Apêndice H

Apresentação detalhada das colecções do Segundo HAREM

Cristina Mota, Diana Santos, Paula Carvalho, Cláudia Freitas e Hugo Gonçalo Oliveira

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Apêndice H, p. 355–377.

Nota das editoras: Este apêndice foi criado especialmente para o presente livro, de modo a complementar as informações sobre as colecções do Segundo HAREM fornecidas nos capítulos 1, 3 e 4.

Neste apêndice apresentamos em mais pormenor a constituição das várias colecções usadas no Segundo HAREM, segundo as seguintes vertentes: variante do português, tipo de texto e fonte do texto.

Para os três casos, apresentamos, na forma gráfica e por tabela, tanto a distribuição dos documentos, como a distribuição das palavras..

Começamos por detalhar a distribuição dos documentos e de palavras entre português do Brasil e de Portugal, nas figuras H.1 e H.2, respectivamente.

Convém no entanto recordar a forma de constituição da colecção do HAREM para explicar a razão da tripartição das figuras ao longo deste apêndice, nomeadamente distinguindo entre colecção HAREM total, colecção HAREM sem documentos da colecção CHAVE e colecção dourada. É que a colecção do HAREM foi construída como a soma da colecção dourada + os exemplos já apresentados e disponibilizados ao público + a CD do Primeiro HAREM + documentos obtidos da colecção CHAVE.

As duas parcelas intermédias foram incluídas para permitir mais tarde comparações entre o desempenho dos sistemas neste e no anterior HAREM, ou para comparar o desempenho entre material de treino e material novo.

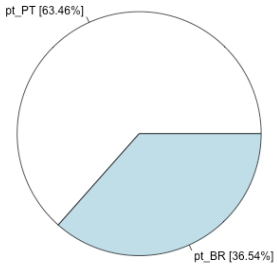
Mas do ponto de vista de distribuição, a colecção CHAVE é homogeneamente composta por documentos jornalísticos e faz sentido retirá-la por exemplo quando se está interessado nos diferentes géneros.

Depois apresentamos a distribuição de documentos e de palavras por tipo de texto, nas figuras H.3 e H.4, respectivamente. Notamos que os documentos podiam ser classificados com mais do que um tipo de texto. Assim, no caso dos documentos com mais de um tipo de texto, esse documento contribuiu com peso $1/n$, sendo n o número de classificações diferentes desse documento.

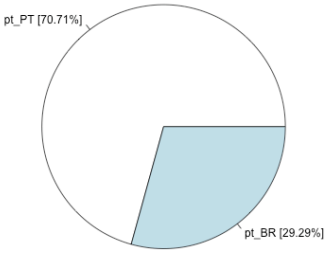
Salientamos que, relativamente à classificação por tipo de texto, não há um conjunto de géneros (ou tipos de texto) consensuais ou sequer amplamente utilizados em português (e mesmo em qualquer outra língua). Por isso de cada vez que é preciso caracterizar uma coleção de textos heterogénea deparamo-nos com problemas e com discordâncias sobre quer a grelha quer a granularidade. Isto pode aliás ver-se nas três avaliações conjuntas organizadas pela Linguateca, que usaram três bitolas diferentes (a das Morfolimpíadas, descrita em Costa et al. (2007), a do Primeiro HAREM, descrita em Rocha e Santos (2007b) e esta que apresentamos aqui). Convém também indicar que, pelo menos do nosso conhecimento, também existem as do Corpus NILC e do Lácio-Web (descrito em Pinheiro e Aluísio (2003); Aluisio et al. (2004) e usado em Aires (2005)) e as do Corpus do Português de Referência do CLUL (Bacelar do Nascimento et al., 2000).

Por último, apresentamos nas figuras H.5 e H.6 a origem dos textos.

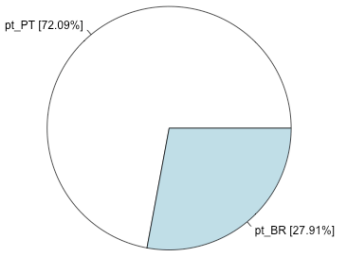
Finalmente, destacamos que todos os valores usados para criar estas tabelas constam do ficheiro meta distribuído na LÂMPADA, o pacote de recursos do Segundo HAREM.



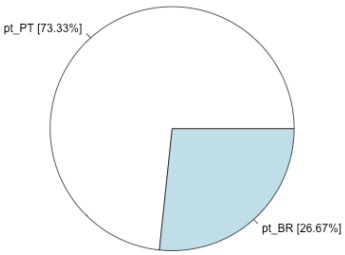
(a) Colecção do Segundo HAREM incluindo os documentos da colecção CHAVE



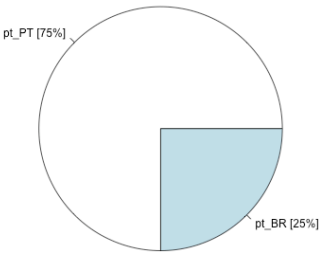
(b) Colecção do Segundo HAREM excluindo os documentos da colecção CHAVE



(c) Colecção dourada



(d) CD do TEMPO



(e) CD do ReRelEM

Figura H.1: Distribuição de documentos por variante de português

Tabela H.1: Colecção do Segundo HAREM: distribuição de documentos por variante de português, incluindo documentos da colecção CHAVE

Variante de português	Total	%
pt_PT	660	63,46%
pt_BR	380	36,54%

Tabela H.2: Colecção do Segundo HAREM: distribuição de documentos por variante de português, excluindo documentos da colecção CHAVE

Variante de português	Total	%
pt_PT	99	70,71%
pt_BR	41	29,29%

Tabela H.3: Colecção dourada: distribuição de documentos por variante de português

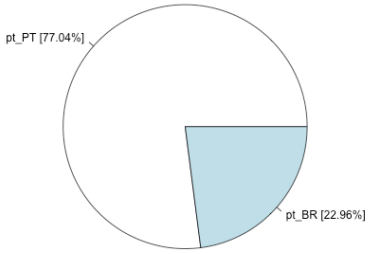
variante de português	Total	%
pt_PT	93	72,09%
pt_BR	36	27,91%

Tabela H.4: CD do TEMPO: distribuição de documentos por variante de português

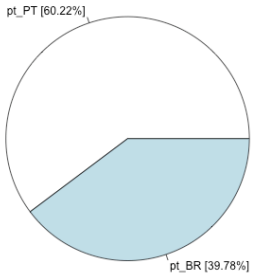
variante de português	Total	%
pt_PT	22	73,33%
pt_BR	8	26,67%

Tabela H.5: CD do ReRelEM: distribuição de documentos por variante de português

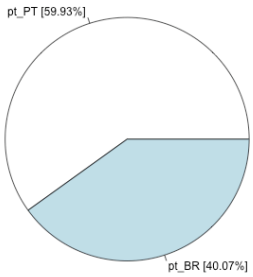
variante de português	Total	%
pt_PT	9	75%
pt_BR	3	25%



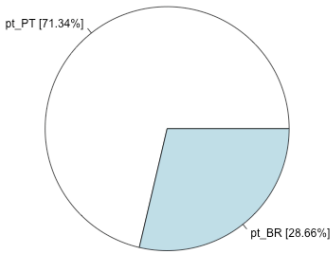
(a) Colecção do Segundo HAREM incluindo os documentos da colecção CHAVE



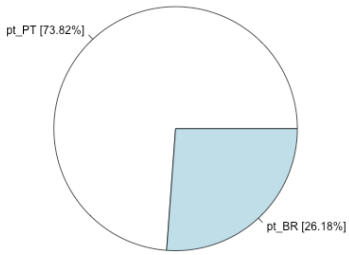
(b) Colecção do Segundo HAREM excluindo os documentos da colecção CHAVE



(c) Colecção dourada



(d) CD do TEMPO



(e) CD do ReRelEM

Figura H.2: Distribuição de palavras por variante de português

Tabela H.6: Colecção do Segundo HAREM: distribuição de palavras por variante de português, incluindo documentos da colecção CHAVE

Variante de português	Total	%
pt_PT	515264	77,04%
pt_BR	153553	22,96%

Tabela H.7: Colecção do Segundo HAREM: distribuição de palavras por variante de português, excluindo documentos da colecção CHAVE

Variante de português	Total	%
pt_PT	47615	60,22%
pt_BR	31448	39,78%

Tabela H.8: Colecção dourada: distribuição de palavras por variante de português

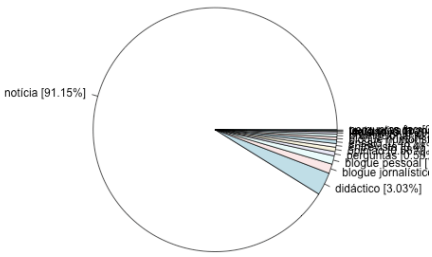
variante de português	Total	%
pt_PT	44555	59,93%
pt_BR	29795	40,07%

Tabela H.9: CD do TEMPO: distribuição de palavras por variante de português

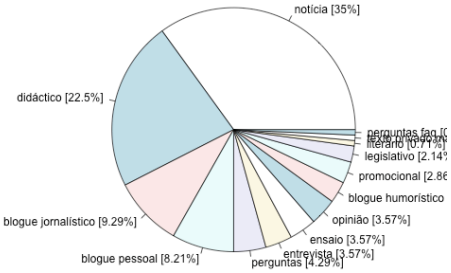
variante de português	Total	%
pt_PT	9268	71,34%
pt_BR	3724	28,66%

Tabela H.10: CD do ReRelEM: distribuição de palavras por variante de português

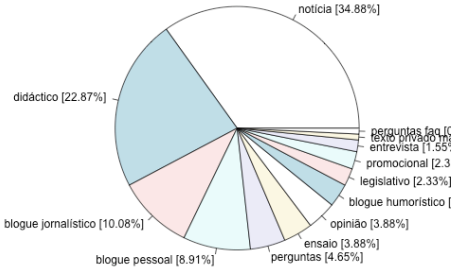
variante de português	Total	%
pt_PT	3271	73,82%
pt_BR	1160	26,18%



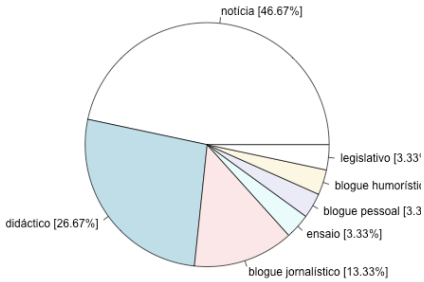
(a) Colecção do Segundo HAREM incluindo os documentos da colecção CHAVE



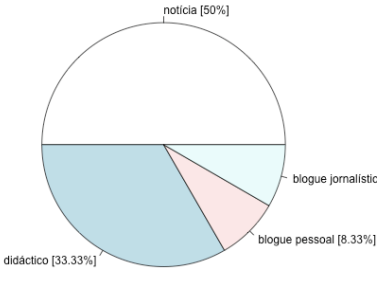
(b) Colecção do Segundo HAREM excluindo os documentos da colecção CHAVE



(c) Colecção dourada



(d) CD do TEMPO



(e) CD do ReRelEM

Figura H.3: Distribuição de documentos por tipo de texto

Tabela H.11: Colecção do Segundo HAREM: distribuição de documentos por tipo de texto, incluindo documentos da colecção CHAVE

Tipo de texto	Total	%
notícia	948	91,15%
didáctico	31,5	3,03%
blogue jornalístico	13	1,25%
blogue pessoal	11,5	1,11%
perguntas	6	0,58%
opinião	6	0,58%
entrevista	5	0,48%
ensaio	5	0,48%
blogue humorístico	4	0,38%
promocional	4	0,38%
legislativo	3	0,29%
literário	1	0,1%
texto privado manuscrito	1	0,1%
perguntas faq	1	0,1%

Tabela H.12: Colecção do Segundo HAREM: distribuição de documentos por tipo de texto, excluindo documentos da colecção CHAVE

Tipo de texto	Total	%
notícia	49	35%
didáctico	31,5	22,5%
blogue jornalístico	13	9,29%
blogue pessoal	11,5	8,21%
perguntas	6	4,29%
entrevista	5	3,57%
ensaio	5	3,57%
opinião	5	3,57%
blogue humorístico	4	2,86%
promocional	4	2,86%
legislativo	3	2,14%
literário	1	0,71%
texto privado manuscrito	1	0,71%
perguntas faq	1	0,71%

Tabela H.13: Colecção dourada: distribuição de documentos por tipo de texto

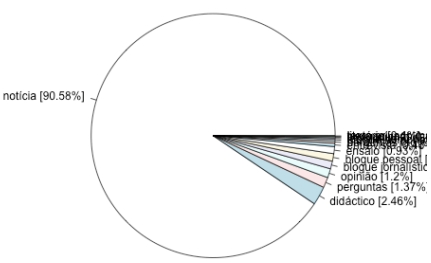
tipo de texto	Total	%
notícia	45	34,88%
didáctico	29,5	22,87%
blogue jornalístico	13	10,08%
blogue pessoal	11,5	8,91%
perguntas	6	4,65%
ensaio	5	3,88%
opinião	5	3,88%
blogue humorístico	4	3,1%
legislativo	3	2,33%
promocional	3	2,33%
entrevista	2	1,55%
texto privado manuscrito	1	0,78%
perguntas faq	1	0,78%

Tabela H.14: CD do TEMPO: distribuição de documentos por tipo de texto

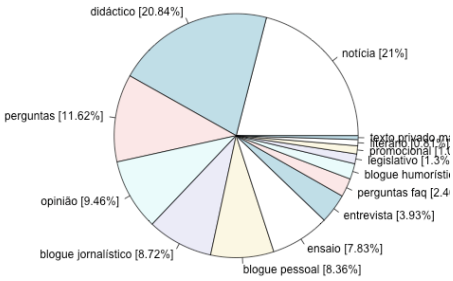
tipo de texto	Total	%
notícia	14	46,67%
didáctico	8	26,67%
blogue jornalístico	4	13,33%
ensaio	1	3,33%
blogue pessoal	1	3,33%
blogue humorístico	1	3,33%
legislativo	1	3,33%

Tabela H.15: CD do ReRelEM: distribuição de documentos por tipo de texto

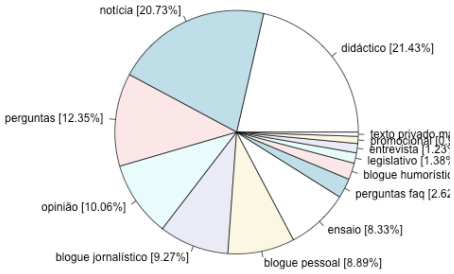
tipo de texto	Total	%
notícia	6	50%
didáctico	4	33,33%
blogue pessoal	1	8,33%
blogue jornalístico	1	8,33%



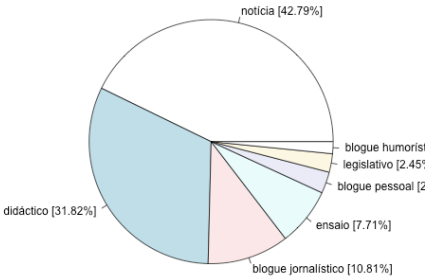
(a) Colecção do Segundo HAREM incluindo os documentos da colecção CHAVE



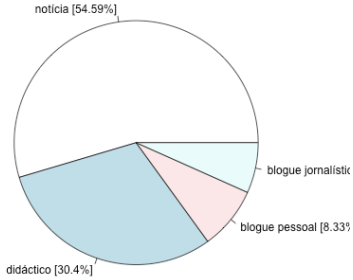
(b) Colecção do Segundo HAREM excluindo os documentos da colecção CHAVE



(c) Colecção dourada



(d) CD do TEMPO



(e) CD do ReRelEM

Figura H.4: Distribuição de palavras por tipo de texto

Tabela H.16: Colecção do Segundo HAREM: distribuição de palavras por tipo de texto, incluindo documentos da colecção CHAVE

Tipo de texto	Total	%
notícia	605815	90,58%
didáctico	16479,5	2,46%
perguntas	9184	1,37%
opinião	8023	1,2%
blogue jornalístico	6893	1,03%
blogue pessoal	6610,5	0,99%
ensaio	6193	0,93%
entrevista	3106	0,46%
perguntas faq	1945	0,29%
blogue humorístico	1639	0,25%
legislativo	1028	0,15%
promocional	850	0,13%
literário	644	0,1%
texto privado manuscrito	407	0,06%

Tabela H.17: Colecção do Segundo HAREM: distribuição de palavras por tipo de texto, excluindo documentos da colecção CHAVE

Tipo de texto	Total	%
notícia	16604	21%
didáctico	16479,5	20,84%
perguntas	9184	11,62%
opinião	7480	9,46%
blogue jornalístico	6893	8,72%
blogue pessoal	6610,5	8,36%
ensaio	6193	7,83%
entrevista	3106	3,93%
perguntas faq	1945	2,46%
blogue humorístico	1639	2,07%
legislativo	1028	1,3%
promocional	850	1,08%
literário	644	0,81%
texto privado manuscrito	407	0,51%

Tabela H.18: Colecção dourada: distribuição de palavras por tipo de texto

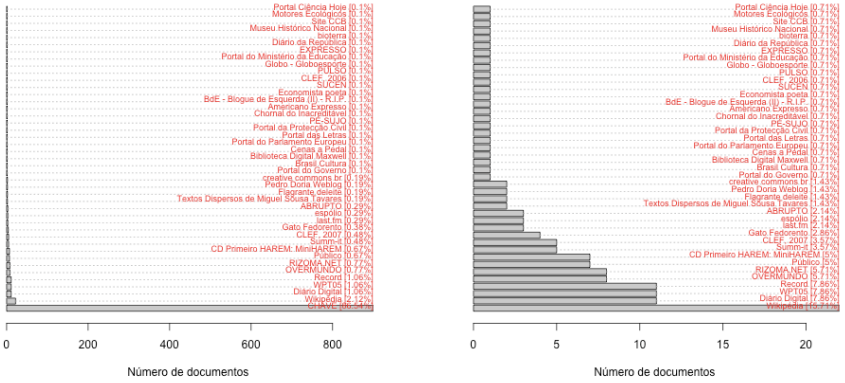
tipo de texto	Total	%
didáctico	15933,5	21,43%
notícia	15416	20,73%
perguntas	9184	12,35%
opinião	7480	10,06%
blogue jornalístico	6893	9,27%
blogue pessoal	6610,5	8,89%
ensaio	6193	8,33%
perguntas faq	1945	2,62%
blogue humorístico	1639	2,2%
legislativo	1028	1,38%
entrevista	916	1,23%
promocional	705	0,95%
texto privado manuscrito	407	0,55%

Tabela H.19: CD do TEMPO: distribuição de palavras por tipo de texto

tipo de texto	Total	%
notícia	5559	42,79%
didáctico	4134	31,82%
blogue jornalístico	1404	10,81%
ensaio	1002	7,71%
blogue pessoal	369	2,84%
legislativo	318	2,45%
blogue humorístico	206	1,59%

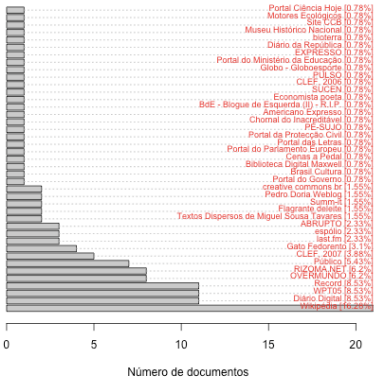
Tabela H.20: CD do ReRelEM: distribuição de palavras por tipo de texto

tipo de texto	Total	%
notícia	2419	54,59%
didáctico	1347	30,4%
blogue pessoal	369	8,33%
blogue jornalístico	296	6,68%

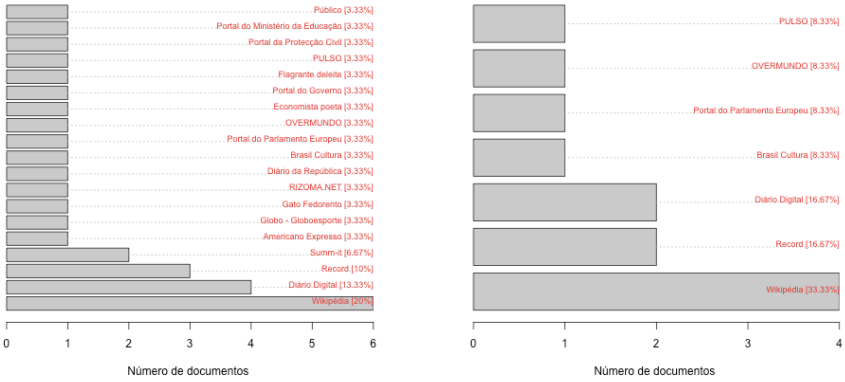


(a) Colecção do Segundo HAREM incluindo os documentos da colecção CHAVE

(b) Colecção do Segundo HAREM excluindo os documentos da colecção CHAVE



(c) Colecção dourada



(d) CD do TEMPO

(e) CD do ReRelEM

Figura H.5: Distribuição de documentos por fonte

Tabela H.21: Colecção do Segundo HAREM: distribuição de documentos por fonte, incluindo documentos da colecção CHAVE

Fonte	Total	%
CHAVE	900	86,54%
Wikipédia	22	2,12%
Diário Digital	11	1,06%
WPT05	11	1,06%
Record	11	1,06%
OVERMUNDO	8	0,77%
RIZOMA.NET	8	0,77%
Público	7	0,67%
CD Primeiro HAREM: MiniHAREM	7	0,67%
Summ-it	5	0,48%
CLEF, 2007	5	0,48%
Gato Fedorento	4	0,38%
last.fm	3	0,29%
espólio	3	0,29%
ABRUPTO	3	0,29%
Textos Dispersos de Miguel Sousa Tavares	2	0,19%
Flagrante deleite	2	0,19%
Pedro Doria Weblog	2	0,19%
creative commons br	2	0,19%
Portal do Governo	1	0,1%
Brasil Cultura	1	0,1%
Biblioteca Digital Maxwell	1	0,1%
Cenas a Pedal	1	0,1%
Portal do Parlamento Europeu	1	0,1%
Portal das Letras	1	0,1%
Portal da Protecção Civil	1	0,1%
PÉ-SUJO	1	0,1%
Chornal do Inacreditável	1	0,1%
Americano Expresso	1	0,1%
BdE - Blogue de Esquerda (II) - R.I.P.	1	0,1%
Economista poeta	1	0,1%
SUCEN	1	0,1%
CLEF, 2006	1	0,1%
PULSO	1	0,1%
Globo - Globoesporte	1	0,1%
Portal do Ministério da Educação	1	0,1%
EXPRESSO	1	0,1%
Diário da República	1	0,1%
bioterra	1	0,1%
Museu Histórico Nacional	1	0,1%
Site CCB	1	0,1%
Motores Ecológicos	1	0,1%
Portal Ciência Hoje	1	0,1%

Tabela H.22: Colecção do Segundo HAREM: distribuição de documentos por fonte, excluindo documentos da colecção CHAVE

Fonte	Total	%
Wikipédia	22	15,71%
Diário Digital	11	7,86%
WPT05	11	7,86%
Record	11	7,86%
OVERMUNDO	8	5,71%
RIZOMA.NET	8	5,71%
Público	7	5%
CD Primeiro HAREM: MiniHAREM	7	5%
Summ-it	5	3,57%
CLEF, 2007	5	3,57%
Gato Fedorento	4	2,86%
last.fm	3	2,14%
espólio	3	2,14%
ABRUPTO	3	2,14%
Textos Dispersos de Miguel Sousa Tavares	2	1,43%
Flagrante deleite	2	1,43%
Pedro Doria Weblog	2	1,43%
creative commons br	2	1,43%
Portal do Governo	1	0,71%
Brasil Cultura	1	0,71%
Biblioteca Digital Maxwell	1	0,71%
Cenas a Pedal	1	0,71%
Portal do Parlamento Europeu	1	0,71%
Portal das Letras	1	0,71%
Portal da Protecção Civil	1	0,71%
PÉ-SUJO	1	0,71%
Chornal do Inacreditável	1	0,71%
Americano Expresso	1	0,71%
BdE - Blogue de Esquerda (II) - R.I.P.	1	0,71%
Economista poeta	1	0,71%
SUCEN	1	0,71%
CLEF, 2006	1	0,71%
PULSO	1	0,71%
Globo - Globoesporte	1	0,71%
Portal do Ministério da Educação	1	0,71%
EXPRESSO	1	0,71%
Diário da República	1	0,71%
bioterra	1	0,71%
Museu Histórico Nacional	1	0,71%
Site CCB	1	0,71%
Motores Ecológicos	1	0,71%
Portal Ciência Hoje	1	0,71%

Tabela H.23: Colecção dourada: distribuição de documentos por fonte

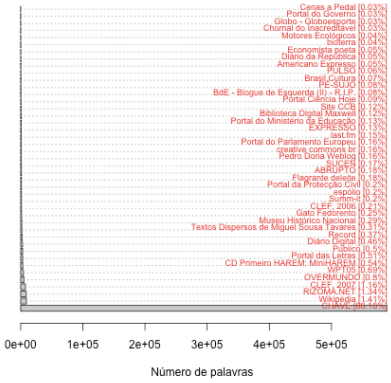
fonte	Total	%
Wikipédia	21	16,28%
Diário Digital	11	8,53%
WPT05	11	8,53%
Record	11	8,53%
OVERMUNDO	8	6,2%
RIZOMA.NET	8	6,2%
Público	7	5,43%
CLEF, 2007	5	3,88%
Gato Fedorento	4	3,1%
last.fm	3	2,33%
espólio	3	2,33%
ABRUPTO	3	2,33%
Textos Dispersos de Miguel Sousa Tavares	2	1,55%
Flagrante deleite	2	1,55%
Summ-it	2	1,55%
Pedro Doria Weblog	2	1,55%
creative commons br	2	1,55%
Portal do Governo	1	0,78%
Brasil Cultura	1	0,78%
Biblioteca Digital Maxwell	1	0,78%
Cenas a Pedal	1	0,78%
Portal do Parlamento Europeu	1	0,78%
Portal das Letras	1	0,78%
Portal da Protecção Civil	1	0,78%
PÉ-SUJO	1	0,78%
Chornal do Inacreditável	1	0,78%
Americano Expresso	1	0,78%
BdE - Blogue de Esquerda (II) - R.I.P.	1	0,78%
Economista poeta	1	0,78%
SUCEN	1	0,78%
CLEF, 2006	1	0,78%
PULSO	1	0,78%
Globo - Globoesporte	1	0,78%
Portal do Ministério da Educação	1	0,78%
EXPRESSO	1	0,78%
Diário da República	1	0,78%
bioterra	1	0,78%
Museu Histórico Nacional	1	0,78%
Site CCB	1	0,78%
Motores Ecológicos	1	0,78%
Portal Ciência Hoje	1	0,78%

Tabela H.24: CD do TEMPO: distribuição de documentos por fonte

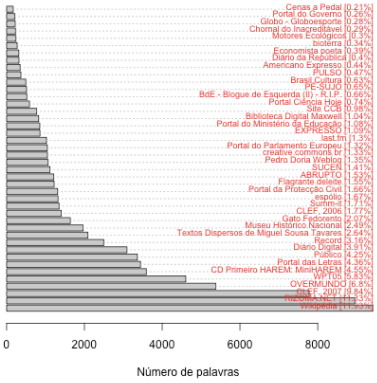
fonte	Total	%
Wikipédia	6	20%
Diário Digital	4	13,33%
Record	3	10%
Summ-it	2	6,67%
Americano Expresso	1	3,33%
Globo - Globoesporte	1	3,33%
Gato Fedorento	1	3,33%
RIZOMA.NET	1	3,33%
Diário da República	1	3,33%
Brasil Cultura	1	3,33%
Portal do Parlamento Europeu	1	3,33%
OVERMUNDO	1	3,33%
Economista poeta	1	3,33%
Portal do Governo	1	3,33%
Flagrante deleite	1	3,33%
PULSO	1	3,33%
Portal da Protecção Civil	1	3,33%
Portal do Ministério da Educação	1	3,33%
Público	1	3,33%

Tabela H.25: CD do ReRelEM: distribuição de documentos por fonte

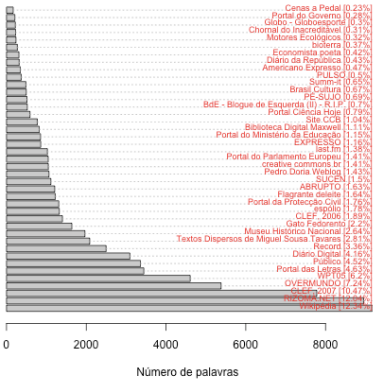
fonte	Total	%
Wikipédia	4	33,33%
Record	2	16,67%
Diário Digital	2	16,67%
Brasil Cultura	1	8,33%
Portal do Parlamento Europeu	1	8,33%
OVERMUNDO	1	8,33%
PULSO	1	8,33%



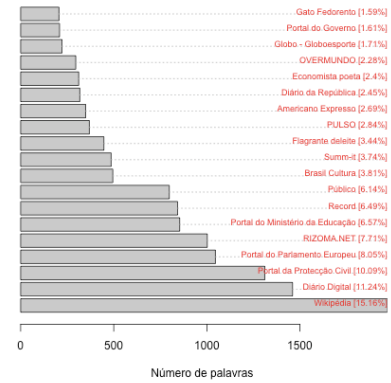
(a) Colecção do Segundo HAREM incluindo os documentos da colecção CHAVE



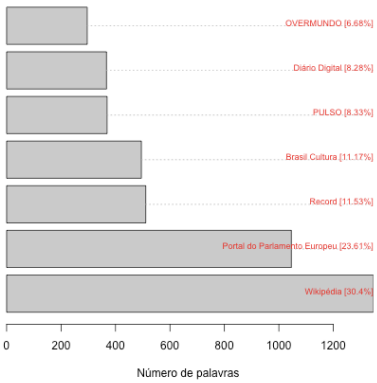
(b) Colecção do Segundo HAREM excluindo os documentos da colecção CHAVE



(c) Colecção dourada



(d) CD do TEMPO



(e) CD do ReRelEM

Figura H.6: Distribuição de palavras por fonte

Tabela H.26: Colecção do Segundo HAREM: distribuição de palavras por fonte, incluindo documentos da colecção CHAVE

Fonte	Total	%
CHAVE	589754	88,18%
Wikipédia	9430	1,41%
RIZOMA.NET	8955	1,34%
CLEF, 2007	7781	1,16%
OVERMUNDO	5380	0,8%
WPT05	4607	0,69%
CD Primeiro HAREM: MiniHAREM	3594	0,54%
Portal das Letras	3444	0,51%
Público	3361	0,5%
Diário Digital	3094	0,46%
Record	2498	0,37%
Textos Dispersos de Miguel Sousa Tavares	2087	0,31%
Museu Histórico Nacional	1966	0,29%
Gato Fedorento	1639	0,25%
CLEF, 2006	1403	0,21%
Summ-it	1350	0,2%
espólio	1323	0,2%
Portal da Protecção Civil	1311	0,2%
Flagrante deleite	1222	0,18%
ABRUPTO	1209	0,18%
SUCEN	1113	0,17%
Pedro Doria Weblog	1064	0,16%
creative commons br	1049	0,16%
Portal do Parlamento Europeu	1046	0,16%
last.fm	1026	0,15%
EXPRESSO	859	0,13%
Portal do Ministério da Educação	854	0,13%
Biblioteca Digital Maxwell	823	0,12%
Site CCB	772	0,12%
Portal Ciência Hoje	589	0,09%
BdE - Blogue de Esquerda (II) - R.I.P.	521	0,08%
PÉ-SUJO	512	0,08%
Brasil Cultura	495	0,07%
PULSO	369	0,06%
Americano Expresso	349	0,05%
Diário da República	318	0,05%
Economista poeta	312	0,05%
bioterra	272	0,04%
Motores Ecológicos	237	0,04%
Chornal do Inacreditável	230	0,03%
Globo - Globoesporte	222	0,03%
Portal do Governo	209	0,03%
Cenas a Pedal	168	0,03%

Tabela H.27: Colecção do Segundo HAREM : distribuição de palavras por fonte, excluindo documentos da colecção CHAVE

Fonte	Total	%
Wikipédia	9430	11,93%
RIZOMA.NET	8955	11,33%
CLEF, 2007	7781	9,84%
OVERMUNDO	5380	6,8%
WPT05	4607	5,83%
CD Primeiro HAREM: MiniHAREM	3594	4,55%
Portal das Letras	3444	4,36%
Público	3361	4,25%
Diário Digital	3094	3,91%
Record	2498	3,16%
Textos Dispersos de Miguel Sousa Tavares	2087	2,64%
Museu Histórico Nacional	1966	2,49%
Gato Fedorento	1639	2,07%
CLEF, 2006	1403	1,77%
Summ-it	1350	1,71%
espólio	1323	1,67%
Portal da Protecção Civil	1311	1,66%
Flagrante deleite	1222	1,55%
ABRUPTO	1209	1,53%
SUCEN	1113	1,41%
Pedro Doria Weblog	1064	1,35%
creative commons br	1049	1,33%
Portal do Parlamento Europeu	1046	1,32%
last.fm	1026	1,3%
EXPRESSO	859	1,09%
Portal do Ministério da Educação	854	1,08%
Biblioteca Digital Maxwell	823	1,04%
Site CCB	772	0,98%
Portal Ciência Hoje	589	0,74%
BdE - Blogue de Esquerda (II) - R.I.P.	521	0,66%
PÉ-SUJO	512	0,65%
Brasil Cultura	495	0,63%
PULSO	369	0,47%
Americano Expresso	349	0,44%
Diário da República	318	0,4%
Economista poeta	312	0,39%
bioterra	272	0,34%
Motores Ecológicos	237	0,3%
Chornal do Inacreditável	230	0,29%
Globo - Globoesporte	222	0,28%
Portal do Governo	209	0,26%
Cenas a Pedal	168	0,21%

Tabela H.28: Colecção dourada: distribuição de palavras por fonte

fonte	Total	%
Wikipédia	9175	12,34%
RIZOMA.NET	8955	12,04%
CLEF, 2007	7781	10,47%
OVERMUNDO	5380	7,24%
WPT05	4607	6,2%
Portal das Letras	3444	4,63%
Público	3361	4,52%
Diário Digital	3094	4,16%
Record	2498	3,36%
Textos Dispersos de Miguel Sousa Tavares	2087	2,81%
Museu Histórico Nacional	1966	2,64%
Gato Fedorento	1639	2,2%
CLEF, 2006	1403	1,89%
espólio	1323	1,78%
Portal da Protecção Civil	1311	1,76%
Flagrante deleite	1222	1,64%
ABRUPTO	1209	1,63%
SUCEN	1113	1,5%
Pedro Doria Weblog	1064	1,43%
creative commons br	1049	1,41%
Portal do Parlamento Europeu	1046	1,41%
last.fm	1026	1,38%
EXPRESSO	859	1,16%
Portal do Ministério da Educação	854	1,15%
Biblioteca Digital Maxwell	823	1,11%
Site CCB	772	1,04%
Portal Ciência Hoje	589	0,79%
BdE - Blogue de Esquerda (II) - R.I.P.	521	0,7%
PÉ-SUJO	512	0,69%
Brasil Cultura	495	0,67%
Summ-it	486	0,65%
PULSO	369	0,5%
Americano Expresso	349	0,47%
Diário da República	318	0,43%
Economista poeta	312	0,42%
bioterra	272	0,37%
Motores Ecológicos	237	0,32%
Chornal do Inacreditável	230	0,31%
Globo - Globoesporte	222	0,3%
Portal do Governo	209	0,28%
Cenas a Pedal	168	0,23%

Tabela H.29: CD do TEMPO: distribuição de palavras por fonte

fonte	Total	%
Wikipédia	1969	15,16%
Diário Digital	1460	11,24%
Portal da Protecção Civil	1311	10,09%
Portal do Parlamento Europeu	1046	8,05%
RIZOMA.NET	1002	7,71%
Portal do Ministério da Educação	854	6,57%
Record	843	6,49%
Público	798	6,14%
Brasil Cultura	495	3,81%
Summ-it	486	3,74%
Flagrante deleite	447	3,44%
PULSO	369	2,84%
Americano Expresso	349	2,69%
Diário da República	318	2,45%
Economista poeta	312	2,4%
OVERMUNDO	296	2,28%
Globo - Globoesporte	222	1,71%
Portal do Governo	209	1,61%
Gato Fedorento	206	1,59%

Tabela H.30: CD do ReRelEM: distribuição de palavras por fonte

fonte	Total	%
Wikipédia	1347	30,4%
Portal do Parlamento Europeu	1046	23,61%
Record	511	11,53%
Brasil Cultura	495	11,17%
PULSO	369	8,33%
Diário Digital	367	8,28%
OVERMUNDO	296	6,68%