Engenho de Busca





# Índices

- Estrutura de dados que aumentam a velocidade da busca
- Importante para aplicações de larga escala: consultas e dados
- Métricas de avaliação
  - Espaço em disco
  - Velocidade de busca
  - Velocidade de indexação



# Abordagem Simples: Grep

- Quais peças de Shakespeare contêm as palavras “Brutus” e “Cesar” mas não “Calpurnia”?
- Fazer um grep por “Brutus” e “Cesar” e remover as linhas com “Calpurnia”
- Por que o grep não é a melhor solução?
  - Devagar (para grandes coleções)
  - grep é orientado à linha, RI é orientado a documento
  - Outros operadores não funcionam (ex. ‘near’)



# Matrix termo-documento

## Documentos

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	...
ANTHONY	1	1	0	0	0	1	
BRUTUS	1	1	0	1	0	0	
CAESAR	1	1	0	1	1	1	
CALPURNIA	0	1	0	0	0	0	
CLEOPATRA	1	0	0	0	0	0	
MERCY	1	0	1	1	1	1	
WORSER	1	0	1	1	1	0	
...							

## Termos





# Matrix termo-documento

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	...
ANTHONY	1	1	0	0	0	1	
BRUTUS	1	1	0	1	0	0	
CAESAR	1	1	0	1	1	1	
CALPURNIA	0	1	0	0	0	0	
CLEOPATRA	1	0	0	0	0	0	
MERCY	1	0	1	1	1	1	
WORSER	1	0	1	1	1	0	
...							

Brutus AND Cesar NOT Calpurnia



# Matrix termo-documento

- Uma coleção de 1 milhão de documentos cada um com 1000 termos -> 1 bilhão de termos
- Assuma  $M=500.000$  termos distintos na coleção
- Tamanho da matrix =  $500.000 \times 10^6 =$  meio trilhão
  - Não mais que 1 bilhão de 1s
- Esparsa e ocupa muito espaço
- Melhor representação: armazenar somente os 1s



# Índice Invertido

BRUTUS	→	1	2	4	11	31	45	173	174	
CAESAR	→	1	2	4	5	6	16	57	132	...
CALPURNIA	→	2	31	54	101					

**Vocabulário**

**Postings**

- Duas partes
  - Vocabulário: conjunto de palavras únicas nos textos
  - Postings:
    - Cada documento tem um identificador único (número)
    - Lista de docs onde a palavra ocorre (ordenada pelo identificador)



# Índice Invertido Básico

BRUTUS → 

1	2	4	11	31	45	173	174
---	---	---	----	----	----	-----	-----

CAESAR → 

1	2	4	5	6	16	57	132	...
---	---	---	---	---	----	----	-----	-----

CALPURNIA → 

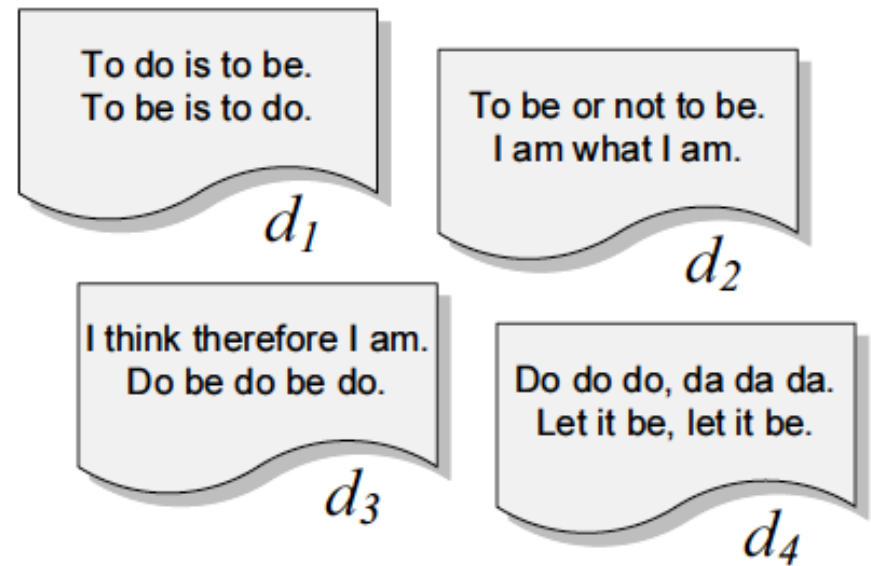
2	31	54	101
---	----	----	-----

- Suporta consultas booleanas
  - Ex.: Brutus and Cesar not Calpurnia
  - And: procura termos e acha intersecção nos postings
  - OR: procura termos e remove duplicatas
  - NOT: remove termos
  - Linear no tamanho dos postings se tiverem ordenados



# Índice Invertido com Frequência

Vocabulary	$n_i$	Occurrences as inverted lists
to	2	[1,4],[2,2]
do	3	[1,2],[3,3],[4,3]
is	1	[1,2]
be	4	[1,2],[2,2],[3,2],[4,2]
or	1	[2,1]
not	1	[2,1]
I	2	[2,2],[3,2]
am	2	[2,2],[3,1]
what	1	[2,1]
think	1	[3,1]
therefore	1	[3,1]
da	1	[4,3]
let	1	[4,2]
it	1	[4,2]



- Suporta consultas com ranqueamento



# Índice Invertido Posicional

Vocabulary	$n_i$
to	2
do	3
is	1
be	4
or	1
not	1
I	2
am	2
what	1
think	1
therefore	1
da	1
let	1
it	1

Occurrences as full inverted lists

[1,4,[1,4,6,9]], [2,2,[1,5]]

[1,2,[2,10]], [3,3,[6,8,10]], [4,3,[1,2,3]]

[1,2,[3,8]]

[1,2,[5,7]], [2,2,[2,6]], [3,2,[7,9]], [4,2,[9,12]]

[2,1,[3]]

[2,1,[4]]

[2,2,[7,10]], [3,2,[1,4]]

[2,2,[8,11]], [3,1,[5]]

[2,1,[9]]

[3,1,[2]]

[3,1,[3]]

[4,3,[4,5,6]]

[4,2,[7,10]]

[4,2,[8,11]]

To do is to be.  
To be is to do.

$d_1$

To be or not to be.  
I am what I am.

$d_2$

I think therefore I am.  
Do be do be do.

$d_3$

Do do do, da da da.  
Let it be, let it be.

$d_4$



# Índice Invertido Posicional

Vocabulary	$n_i$
to	2
do	3
is	1
be	4
or	1
not	1
I	2
am	2
what	1
think	1
therefore	1
da	1
let	1
it	1

Occurrences as full inverted lists

[1,4,[1,4,6,9]], [2,2,[1,5]]

[1,2,[2,10]], [3,3,[6,8,10]], [4,3,[1,2,3]]

[1,2,[3,8]]

[1,2,[5,7]], [2,2,[2,6]], [3,2,[7,9]], [4,2,[9,12]]

[2,1,[3]]

[2,1,[4]]

[2,2,[7,10]], [3,2,[1,4]]

[2,2,[8,11]], [3,1,[5]]

[2,1,[9]]

[3,1,[2]]

[3,1,[3]]

[4,3,[4,5,6]]

[4,2,[7,10]]

[4,2,[8,11]]

consulta: "to do"

To do is to be.  
To be is to do.

$d_1$

To be or not to be.  
I am what I am.

$d_2$

I think therefore I am.  
Do be do be do.

$d_3$

Do do do, da da da.  
Let it be, let it be.

$d_4$



# Índice Invertido Posicional

- Suporta consultas com ranqueamento e proximidade
- Demanda espaço extra: 40% sem stopwords e 80% com stopwords
- É o padrão usado por engenhos de busca





# Outras Formas

- Lista pré-computada com scores
  - Ex.: carro [(1:3.6),(3:2.2)], 3.6 é o score do termo carro no documento 1
  - Melhora a velocidade mas reduz flexibilidade
- Listas de scores ordenadas
  - Processamento da consulta foca somente no começo das listas
  - Bastante eficiente para consultas de palavras únicas



# Construção do Índice

- Tokenização e pré-processamento

**Doc 1.** I did enact Julius Caesar: I was killed i' the Capitol; Brutus killed me.

**Doc 2.** So let it be with Caesar. The noble Brutus hath told you Caesar was ambitious:



**Doc 1.** i did enact julius caesar i was killed i' the capitol brutus killed me

**Doc 2.** so let it be with caesar the noble brutus hath told you caesar was ambitious



# Criação dos IDs dos Documentos

**Doc 1.** i did enact julius caesar i was  
killed i' the capitol brutus killed me

**Doc 2.** so let it be with caesar the  
noble brutus hath told you caesar was  
ambitious



term	docID
i	1
did	1
enact	1
julius	1
caesar	1
i	1
was	1
killed	1
i'	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2
told	2
you	2
caesar	2
was	2
ambitious	2



# Ordenar os Termos

term	docID		term	docID
i	1		ambitious	2
did	1		be	2
enact	1		brutus	1
julius	1		brutus	2
caesar	1		capitol	1
i	1		caesar	1
was	1		caesar	2
killed	1		caesar	2
i'	1		did	1
the	1		enact	1
capitol	1		hath	1
brutus	1		i	1
killed	1		i	1
me	1	⇒	i'	1
so	2		it	2
let	2		julius	1
it	2		killed	1
be	2		killed	1
with	2		let	2
caesar	2		me	1
the	2		noble	2
noble	2		so	2
brutus	2		the	1
hath	2		the	2
told	2		told	2
you	2		you	2
caesar	2		was	1
was	2		was	2
ambitious	2		with	2



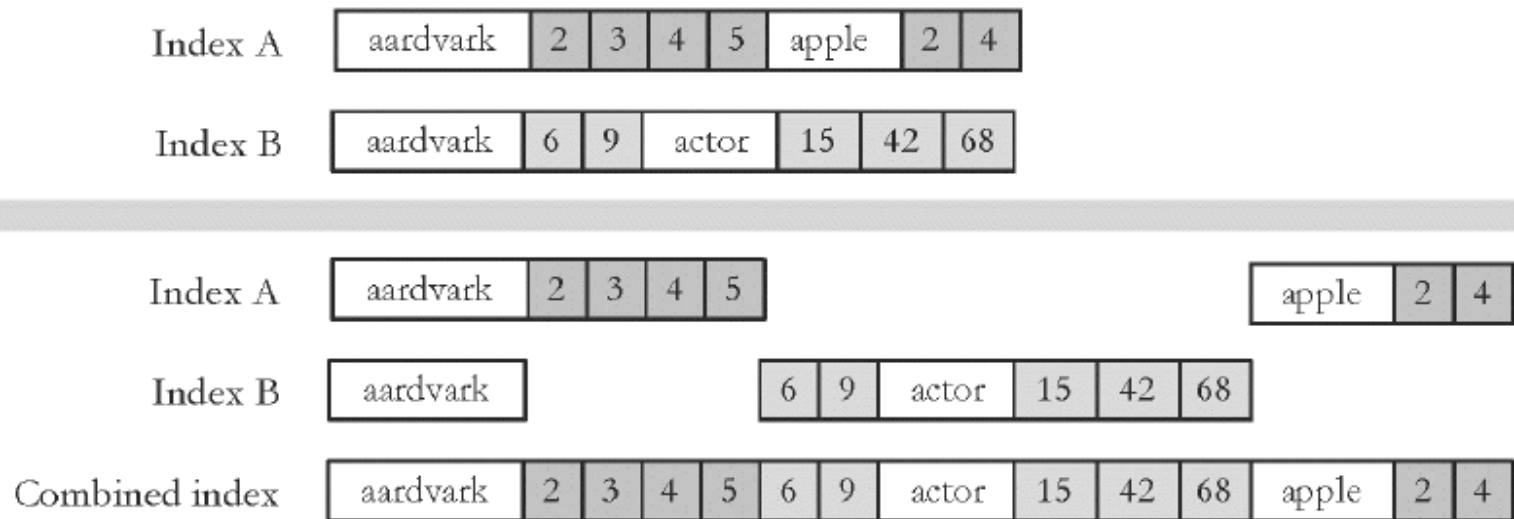
# Construção do Índice

term	docID		term	doc.	freq.	→	postings lists
ambitious	2		ambitious	1		→	2
be	2		be	1		→	2
brutus	1		brutus	2		→	1 → 2
brutus	2		capitol	1		→	1
capitol	1		caesar	2		→	1 → 2
caesar	1		did	1		→	1
caesar	2		enact	1		→	1
caesar	2		hath	1		→	2
did	1		i	1		→	1
enact	1		i'	1		→	1
hath	1		it	1		→	2
i	1	⇒	julius	1		→	1
i	1		killed	1		→	1
i'	1		let	1		→	2
it	2		me	1		→	1
julius	1		noble	1		→	2
killed	1		so	1		→	2
killed	1		the	2		→	1 → 2
let	2		told	1		→	2
me	1		you	1		→	2
noble	2		was	2		→	1 → 2
so	2		with	1		→	2
the	1						
the	2						
told	2						
you	2						
was	1						
was	2						
with	2						



# Merging de Índices

- Lida com memória limitada
  1. Constroi índice invertido até o limite da memória
  2. Escreve o índice parcial para o disco, e começa um novo
  3. No fim do processo, o disco possui vários índices parciais





# Merging de Resultados

- Merging de índice funciona bem em grandes batches (offline)
- Alternativa:
  - Docs novos vão para um índice auxiliar
  - Busca é feita em ambos
  - Resultados são unidos
  - Periodicamente índices são unidos
  - Desempenho da busca cai



# Compressão de Índices

- Motivação:
  - Usar menos disco
  - Colocar mais dados em memória
  - Ler dados compactados em memória mais rápido do que descompactado em disco
- Dois tipos:
  - Lossy: pré-processamento (ex.: stemming e stopword)
  - Lossless: nenhuma informação perdida





# Compressão de Dicionário

- Dicionário é pequeno comparado aos postings
- Motivação: colocá-lo em memória



# Compressão de Dicionário

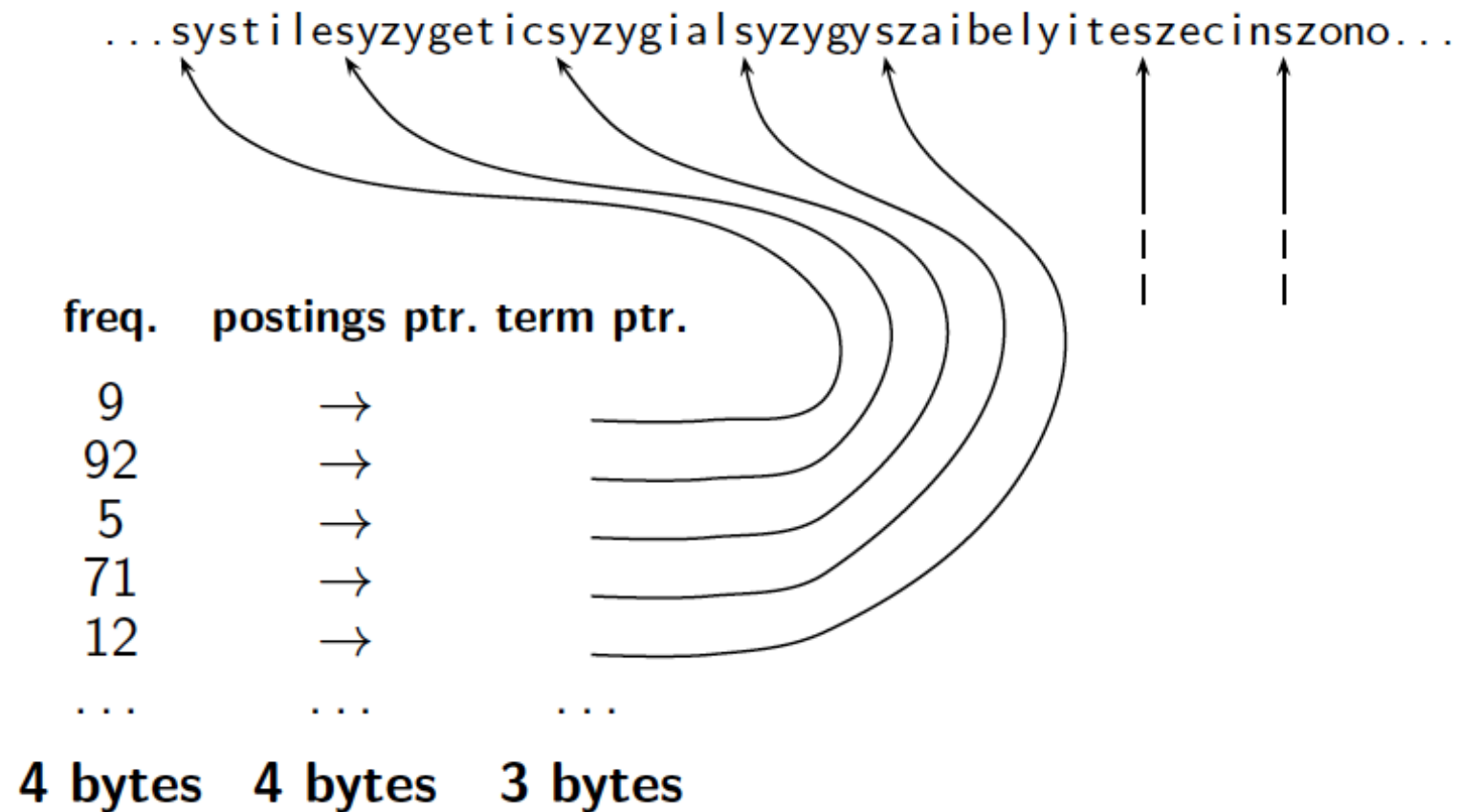
- Dicionário é pequeno comparado aos postings
- Motivação: colocá-lo em memória
- Abordagem simples: array com tamanho fixo de bytes

term	document frequency	pointer to postings list
a	656,265	→
aachen	65	→
...	...	...
zulu	221	→
20 bytes	4 bytes	4 bytes

- Para termos pequenos: desperdício de espaço
- Não consegue representar termos longos



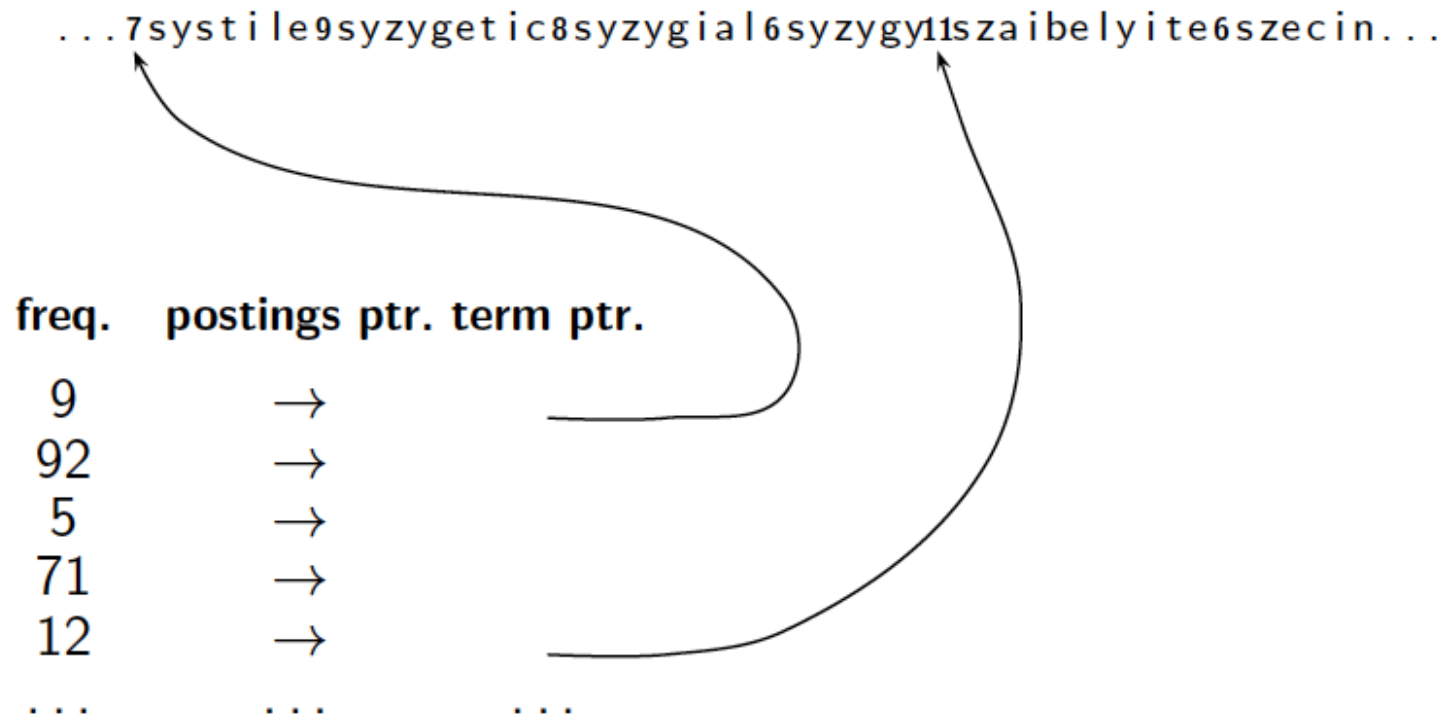
# Dicionário como String



- Ponteiro no termo mostra sua posição inicial e final do anterior
- Ao invés de armazenar o termo, armazena-se o ponteiro



# Dicionário como String com Blocos



- Adiciona 1 byte contendo o tamanho do termo no string do dicionário



# Compressão no Corpus Reuters

data structure	size in MB
dictionary, fixed-width	11.2
dictionary, term pointers into string	7.6
$\sim$ , with blocking, $k = 4$	7.1



# Compressão de Postings

- Postings são bem maiores que dicionários (pelo menos 10 vezes)
- Guardar intervalos ao invés de ids

COMPUTER: 283154, 283159, 283202,



COMPUTER: 283154, 5, 43



# Compressão de Postings

- Problema: intervalos para termos frequentes são pequenos e grandes para raros

	encoding	postings list				
THE	docIDs	...	283042	283043	283044	283045 ...
	gaps		1	1	1	...
COMPUTER	docIDs	...	283047	283154	283159	283202 ...
	gaps		107	5	43	...
ARACHNOCENTRIC	docIDs	252000	500100			
	gaps	252000	248100			



# Codificação de Tamanho Variável

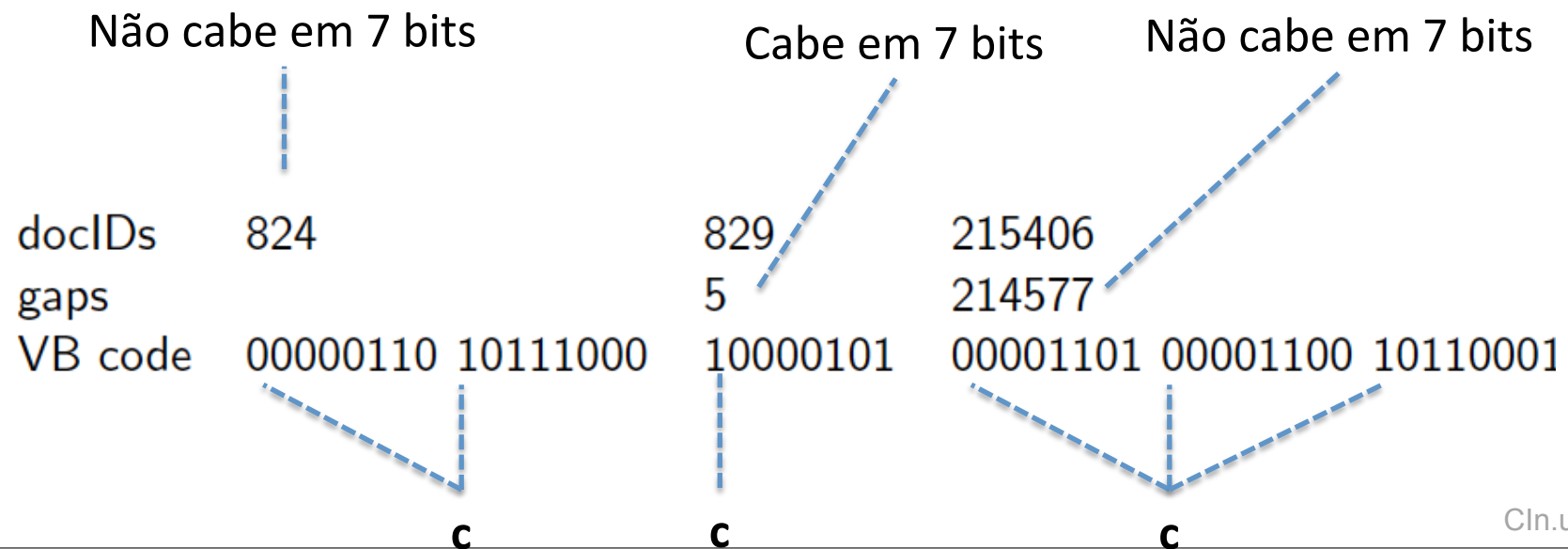
- Solução: poucos bits para termos frequentes e muitos para raros
- Bastante usado na prática






# Codificação de Tamanho Variável

- Se: o intervalo está dentro de 7 bits, coloque-o em binário nos 7 bits disponíveis e  $c = 1$
- 1 bit (mais alto) para ser o de continuação  $c$
- Senão: coloque o intervalo em binário e os 7 bits menores no primeiro byte, e o restante nos próximos,  $c=1$  no último  $c=0$  nos restantes





# Código Gamma

- Ex: 13 -> **1110101**  


Tamanho  
do offset  
em unário

Offset
- Concatenação do tamanho do offset em unário e offset
- Offset: número em binário sem o maior bit
  - Ex.: 13 -> 1101 -> 101
- Usa código unário: representa n com n1s e 0 no final
  - Ex.: 3 -> 1110
- 13 em inteiro: 32 bits
- 13 em unário: 7 bits
- 13 em código variável: 8 bits



# Código Gamma

number	unary code	length	offset	$\gamma$ code
0	0			
1	10	0		0
2	110	10	0	10,0
3	1110	10	1	10,1
4	11110	110	00	110,00
9	1111111110	1110	001	1110,001
13		1110	101	1110,101
24		11110	1000	11110,1000
511		111111110	11111111	111111110,11111111
1025		11111111110	0000000001	11111111110,0000000001



# Código Gamma

- Decodificação: 1110101...
  - Lê o unário até o 0
  - Determina o tamanho do offset
  - Adiciona o bit mais relevante retirado
- Código variável é mais simples e melhor para valores maiores



# Compressão da Coleção Reuters

data structure	size in MB
collection (text, xml markup etc)	3600.0
collection (text)	960.0
T/D incidence matrix	40,000.0
postings, uncompressed (32-bit words)	400.0
postings, variable byte encoded	116.0
postings, $\gamma$ encoded	101.0