A statistical study of the WPT05 crawl of the Portuguese Web

David Batista, Mário J. Silva

LaSIGE, Faculty of Sciences, Lisbon, University of Lisbon, Portugal

dsbatista@xldb.di.fc.ul.pt

Abstract

This article presents a statistical study of WPT05, a text corpus derived from a crawl of the Portuguese Web performed in 2005. This corpus is a valuable resource for researchers in Natural Language Processing (NLP). As one of the biggest publicly available collections of European Portuguese texts, we provide statistical analyses of the contents, covering the languages identified, the representativity of the top-level domains crawled and terms frequency and size. An analysis of an n-grams collection extracted from the Portuguese documents in the corpus is also presented. We analyze the occurrence of first names, surnames and geographic names in the corpus. Since some toponyms are named after personal names, we show the overlap of Portuguese names with geographic entities corresponding to places in Portugal.

Index Terms: web corpus, resources, Portuguese

1. Introduction

This study presents a statistical analysis of the textual contents of WPT05, a 2005 crawl of the Portuguese Web. WPT05 is the successor to WPT03 [1], a crawl from 2003 released earlier. WBR-99, a crawl from 1999 of the Brazilian Web, is another large collection of 6 million documents [2].

The Web pages that are part of the WPT05 Collection were retrieved by the crawler of the Tumba! search engine [3]. This crawl targeted documents written in Portuguese, hosted in a .PT domain, or hosted in the .COM, .NET, .TV, .INFO, .BIZ, .TK, .CC and .FM domains and referenced by a hyperlink from, at least, one page hosted in a .PT domain. In addition to these domains, a set of individual sites considered relevant by the developers of the crawler as well.

The content of WPT05 is available in 3 formats: as raw data, as text only documents with the metadata associated and as an n-grams collection.

The raw format includes the documents as they were crawled, without any sort of post-processing, such as filtering of some document types, elimination of duplicates, or text encoding normalization. We adopted the Internet Archive file format (ARC), designed for the specific purpose of preserving web pages as they were crawled [4].

The text-only format of the collection contains metadata associated with each document. Its production is described in the Master dissertation of David Cruz [5]. This format uses the Resource Description Framework (RDF) technology and the Open Archives Initiative Object Reuse and Exchange (OAI-ORE) specification [6]. It allows the preservation of the hierarchy of pages within each domain and the flagging of duplicate documents, for wich we mark the additional URL where the same contents were found in the case of duplicates instead of including a replica. We provide, for each document, the hierarchy of domains and duplicate information along with the identified language and crawling metadata, such as the IP address, the HTTP server running and the date and time when the document was fetched. This format contains text-rich documents only, namely, documents of the following MIME types:

- application/pdf
- application/postscript
- application/vnd.ms-office
- text/html, text/plain
- text/rtf

All the extracted text is encoded in the UTF-8 format and each file of the distributed collection is a valid XML file, enabling its handling by the tools commonly available for RDF and XML processing.

A third format of the collections is an n-grams dataset, which is described in detail in the next section. The n-grams collection was extracted from the collected documents whose identified language was Portuguese. We extracted word n-grams up to the fifht order (5-grams) using the Ngram Statistics Package [7]. A set of regular expressions to tokenize the text were applied. These regular expressions are part of the Lingua-PT-PLNbase-0.21 [8], a Perl extension for NLP of Portuguese which include a tokenizer available from Linguateca [9]. After the extraction, all n-grams with tokens having more than 32 characters were discarded. N-grams with frequencies below 5 were discarded as well. The n-grams collection is available as a set of UTF-8 encoded files, containing the n-grams and their frequencies.

2. Statistics

We analyzed the languages present in each document and the top-level domains from where documents were crawled and obtained several statistics concerning the number of unique extracted terms and the frequency and length of each, the top n-grams and also the amount of geographic information present in the WPT05 crawl. All the statistical data presented in this section was obtained from the RDF format of the collection or from the extracted n-grams. The RDF version of the collection has a total of 12,523,110 URLs, of these 9,483,489 with unique textual content.

2.1. URLs

Table 1 shows the percentage of the most targeted top-level domains (TLD) from which documents were crawled. Almost 70% of the crawl comes from .PT followed by .COM and .NET.

2.2. Language

The language for each document was detected with a popular n-gram analysis algorithm [10]. NGramJ [11], a software tool

TLD	Percentage of URLs
.pt	69.92%
.com	23.26%
.net	3.76%
.vu	1.73%
.org	1.13%
.others	0.21%

Table 1: *Top Level Domains of URLs crawled.*

implementing the algorithm was used to perform language detection in the extracted text from each document. NGramJ contains profiles for about 70 languages using up to 4-grams for identification. Only documents with more than 200 bytes in size were considered, which totals 8,877,430 documents. Documents classified as unknown correspond to harvested pages that despite presenting rich-text, the contents only contain URLs, email addresses, web server directory listings or similar contents.

Language	Nº Documents	MBytes	Percentage
Portuguese	7 412 778	24 707	83.50 %
English	941 711	3 423	10.61 %
Spanish	206 732	800	2.33 %
Others	210 014	720	2.37 %
Unknown	106 195	308	1.20 %

Table 2: *Language Distribution over documents.*

The languages per document distribution presented in Table 2 has a similar pattern as that of the crawl of the Portuguese Web from 2003 [12], although in this study the percentage of Portuguese documents was higher. The amount of Portuguese text in the collection is around 25 Gbytes, from almost 7.5 millions documents. There is no distinction between the different variations of Portuguese, such as European, Brazilian or African.

2.3. Terms

An n-gram is a subsequence of n items from a given sequence. The items can be, for example, words from a sentence, characters from a word or phonemes from a sound, depending on the application. We extracted up to 5 word n-grams from the Portuguese documents. Table 3 lists the number of unique identified n-grams from unigrams up to pentagrams, as well as the size of each set. We used the extracted unigrams to calculate the number and frequency of individual terms. As the n-grams were extracted only from documents identified as Portuguese, most of the terms have a high likelihood of being used in Portuguese.

Table 4 shows the average, median, standard deviation and mode for the frequency of terms and size of terms. Regarding the frequency, the median of 16 and the mode of 5 show that most of the identified terms have the cut-off frequency (5) of the collection. Half of all the identified terms have a frequency of 16 or less. This suggest that the term frequency, as in the crawl of 2003 [1], and other web crawls, follows a Zipf law [13].

2.4. Top N-Grams

We present in Table 5 the top 25 most frequent unigrams and bigrams. Only n-grams with tokens that do not contain any

N-Grams	Count	Size
Unigrams	2 111 004	25 Mb
Bigrams	27 674 092	432 Mb
Trigrams	71 307 404	1 400 Mb
Tetragrams	89 668 947	2 100 Mb
Pentagrams	84 378 473	2 300 Mb

Table 3: *Statistics of the WPT05 Portuguese N-Grams collection*

punctuation mark are included. These n-grams are potential candidates to a Portuguese stop-words list. Table 6 lists the top 25 trigrams. Some of the n-grams contain terms which are not Portuguese. This happens because a large portion of documents identified as Portuguese also contain English terms. These terms, such as *Blog* or *Next Blog*, are most likely part of English interfaces of content publishing systems, such as blogs.

3. Personal Names and Toponyms prevalence

We analyzed the occurrence of personal names, surnames and toponyms in the extracted n-grams. We were interested in discovering the overlap between person names and toponyms, as traditionally many geographic references, such as streets, are named after a personality's name, and many people have a placename as their surname in Portuguese.

3.1. Geographic Entities

The corpus was analyzed for the presence of geographic references. We did a search with base on Geo-Net-PT02 [14] [15] a public geographic ontology of Portugal, that contains the geographic administrative and physical data about districts, municipalities and streets, rivers, beaches, among others. We looked up in the n-grams collections for occurrences of names which correspond to geographic concepts in the geographic ontology.

Each geographic concept in Geo-Net-PT02 is associated to a name. The name is represented by 3 different variations: capitalized, non-capitalized, and simple ASCII. Table 7 shows an example of the representations. Geo-Net-PT02 contains 51,292 unique names for different geographic concepts.

We searched in n-grams for occurrences of the three different representations, 97.82% of the geographic concept names were found in WPT05 in a capitalized representation. This evidences the use of capitalization to refer to geographic place names.

Table 8 shows the coverage of geographic names in WPT05, that is, the percentage of geographic concept names found for each representation, as well as the average number of occurrences, the median, standard deviation and mode.

This approach is naive, as the occurrences of these names in WPT05 might be references to other concepts rather than only

Measure	Term Frequency	Term Size
Average	2.14	8.29
Median	16	11
Standard deviation	2 421 778	3.18
Mode	5	7

Table 4: *Term size and term occurrences statistical characterization*

Unigram	Count	Bigram	Count
de	151 331 293	para o	3 654 827
a	80 751 534	o que	3 588 803
e	78 057 840	que o	3 510 621
o	59 632 368	para a	3 450 908
que	58 002 495	e a	3 353 043
do	48 119 636	com a	3 156 764
da	39 445 585	com o	3 131 003
em	31 807 331	de um	3 122 238
para	30 871 814	que se	2 930 294
com	29 709 820	que a	2 763 518
um	24 032 617	e o	2 714 772
se	23 482 819	Todos os	2 603 578
os	21 718 820	que não	2 510 046
não	19 841 653	a sua	2 412 501
é	19 392 183	de uma	2 408 046
por	19 273 135	todos os	2 090 440
no	18 954 414	o seu	2 089 973
A	17 753 909	Powered by	1 813 626
uma	17 575 533	Responder com	1 763 910
O	17 201 084	Enviar Mensagem	1 729 514
na	15 501 678	Ver o	1 649 408
as	14 618 221	com Citação	1 636 297
dos	14 265 211	é o	1627 771
mais	13 425 740	os direitos	1 568 047
ao	11 609 150	em que	1 543 058

Table 5: *Top 25 occurring unigrams and bigrams in WPT05 corpus. N-grams with punctuation marks were removed*

Trigrams	Count
Responder com Citação	1 630 843
Ver o perfil	1 516 648
o perfil de	1 503 227
os direitos reservados	1 460 648
Enviar Mensagem Privada	1 414 293
Todos os direitos	1 366 069
perfil de utilizadores	1 196 436
de utilizadores Enviar	1 176 949
utilizadores Enviar Mensagem	1 174 793
Get your own	939 337
your own blog	934 967
Next blog BlogThis	934 480
Blogger Get your	934 450
own blog Next	915 500
blog Next blog	915 500
Voltar acima Ver	911 284
Índice do Fórum	763 560
de Julho de	759 366
Não há mensagens	756 468
há mensagens novas	731 423
a um amigo	700 625
Julho de 2005	675 378
Powered by Blogger	650 165
a última mensagem	560 605
mensagem Não há	489 854

Table 6: *Top 25 trigrams in WPT05 corpus. Trigrams with punctuation characters were removed*

geographic locations, such as personal names, organizations. However, it still provides a relevant measure of the prevalence of geographic names in WPT05.

3.2. Personal Names and Surnames

We gathered Portuguese personal names and surnames from a public list and looked for its occurrences in the WPT05 unigrams. Our list consists of 1,786 unique personal names and surnames. These were collected from the public lists of placed secondary teacher names in the 2009 recruitment, available from the Portuguese Ministry of Education website. Table 9 lists the top twenty most frequent first names and surnames. Here is also important to note that some surnames might have other semantic meanings, for instance a reference to a month.

3.3. Overlap of Personal Names, Surnames and Toponyms

Typically many first names and surnames are used as toponyms. We looked for the overlap between Portuguese names and toponyms, based on the occurrences in WPT05. From the 1,786 names, 1,030 where found to have a correspondent geographic name in Geo-Net-PT02, around 57%. Table 10 shows the top 20 most frequent Portuguese names in WPT05 that also represent a geographic concept names, and the number of geographic concepts having that name. This information could be useful for word-sense-disambiguation systems on words that can represent both a geographic concept and a person's name.

Capitalized	Non-Capitalized	Simple ASCII
Alcácer do Sal	alcácer do sal	alcacer do sal
Dão-Lafões	dão-lafões	dao-lafoes
Lisboa	lisboa	lisboa

Table 7: *Different representations of a geographic concept's name*

4. Conclusions

This was a first statistic study over the text extracted from the WPT05 collection. The raw format of WPT05 collection was produced by the XLDB Node of Linguateca in 2005. The RDF/XML was produced in 2008 and the n-grams collection was extracted in 2010. By the size of the collection and being the most part of the contents crawled from the .PT top-level domain, this is currently one of the biggest available collections in European Portuguese. The provenance of the extracted textual contents are diverse websites, spreading from personal blogs to newspapers and institutional organizations or forums. This gives a diverse and rich genera of texts, capturing different lin-

Measure	Capitalized	Non-Capitalized	ASCII
Coverage	97.8%	43.6%	42.0%
Average	5.4	3.0	5.4
Median	21	0	0
Standard deviation	62.9	58.4	199.6
Mode	1	0	0

Table 8: *Statistical characterization of occurrences of geographic names in WPT05*

Names	# Occurrences
Portugal	4 340 513
Porto	2 074 629
João	1 886 903
São	1 701 404
Pedro	1 643 292
Paulo	1 587 559
José	1 580 473
Maio	1 512 650
Janeiro	1 403 262
Novo	1 329 434
Maria	1 278 973
Silva	1 178 842
Dias	1 061 872
Bem	1 045 555
Nuno	1 034 905
Miguel	1 003 402
Carlos	971 723
Rui	969 096
Jorge	961 599
Nova	923 395
Rio	913 218
Deus	913 098
António	901 979
Santos	845 191
Manuel	834 351

Table 9: *Top 25 occurring Portuguese first names and surnames in WPT05*

Names	# Occurrences
Portugal	4 340 513
Porto	2 074 629
Pedro	1 643 292
Paulo	1 587 559
Maio	1 512 650
Janeiro	1 403 262
Novo	1 329 434
Maria	1 278 973
Silva	1 178 842
Dias	1 061 872
Miguel	1 003 402
Carlos	971 723
Jorge	961 599
Nova	923 395
Rio	913 218
Deus	913 098
Santos	845 191
Saúde	832 797
Costa	770 628
Rua	769 114
Ferreira	748 912
Luís	717 840
Ana	707 308
Tiago	692 283
Pereira	674 330

Table 10: *Top 25 overlapping Portuguese names with Portuguese geographic place names*

guistic styles.

All the three forms of the web crawl are available upon request through the Linguateca [1] and XLDB [2] websites. WPT05 is made available exclusively for research purposes.

5. Acknowledgements

We wish to thank Daniel Gomes for harvesting the documents in WPT05 and also to David Cruz for generating the RDF/XML format of the collection. This work was supported by FCT (Portugal), through the project PTDC/EIA/73614/2006 (GREASE-II) and the Multiannual Funding Programme.

6. References

[1] B. Martins and M. J. Silva, "A Statistical Study of the Tumba! Corpus," in *Advances in Natural Language Processing, 4th International Conference, EsTAL 2004, Alicante, Spain, October 20-22, 2004, Proceedings*, 2004, pp. 384–394, also available as University of Lisbon, Faculty of Sciences, Technical Report DI/FCUL TR 4-4.

[2] Pável Calado, "The WBR-99 collection: Data-structures and file formats," Department of Computer Science, Federal University of Minas Gerais, Tech. Rep., 1999. [Online]. Available: http://www.linguateca.pt/Repositorio/WBR-99/wbr99.pdf

[3] Daniel Gomes and Mário J. Silva, "The Viúva Negra crawler: an experience report," *Software: Practice and Experience (SPE)*, vol. 38, no. 2, pp. 161–168, February 2008. [Online]. Available: http://dx.doi.org/10.1002/spe.825

[4] "Internet Archive ARC File Format," http://www.digitalpreservation.gov/formats/fdd/fdd000235.shtml.

[5] D. Cruz, "Sidra5: A search system with geographic signatures," Master's thesis, University of Lisbon, Faculty of Sciences, November 2007.

[6] "Open Archives Initiative Object Reuse and Exchange," http://www.openarchives.org/ore/.

[7] "Ngram Statistics Package (NSP)," http://ngram.sourceforge.net/.

[8] "Lingua::PT::PLNbase - Perl extension for NLP of the Portuguese," http://search.cpan.org/~ambs/Lingua-PT-PLNbase-0.21/.

[9] D. Santos, "Caminhos percorridos no mapa da portuguesificação: A Linguateca em perspectiva," *Linguamática*, vol. 1, no. 1, pp. 25–58, May 2009. [Online]. Available: http://linguamatica.com/index.php/linguamatica/article/view/20/9

[10] William B. Cavnar and John M. Trenkle, "N-Gram-Based Text Categorization," in *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, 1994, pp. 161–175.

[11] "NGramJ, Smart Scanning for Document Properties," http://ngramj.sourceforge.net/.

[12] B. Martins and M. J. Silva, "Language Identification in Web Pages," in *ACM-SAC-DE, 20th ACM Symposium on Applied Computing, Document Engeneering Track*, April 2005, pp. 764–768. [Online]. Available: http://doi.acm.org/10.1145/1066677.1066852

[13] G. Zipf, *Human Behavior and the Principle of Least Effort*. Addison-Wesley (Reading MA), 1949.

[14] F. J. Lopez-Pellicer, M. Chaves, C. Rodrigues, and M. J. Silva, "Geographic ontologies production in grease-ii," University of Lisbon, Faculty of Sciences, LASIGE, Tech. Rep. TR 09-18, November 2009. [Online]. Available: http://hdl.handle.net/10455/3256

[15] M. S. Chaves, "Uma metodologia para construção de geo-ontologias," Ph.D. dissertation, Faculty of Sciences, University of Lisbon, September 2009. [Online]. Available: http://www.linguateca.pt/documentos/TeseDoutMarcirioChaves2009.pdf

[1] `http://www.linguateca.pt`

[2] `http://xldb.fc.ul.pt/wiki/WPT_05_in_English`