A statistical study of the WPT05 crawl of the Portuguese Web

David Batista, Mário J. Silva

LaSIGE, Faculty of Sciences, University of Lisbon, Portugal
dsbatista@xldb.di.fc.ul.pt

Crawl characteristics

The Web pages that are part of the WPT05 Collection were retrieved by the crawler of the Tumba! search engine. 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 was crawled as well. The collection has a total of 12,523,110 documents, of these 9,483,489 with unique textual content.

Different formats of distribution

Raw

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.

Text-Only

Contains extracted text encoded in the UTF-8 from text-rich documents, with identified language and crawling meta data. We also provide the hierarchy of domains and duplicate information allowing the preservation of the hierarchy of pages within each domain and the flagging of duplicate documents. Each file of the collection is a valid XML file, enabling its handling by RDF and XML processing tools.

N-Grams Dataset

The n-grams were extracted from the collected documents whose identified language was Portuguese. We extracted word n-grams up to the fifth order (5-grams) using the Ngram Statistics Package with a set of regular expressions from Lingua-PT-PLNbase-0.2. N-grams with tokens having more than 32 characters were discarded, as well as N-grams with frequencies below 5.

Request a copy at: http://xldb.fc.ul.pt/wiki/WPT_05_in_English

Statistical Analysis

Domains crawled

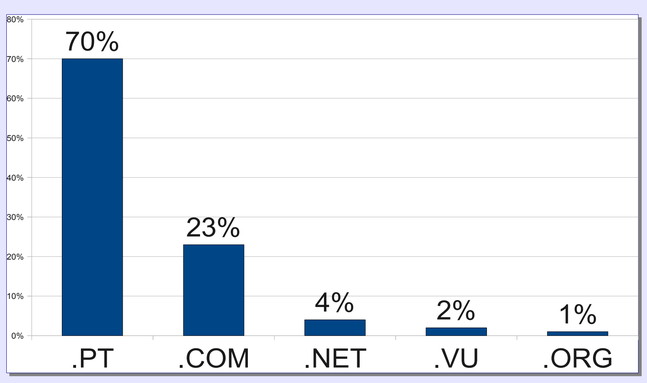

Figure 1: *Top Level Domains of URLs crawled*

Identified Languages

Language	Nº Documents	Mbytes	Percentage
Portuguese	7 412 778	24 707	83.5%
English	941 711	3 423	10.6%
Spanish	206 732	800	2.4%
Others	210 014	720	2.3%
Unknown	106 195	308	1.2%

Table 1: *Identified Languages*

N-Grams Dataset

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

Table 2: *Statistics of the N-grams dataset*

Personal Names and Toponyms prevalence in WPT05

Creative Commons Attribution 3.0 License
http://linguateca.pt/geonetpt

Geo-Net-PT02 is a public geographic ontology covering the territory of Portugal. It is divided in two domains: administrative and physical. The administrative domain contains the administrative divisions of the territory and the physical domain includes physical geography features, such as natural regions and manmade spots.

It contains 51,292 unique names for different geographic concepts represented by 3 different variations: capitalized, non-capitalized, and simple ASCII.

Capitalized	Non-Capitalized	ASCII
Dão-Lafões	dão-lafões	dao-lafoes

Table 3: *Example of a representation of geographic names in Geo-Net-PT02*

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

Measure	Capitalized	Non-Capitalized	ASCII
Coverage	97.80%	43.60%	42.00%

Table 4: *Statistical characterization of occurrences of Portuguese geographic names in WPT05*

We gathered a list of 1,786 Portuguese personal names and surnames from the public lists of placed secondary teacher names in the 2009 recruitment, available from the Portuguese Ministry of Education website, and looked for occurrences of these names in WPT05. It is important to note that some names and surnames might have other semantic meanings.

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

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%.
This information could be useful for word-sense-disambiguation systems on words that can represent both a geographic concept and a person's name.

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 6: *Top 25 most frequent Portuguese names in WPT05 that also represent a geographic concept*

This work was supported by FCT (Portugal), through the project PTDC/EIA/73614/2006 (GREASE-II) and the Multiannual Funding Programme.

We thank Daniel Gomes for harvesting the documents in WPT05 and also to David Cruz for generating the RDF/XML format of the collection.