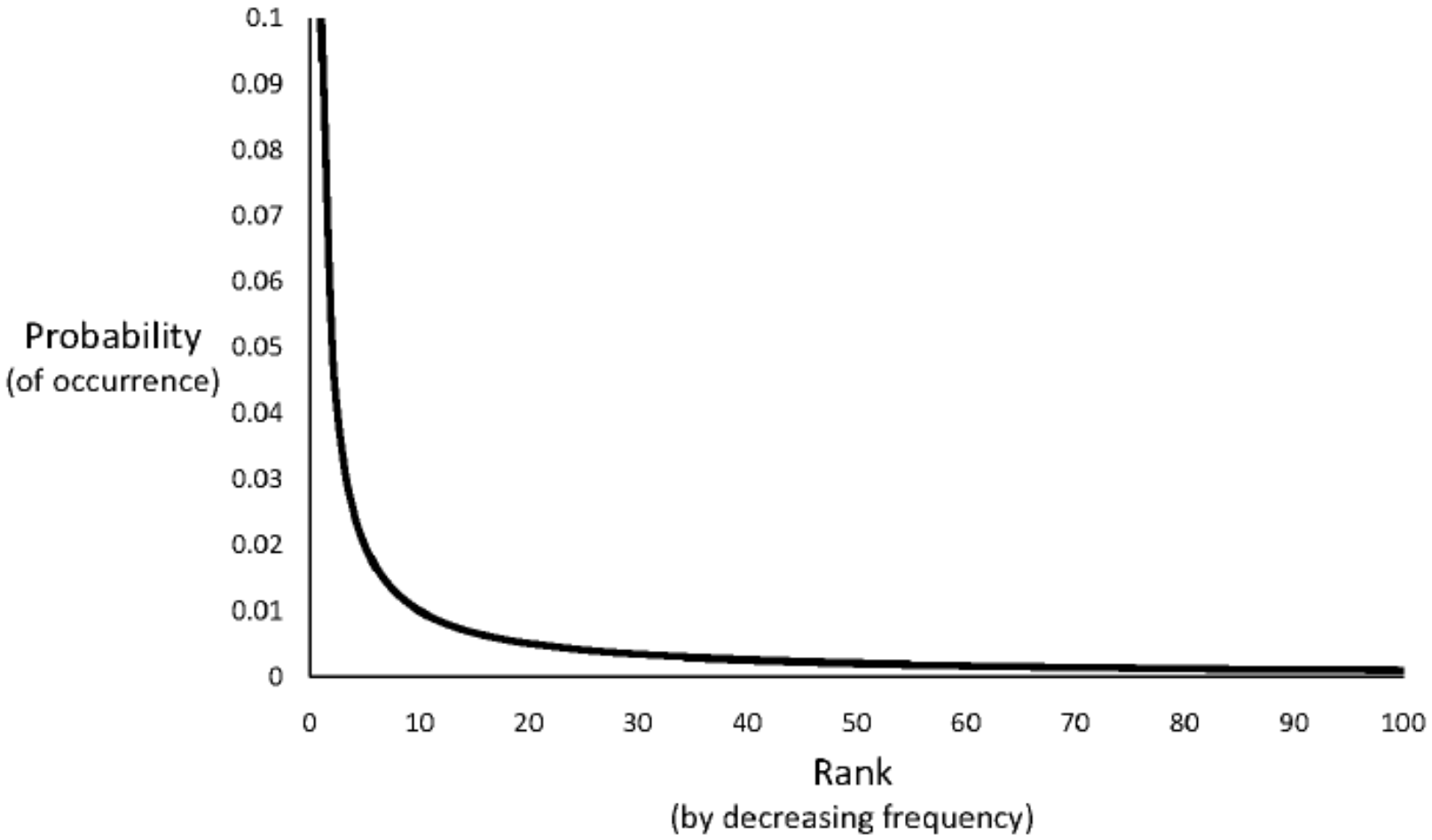
Recuperação de Informação: Processamento de Texto

Prof. Luciano Barbosa

(Parte do material retirado dos slides dos livros adotados)

UNIVERSIDADE
FEDERAL
DE PERNAMBUCO

Zipf's Law


0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
Probability
(of occurrence)
0
10
20
30
40
50
60
70
80
90
100
Rank
(by decreasing frequency)

Zipf's Law

- A distribuição da frequência das palavras em uma coleção é desigual
  - Poucas palavras aparecem muito e muitas pouco
  - Ex.: as duas palavras mais comuns em inglês (the, of) compõem 10% de todas ocorrências em textos
- Zipf's law:
  - Palavras ranqueadas em ordem decrescente de frequência
  - O ranking (r) de uma palavra vezes sua frequência (f) é aproximadamente uma constante (k): $r.f \approx k$ ou $r.P_r \approx c$

Coleção de Notícias: AP89

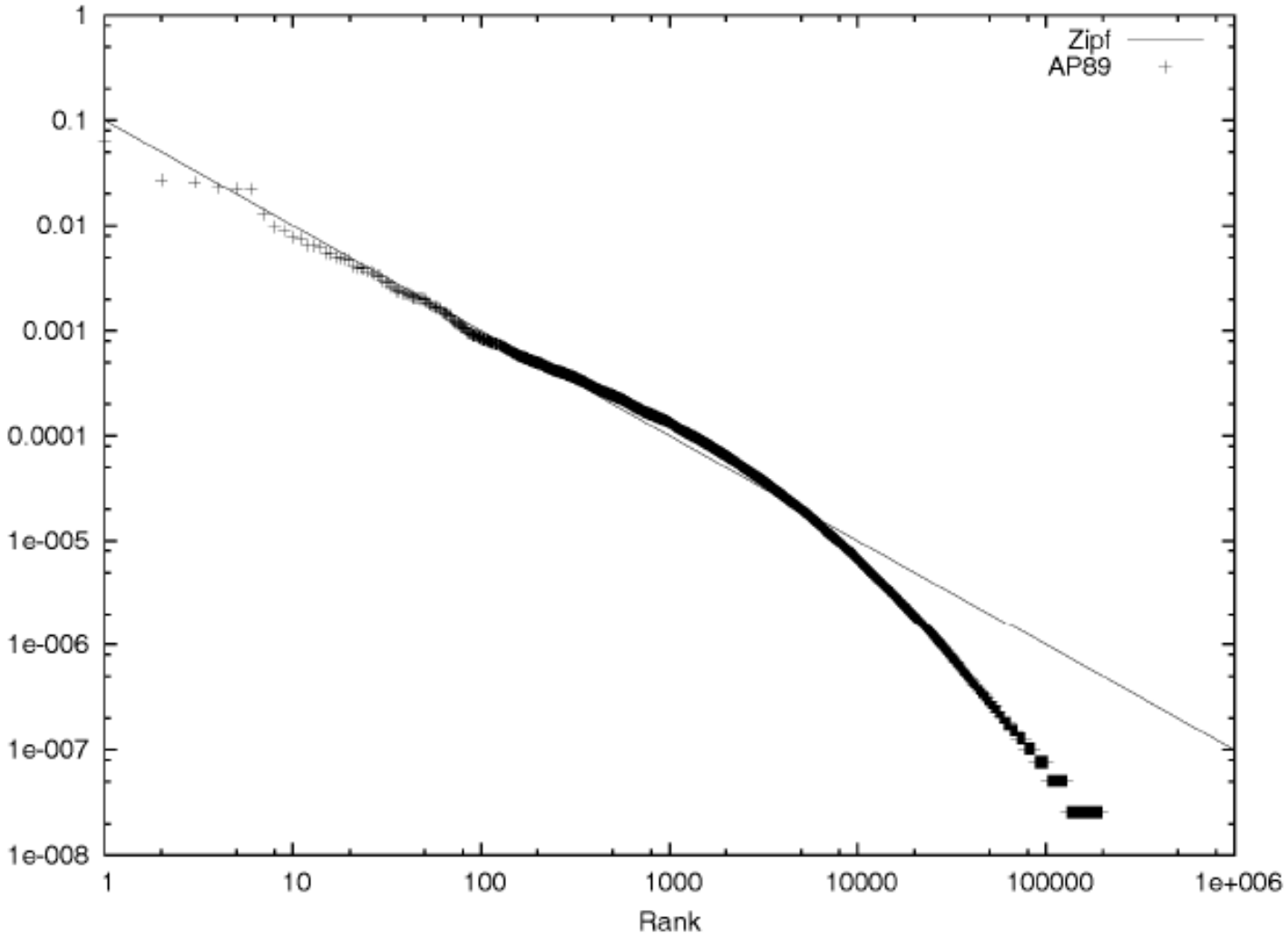
Total documents	84,678
Total word occurrences	39,749,179
Vocabulary size	198,763
Words occurring > 1000 times	4,169
Words occurring once	70,064

Word	*Freq.*	*r*	*Pr(%)*	*r.Pr*
assistant	5,095	1,021	.013	0.13
sewers	100	17,110	$2.56 \times 10{-}4$	0.04
toothbrush	10	51,555	$2.56 \times 10{-}5$	0.01
hazmat	1	166,945	$2.56 \times 10{-}6$	0.04

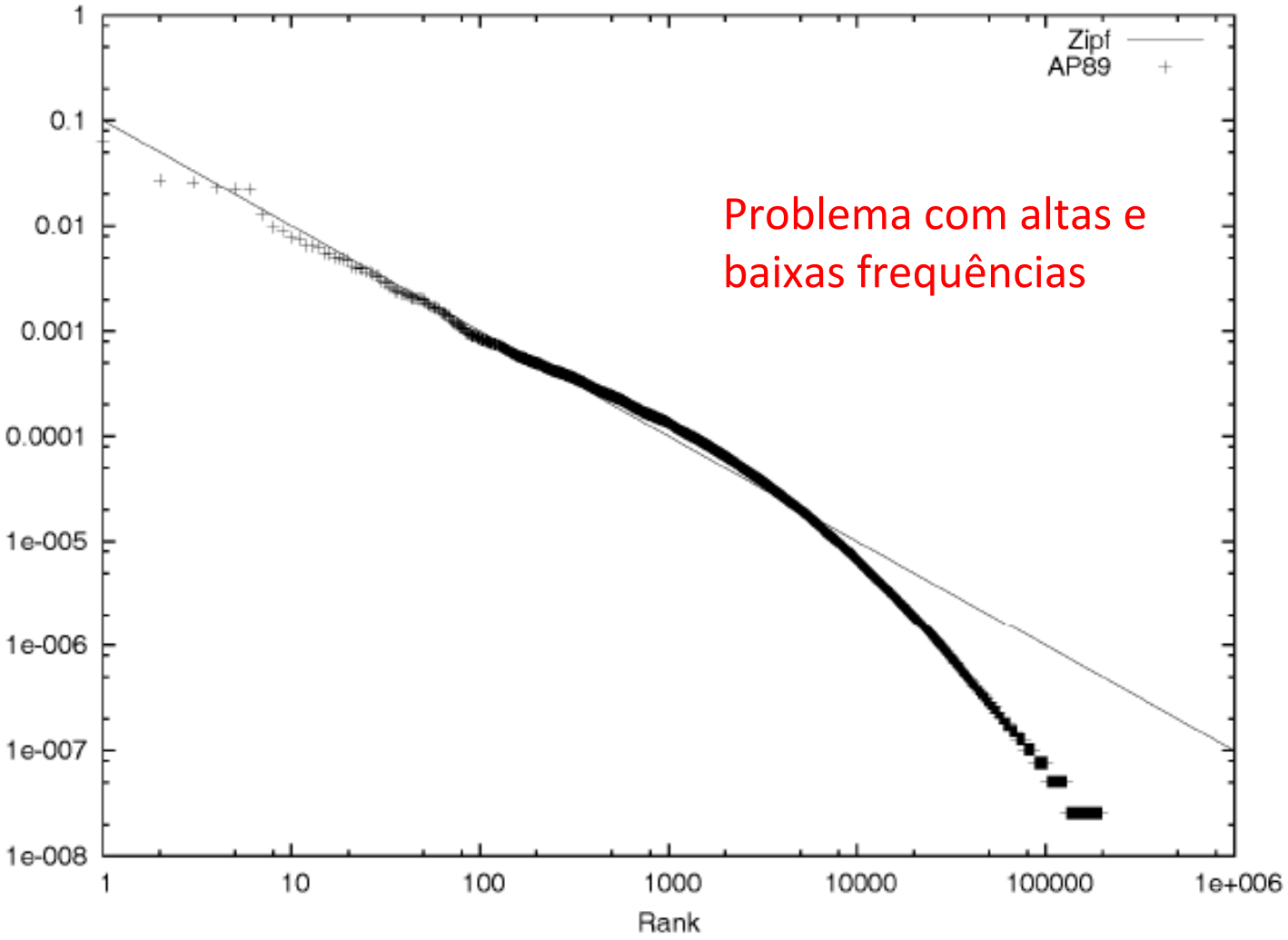
50 Palavras mais Frequentes

Word	*Freq.*	*r*	P_r(%)	$r.P_r$	*Word*	*Freq*	*r*	P_r(%)	$r.P_r$
the	2,420,778	1	6.49	0.065	has	136,007	26	0.37	0.095
of	1,045,733	2	2.80	0.056	are	130,322	27	0.35	0.094
to	968,882	3	2.60	0.078	not	127,493	28	0.34	0.096
a	892,429	4	2.39	0.096	who	116,364	29	0.31	0.090
and	865,644	5	2.32	0.120	they	111,024	30	0.30	0.089
in	847,825	6	2.27	0.140	its	111,021	31	0.30	0.092
said	504,593	7	1.35	0.095	had	103,943	32	0.28	0.089
for	363,865	8	0.98	0.078	will	102,949	33	0.28	0.091
that	347,072	9	0.93	0.084	would	99,503	34	0.27	0.091
was	293,027	10	0.79	0.079	about	92,983	35	0.25	0.087
on	291,947	11	0.78	0.086	i	92,005	36	0.25	0.089
he	250,919	12	0.67	0.081	been	88,786	37	0.24	0.088
is	245,843	13	0.65	0.086	this	87,286	38	0.23	0.089
with	223,846	14	0.60	0.084	their	84,638	39	0.23	0.089
at	210,064	15	0.56	0.085	new	83,449	40	0.22	0.090
by	209,586	16	0.56	0.090	or	81,796	41	0.22	0.090
it	195,621	17	0.52	0.089	which	80,385	42	0.22	0.091
from	189,451	18	0.51	0.091	we	80,245	43	0.22	0.093
as	181,714	19	0.49	0.093	more	76,388	44	0.21	0.090
be	157,300	20	0.42	0.084	after	75,165	45	0.20	0.091
were	153,913	21	0.41	0.087	us	72,045	46	0.19	0.089
an	152,576	22	0.41	0.090	percent	71,956	47	0.19	0.091
have	149,749	23	0.40	0.092	up	71,082	48	0.19	0.092
his	142,285	24	0.38	0.092	one	70,266	49	0.19	0.092
but	140,880	25	0.38	0.094	people	68,988	50	0.19	0.093

Zipf's Law para AP89



Zipf's Law para AP89



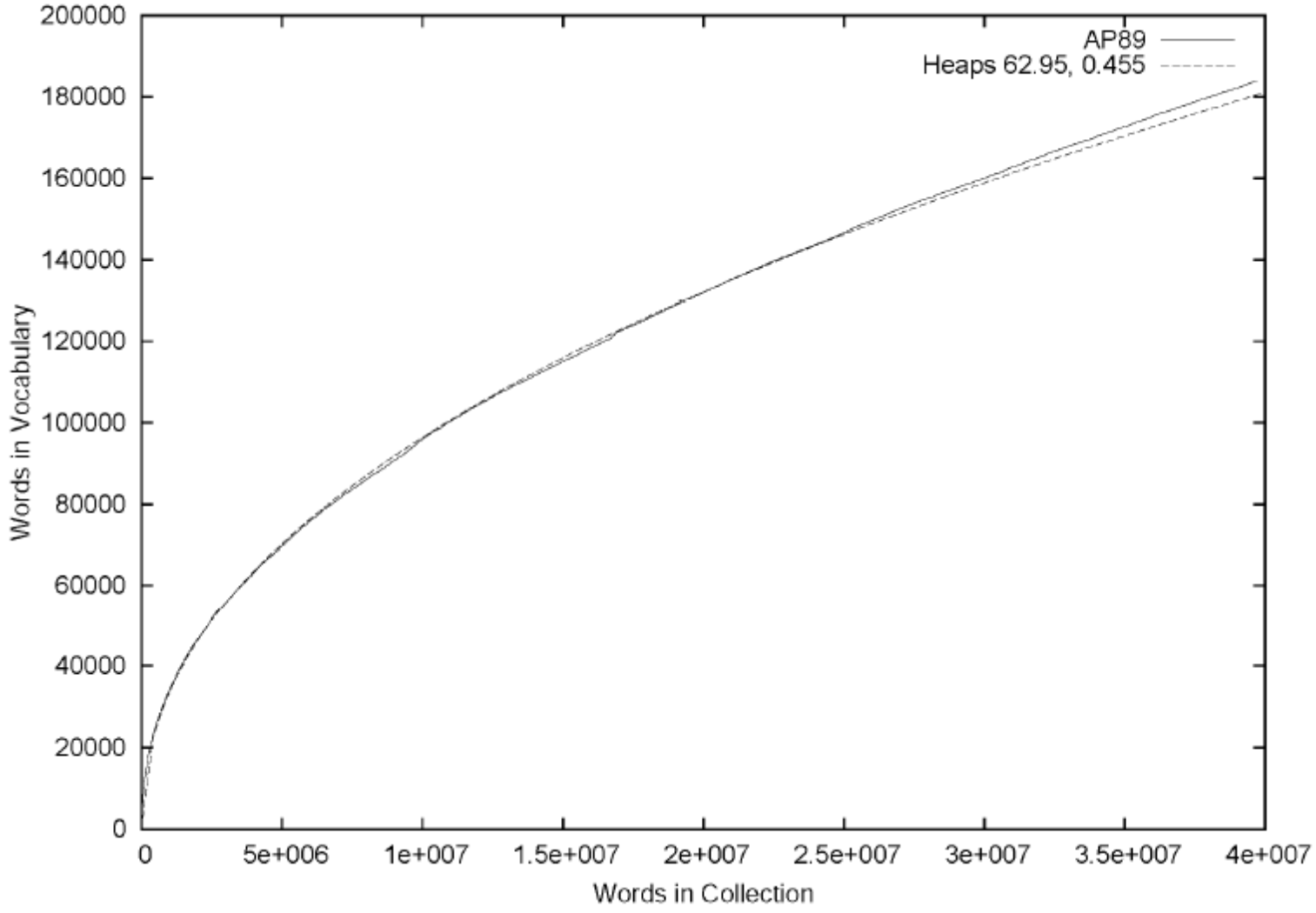
Heap's Law

- Estima o número de termos distintos em uma coleção

$$M = kT^b$$

  - M: tamanho do vocabulário (palavras únicas)
  - T: número de palavras no corpus
  - k,b: parâmetros que variam por corpus (valores típicos $30 \leq k \leq 100$ e $b \approx 0.5$)
- Útil por exemplo para o processo de indexação: ex., alocação de memória

Exemplo: AP89



Predições para Coleções da TREC

- Funciona bem para um número grande de palavras
  - Ex.: 10.879.522 palavras da AP89
    - Predição: 100.151 palavras únicas
    - Número real: 100.024
- Predições para números pequenos (<1000) de palavras são muito piores

Heap's Law

- Funciona bem com grandes corpora
- Parâmetros dependem do corpus
- Novas palavras aparecem mesmo depois de 30 milhões
  - Razão: erros de ortografia, palavras inventadas, código, outras línguas
- Conclusões:
  - O tamanho do vocabulário sempre cresce
  - Para grandes coleções de documentos o vocabulário será muito grande

Estimando o Tamanho do Resultado

- Quantas páginas contêm todos os termos da consulta?
- Para a consulta a ^ b ^ c

- Baseado em contadores individuais
  - Suposição: termos ocorrem independemente
  - f_a, f_b, f_c: número de documentos onde a,b e c ocorrem
  - N: número de documentos na coleção

$$P_{abc} = P_a * P_b * P_c$$

$$f_{abc} = N \cdot f_a/N \cdot f_b/N \cdot f_c/N = (f_a \cdot f_b \cdot f_c)/N^2$$

Exemplo GOV2

Word(s)	*Document Frequency*	*Estimated Frequency*
tropical	120,990	
fish	1,131,855	
aquarium	26,480	
breeding	81,885	
tropical fish	18,472	5,433
tropical aquarium	1,921	127
tropical breeding	5,510	393
fish aquarium	9,722	1,189
fish breeding	36,427	3,677
aquarium breeding	1,848	86
tropical fish aquarium	1,529	6
tropical fish breeding	3,629	18

Collection size (N) is 25,205,179

Estimando o Tamanho do Resultado

- Ruim porque os termos não são independentes
- Melhores estimativas se co-ocorrência estiver disponível

Estimando o Tamanho do Resultado

- Baseado na amostra de documentos processados na consulta
- C/s
  - C: número de docs com todas as palavras
  - s: proporção de documentos processados
- Ex.: “tropical fish aquarium” in GOV2
  - Depois de processar 3000 de 26480 que contêm “aquarium” (termo menos frequente), 258 contêm todos os termos

$$f_{tropical \cap fish \cap aquarium} = 258/(3000 \div 26480) = 2,277$$

- Após processar 20%

$$f_{tropical \cap fish \cap aquarium} = 1,778 \quad (1,529 \text{ is real value})$$

Aplicação: Estimando Associação entre Palavras

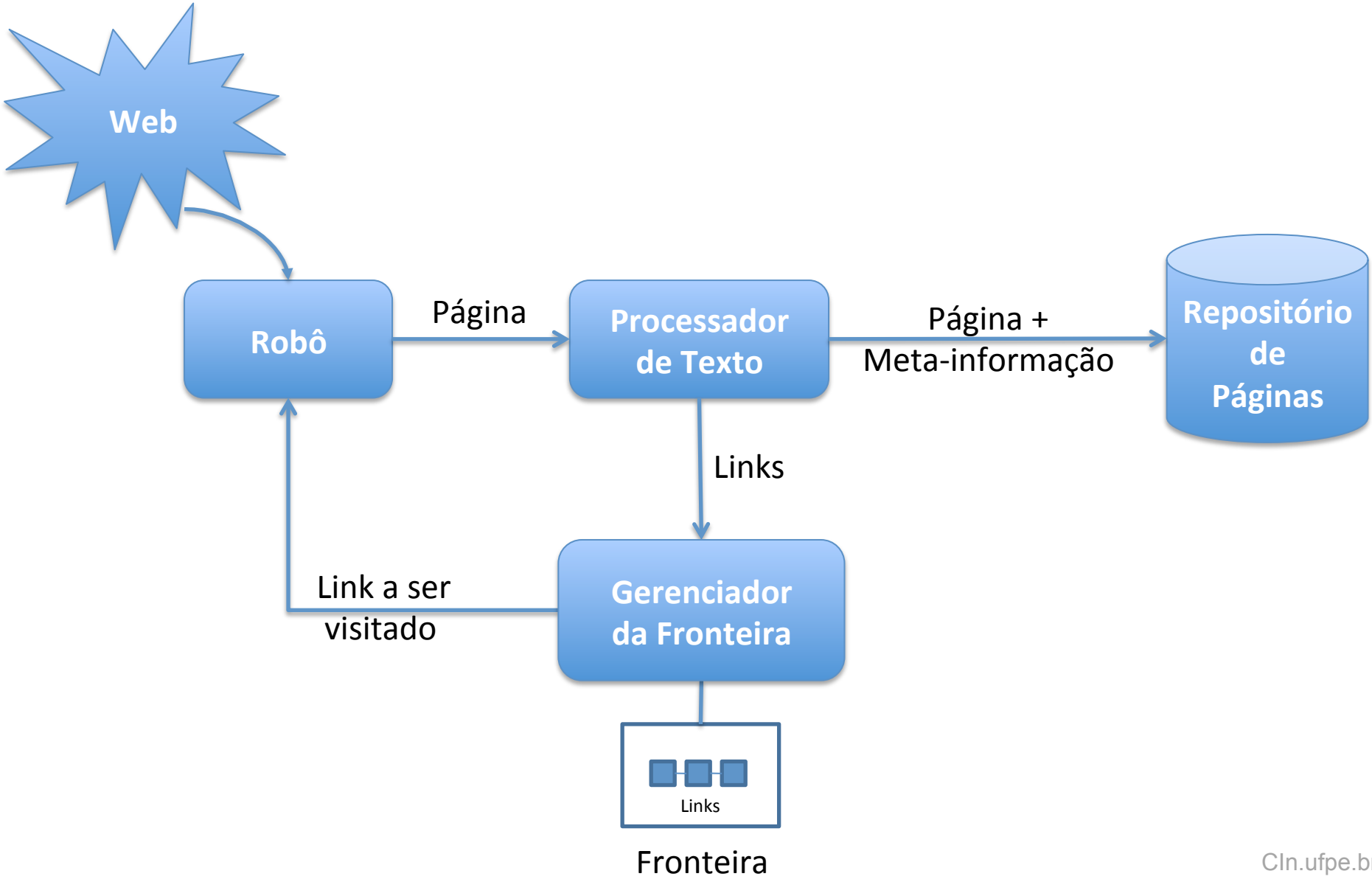
- PMI-IR (pointwise mutual information - IR) =

$$\frac{|HITS(Label_r, Label_c))|}{|HITS(Label_c)|}$$

Company's name	Anchor	PMI-IR
apple	iphone	0.069
apple	environment	0.00014
motorola	motoblur	0.04
motorola	news	0.00005

Chamada

Funcionamento de um Coletor Geral



Processamento de Texto em Engenhos de Busca

- Objetivo: converter documentos para termos no índice
- Motivação
  - Casar exatamente os strings digitados pelos usuários é muito restritivo
  - Nem todas as palavras têm a mesma importância na busca
  - Algumas vezes não é claro quando uma palavra começa ou termina (ex.: chinês, coreano)

Diferentes Formatos

- MIME: Multipurpose Internet Mail Extensions
  - Padrão para identificar arquivos na internet
  - Estrutura: <tipo>/<subtipo>
  - Ex: text/plain, image/jpeg, audio/mp3, video/mp4 e application/msword
  - Lista: http://www.freeformatter.com/mime-types-list.html#mime-types-list

Formatos de Documentos

- HTTP header: campo Content-Type
  - Default: text/plain
  - Em Java: método getContentType da classe java.net.URLConnection

java.net

Class URLConnection

java.lang.Object
    java.net.URLConnection

Direct Known Subclasses:

HttpURLConnection, JarURLConnection

```
public abstract class URLConnection
extends Object
```

The abstract class `URLConnection` is the superclass of all classes that represent a communications link between the application and a URL. Instances of this class can be used both to read from and to write to the resource referenced by the URL. In general, creating a connection to a URL is a multistep process:

getContentType

```
public String getContentType()
```

Returns the value of the `content-type` header field.

Returns:

the content type of the resource that the URL references, or `null` if not known.

Diferentes Charsets

- Conjunto de caracteres usados em uma língua
  - Ex: Caracteres latinos usados pela maioria das línguas europeias
- ASCII original: 128 caracteres usando 7 bits

Dec	Hx	Oct	Char	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
0	0	000	NUL (null)	32	20	040		Space	64	40	100	@	@	96	60	140	`	`
1	1	001	SOH (start of heading)	33	21	041	!	!	65	41	101	A	A	97	61	141	a	a
2	2	002	STX (start of text)	34	22	042	"	"	66	42	102	B	B	98	62	142	b	b
3	3	003	ETX (end of text)	35	23	043	#	#	67	43	103	C	C	99	63	143	c	c
4	4	004	EOT (end of transmission)	36	24	044	$	$	68	44	104	D	D	100	64	144	d	d
5	5	005	ENQ (enquiry)	37	25	045	%	%	69	45	105	E	E	101	65	145	e	e
6	6	006	ACK (acknowledge)	38	26	046	&	&	70	46	106	F	F	102	66	146	f	f
7	7	007	BEL (bell)	39	27	047	'	'	71	47	107	G	G	103	67	147	g	g
8	8	010	BS (backspace)	40	28	050	(	(	72	48	110	H	H	104	68	150	h	h
9	9	011	TAB (horizontal tab)	41	29	051	)	)	73	49	111	I	I	105	69	151	i	i
10	A	012	LF (NL line feed, new line)	42	2A	052	*	*	74	4A	112	J	J	106	6A	152	j	j
11	B	013	VT (vertical tab)	43	2B	053	+	+	75	4B	113	K	K	107	6B	153	k	k
12	C	014	FF (NP form feed, new page)	44	2C	054	,	,	76	4C	114	L	L	108	6C	154	l	l
13	D	015	CR (carriage return)	45	2D	055	-	-	77	4D	115	M	M	109	6D	155	m	m
14	E	016	SO (shift out)	46	2E	056	.	.	78	4E	116	N	N	110	6E	156	n	n
15	F	017	SI (shift in)	47	2F	057	/	/	79	4F	117	O	O	111	6F	157	o	o
16	10	020	DLE (data link escape)	48	30	060	0	0	80	50	120	P	P	112	70	160	p	p
17	11	021	DC1 (device control 1)	49	31	061	1	1	81	51	121	Q	Q	113	71	161	q	q
18	12	022	DC2 (device control 2)	50	32	062	2	2	82	52	122	R	R	114	72	162	r	r
19	13	023	DC3 (device control 3)	51	33	063	3	3	83	53	123	S	S	115	73	163	s	s
20	14	024	DC4 (device control 4)	52	34	064	4	4	84	54	124	T	T	116	74	164	t	t
21	15	025	NAK (negative acknowledge)	53	35	065	5	5	85	55	125	U	U	117	75	165	u	u
22	16	026	SYN (synchronous idle)	54	36	066	6	6	86	56	126	V	V	118	76	166	v	v
23	17	027	ETB (end of trans. block)	55	37	067	7	7	87	57	127	W	W	119	77	167	w	w
24	18	030	CAN (cancel)	56	38	070	8	8	88	58	130	X	X	120	78	170	x	x
25	19	031	EM (end of medium)	57	39	071	9	9	89	59	131	Y	Y	121	79	171	y	y
26	1A	032	SUB (substitute)	58	3A	072	:	:	90	5A	132	Z	Z	122	7A	172	z	z
27	1B	033	ESC (escape)	59	3B	073	;	;	91	5B	133	[	[	123	7B	173	{	{
28	1C	034	FS (file separator)	60	3C	074	<	<	92	5C	134	\	\	124	7C	174			|
29	1D	035	GS (group separator)	61	3D	075	=	=	93	5D	135	]	]	125	7D	175	}	}
30	1E	036	RS (record separator)	62	3E	076	>	>	94	5E	136	^	^	126	7E	176	~	~
31	1F	037	US (unit separator)	63	3F	077	?	?	95	5F	137	_	_	127	7F	177	&#127;	DEL

Source: www.LookupTables.com

ASCII Estendido

- Acomodar caracteres extras (8 bits)
  - Ex: caracteres com acento (á,ü etc)
- Mapeamento depende de região, SO etc

Code	Win	Mac	Code	Win	Mac	Code	Win	Mac	Code	Win	Mac
128	€	Ä	160		†	192	À	¿	224	à	‡
129	￮	Å	161	¡	°	193	Á	¡	225	á	·
130	‚	Ç	162	¢	¢	194	Â	¬	226	â	‚
131	ƒ	É	163	£	£	195	Ã	√	227	ã	„
132	„	Ñ	164	¤	§	196	Ä	ƒ	228	ä	‰
133	…	Ö	165	¥	•	197	Å	≈	229	å	Â
134	†	Ü	166	¦	¶	198	Æ	Δ	230	æ	Ê
135	‡	á	167	§	ß	199	Ç	«	231	ç	Á
136	ˆ	à	168	¨	®	200	È	»	232	è	Ë
137	‰	â	169	©	©	201	É	…	233	é	È
138	Š	ä	170	ª	™	202	Ê		234	ê	Í
139	‹	ã	171	«	´	203	Ë	À	235	ë	Î
140	Œ	å	172	¬	¨	204	Ì	Ã	236	ì	Ï
141	￮	ç	173	-	≠	205	Í	Õ	237	í	Ì
142	Ž	é	174	®	Æ	206	Î	Œ	238	î	Ó
143	￮	è	175	¯	Ø	207	Ï	œ	239	ï	Ô
144	￮	ê	176	°	∞	208	Ð	–	240	ð	□
145	‘	ë	177	±	±	209	Ñ	—	241	ñ	Ò
146	’	í	178	²	≤	210	Ò	“	242	ò	Ú
147	“	ì	179	³	≥	211	Ó	”	243	ó	Û
148	”	î	180	´	¥	212	Ô	‘	244	ô	Ù
149	•	ï	181	µ	µ	213	Õ	’	245	õ	ı
150	–	ñ	182	¶	∂	214	Ö	÷	246	ö	ˆ
151	—	ó	183	·	∑	215	×	◊	247	÷	˜
152	˜	ò	184	¸	∏	216	Ø	ÿ	248	ø	¯
153	™	ô	185	¹	π	217	Ù	Ÿ	249	ù	˘
154	š	ö	186	º	∫	218	Ú	⁄	250	ú	˙
155	›	õ	187	»	ª	219	Û	€	251	û	°
156	œ	ú	188	¼	º	220	Ü	‹	252	ü	¸
157	￮	ù	189	½	Ω	221	Ý	›	253	ý	˝
158	ž	û	190	¾	æ	222	Þ	fi	254	þ	˛
159	Ÿ	ü	191	¿	ø	223	ß	fl	255	ÿ	ˇ

Windows-1252
vs.
Mac Roman Chart

Unicode

- Tenta padronizar e incluir todos os caracteres em todos sistemas de escrita
  - 0-127: ASCII
  - 128-255: Latin-1
  - Resto organizado em blocos de “scripts”
  - http://www.unicode.org/charts/
  - Mais de 1 milhão de caracteres (code points)
- Arquivos de texto podem ser compartilhados sem tradução de encoding
- Mesma página pode conter caracteres de diferentes línguas
  - http://www.trigeminal.com/samples/provincial.html

UTF-8

- Método para codificar entradas no Unicode
- Codifica um caracter em sequências de bytes
- Usado por cerca de 87% das páginas Web
- Codifica mais de 1 milhão de code points

UTF-8

- 0-128 igual ao ASCII (1 byte)
- >128 usa mais de 1 byte
- Começa com sequência de bits especiais
  - 0…: single byte
  - 10…: continuation bytes
  - 11…: leading bytes
- Exemplo: “à”
  - Code point: U+00E1, decimal: 255, binário: 11100001
  - UTF-8: 11000011 10100001 (C3 A1 em hexadecimal)
- Mozila charsetdetector
  - http://www-archive.mozilla.org/projects/intl/detectorsrc.html

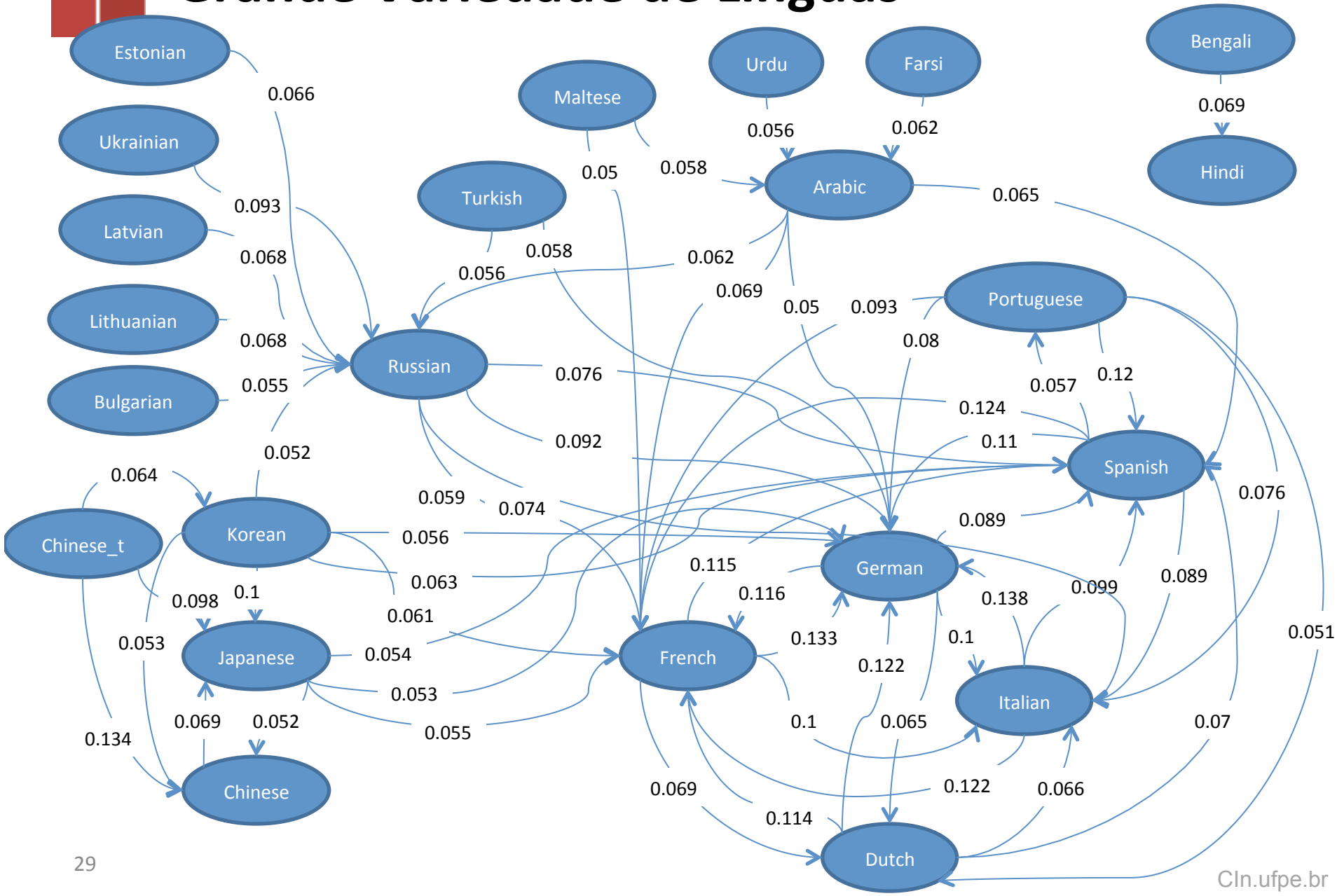
Diferentes Línguas

- 2/3 da Web é em inglês
- Cerca de 50% dos usuários não usam inglês como primeira língua
- A maioria das aplicações de busca na Web lidam com várias línguas
  - Busca monolingual
  - Busca entre línguas

Internationalization

- Muitos aspectos dos engenhos de busca são indiferentes a línguas
- Diferenças:
  - Text encoding
  - Tokenização
  - Stemming
- Diferenças culturais também podem impactar interface
- API em Java
  - https://github.com/shuyo/language-detection

Grande Variedade de Línguas



Tokenização

- Quebrar sequência de caracteres em palavras
- Exemplo de aborgadem simples
  - Qualquer sequência de caracteres alfanuméricos de tamanho mínimo 3
  - Terminado em espaço ou algum caracter especial
  - Letras maiúsculas convertidas para minúsculas

Tokenização

- Quebrar sequência de caracteres em palavras
- Exemplo de aborgadem simples
  - Qualquer sequência de caracteres alfanuméricos de tamanho mínimo 3
  - Terminado em espaço ou algum caracter especial
  - Letras maiúsculas convertidas para minúsculas
  - Ex 1: “Bigcorp’s 2007 bi-annual report showed profits rose 10%” -> “bigcorp 2007 annual report showed profits rose”
  - Ex 2: “Mr. O’Neill thinks that the boys’ stories about Chile’s capital aren’t amusing” - > “mr neill thinks that the boys stories about chile capital aren amusing”

Problemas em Tokenização

- Palavras pequenas podem ser importantes em algumas consultas
  - Ex: am, pm, el (paso), (world war) II
- Hífens
  - Algumas vezes são necessários
    - Ex: e-bay, wal-mart, cd-rom, t-shirts
  - Separam palavras
    - Ex: Dallas-Fort Worth, spanish-speaking

Problemas em Tokenização

- Caracteres especiais são importantes para URL, tags e código em documentos
- Palavras com letras maiúsculas podem ter significados diferentes
  - Bush, Apple
- Apóstrofo pode ser parte de uma palavra, parte de um possessivo (inglês), ou erro

rosie o'donnell, can't, don't, 80's, 1890's, men's straw hats, master's degree, england's ten largest cities, shriner's

Problemas em Tokenização

- Números podem ser importantes
  - Ex: nokia 3250, united 93, quicktime 6.5 pro
- Pontos podem estar em números, abreviações, URLs, fim de sentenças etc
  - Ex: I.B.M., Ph.D.

Tokenização em Outras Línguas

- Chinês

莎拉波娃现在居住在美国东南部的佛罗里达。今年４月９日，莎拉波娃在美国第一大城市纽约度过了１８岁生日。生日派对上，莎拉波娃露出了甜美的微笑。

Tokenização: Ambiguidade

The two characters can be treated as one word meaning ‘monk’ or as a sequence of two words meaning ‘and’ and ‘still’.

Tokenização Sem Espaço em Branco

Compounds in Dutch, German, Swedish

Computerlinguistik → Computer + Linguistik

Lebensversicherungsgesellschaftsangestellter

→ leben + versicherung + gesellschaft + angestellter

Life insurance company employee

Tokenização Sem Espaço em Branco: Japonês

ノーベル平和賞を受賞したワンガリ・マータイさんが名誉会長を務めるＭＯＴＴＡＩＮＡＩキャンペーンの一環として、毎日新聞社とマガジンハウスは「私の、もったいない」を募集します。皆様が日ごろ「もったいない」と感じて実践していることや、それにまつわるエピソードを８００字以内の文章にまとめ、簡単な写真、イラスト、図などを添えて１０月２０日までにお送りください。大賞受賞者には、５０万円相当の旅行券とエコ製品２点の副賞が贈られます。

Processo de Tokenização

- Simples
  - Palavra é qualquer sequência de caracters alfa-numéricos terminado por espaço ou caracter especial, tudo convertido para minúsculo
- Stanford tokenizer para inglês
  - http://nlp.stanford.edu/software/tokenizer.shtml

Stopwords

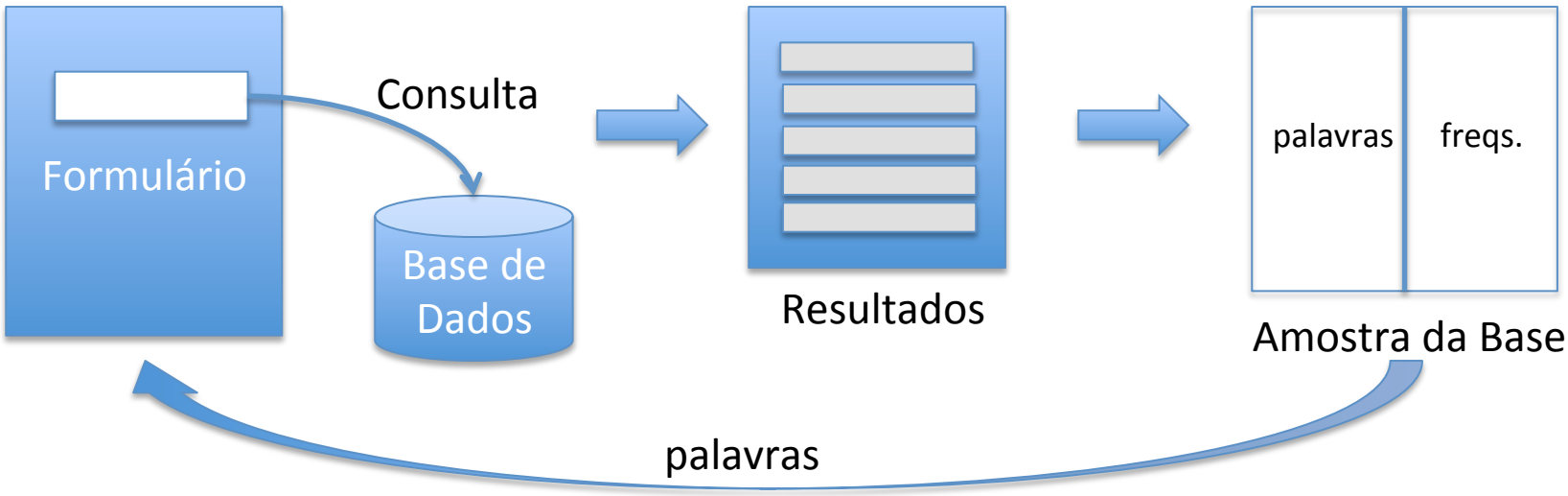
- Palavras que aparecem muito na coleção
- Não possuem muito significado
- Usualmente, não são boas para diferenciar
- Artigos, preposições, conjunções etc
- Remoção usada em sistemas de RI antigos
- Eliminação
  - Reduz o tamanho do índice em 40% ou mais
  - Problema para consultas por frase exata: “no limite”

Stopwords

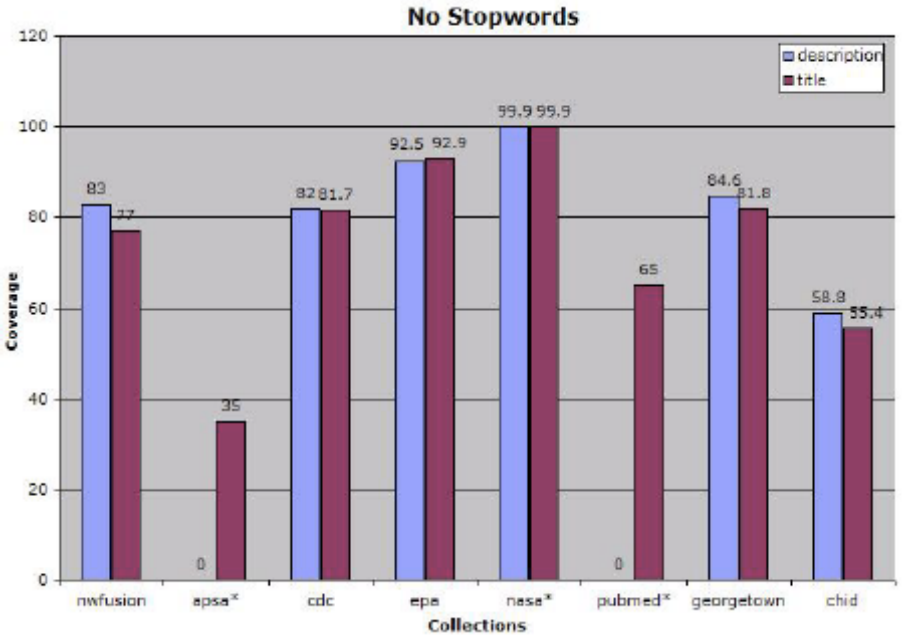
- Criadas a partir de palavras com alta frequência ou de listas existentes
  - Ex: https://gist.github.com/alopes/5358189
- Podem ser customizadas por aplicação e domínio
  - “Click” é uma stopword para âncoras
- Melhor política: indexar todos os documentos e tomar decisões em tempo de consulta
- A maioria dos engenhos de busca não as remove

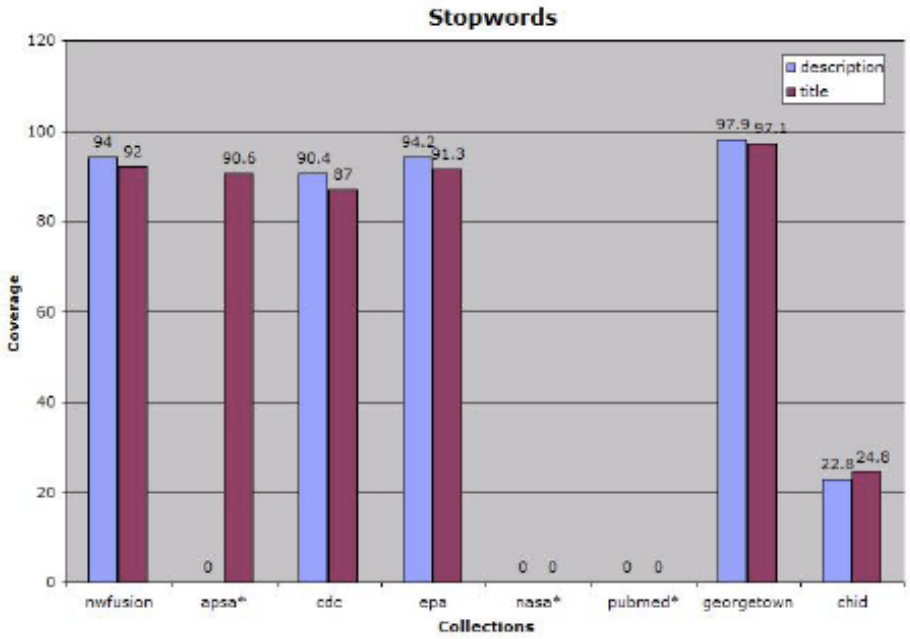
Stopwords para Hidden-Web Crawling

- Siphon



Stopwords para Hidden-Web Crawling





Stemming

- Reduz variações morfológicas das palavras para um stem em comum
- Remove prefixo ou sufixo -> stem
  - Ex: connect – connected, connecting, connection, connections
- Reduz o tamanho do índice
- Não há um consenso sobre benefícios (depende da língua)
- Aumenta a cobertura e pode piorar a precisão
- Muitos engenhos de busca não usam

Porter Stemmer

- Usado em experimentos de RI desde 1970
- Consiste de uma série de regras
- Comete erros e difícil de modificar

Original text:
Document will describe marketing strategies carried out by U.S. companies for their agricultural chemicals, report predictions for market share of such chemicals, or report market statistics for agrochemicals, pesticide, herbicide, fungicide, insecticide, fertilizer, predicted sales, market share, stimulate demand, price cut, volume of sales.

Porter stemmer:
document describ market strategi carri compani agricultur chemic report predict market share chemic report market statist agrochem pesticid herbicid fungicid insecticid fertil predict sale market share stimul demand price cut volum sale

- https://github.com/stanfordnlp/CoreNLP

Remoção de Stopwords e Stemming na Reuters-RCV1

	number
unfiltered	484,494
no numbers	473,723
case folding	391,523
30 stop words	391,493
150 stop words	391,373
stemming	322,383