HPLT Analytics report

General overview

Corpus	Analytics date	Language
HPLT-docslite.es.tsv	7/20/2024	Spanish (es)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
129,291,075	18,462,293,855			1.04 TB	

Top 10 domains

Domain	Docs	% of total
blogspot.com.es	6.4M	4.99
blogspot.com.ar	3.9M	3.05
blogspot.mx	3.5M	2.68
blogspot.com	3.3M	2.55
buenastareas.com	2.2M	1.67
wordpress.com	864K	0.67
dhgate.com	577K	0.45
blogspot.pe	531K	0.41
alibaba.com	522K	0.40
blogspot.cl	372K	0.29

Top 10 TLDs

Domain	Docs	% of total
com	63M	48.40
es	14M	10.77
com.ar	9M	6.96
com.es	6.7M	5.15
mx	5.2M	4.06
org	5.2M	4.04
net	3.8M	2.93
cl	3.3M	2.55
com.mx	1.7M	1.31
info	1.2M	0.94

Documents size (in segments)

<= 25 segments **7.05%** (9.1M documents)
> 25 segments **92.95%** (120M documents)
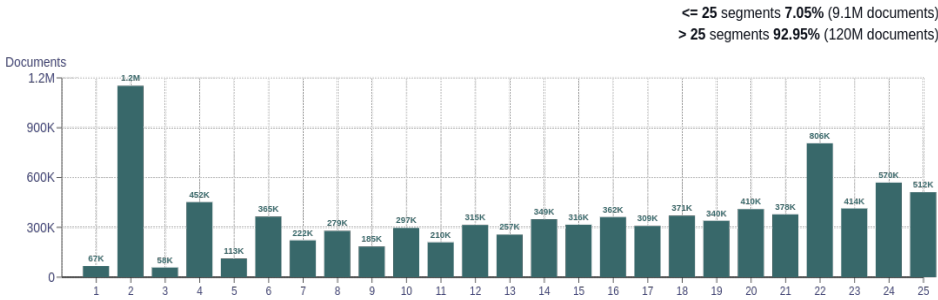
Documents by collection

cc40 (43M), wide15 (37M), wide17 (17M), wide16 (32M)
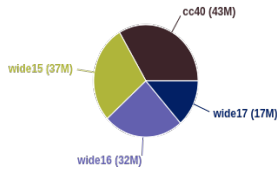
Language Distribution

Number of segments

- Spanish (es) - 13B
- English (en) - 2.3B
- Portuguese (pt) - 572M
- Italian (it) - 547M
- Czech (cs) - 439M
- French (fr) - 414M
- Galician (gl) - 325M
- German (de) - 256M
- Catalan (ca) - 119M
- Polish (pl) - 57M
- 164 Others - 677M
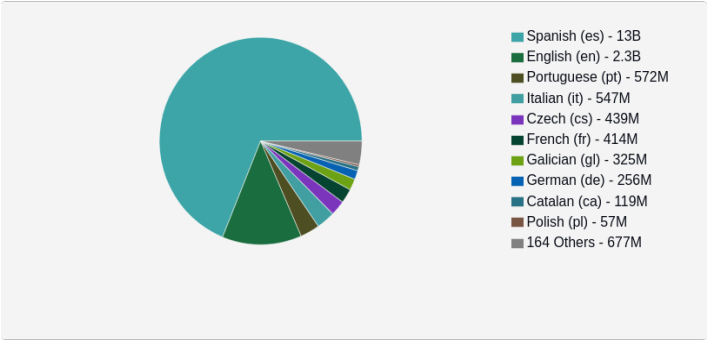
Percentage of segments in Spanish (es) inside documents
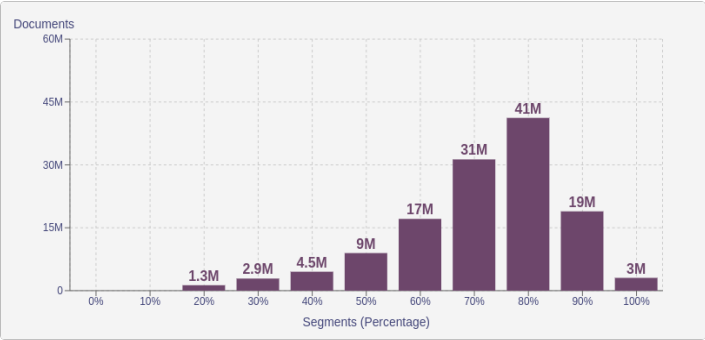
Distribution of documents by document score

score < 5 - **23.27%** (30M documents)
score >= 5 - **76.73%** (99M documents)

Segment noise distribution

- Too long: 0.27 %
- Too short: 34.94 %
- URLs: 1.69 %
- Bad encoding: 0.00 %
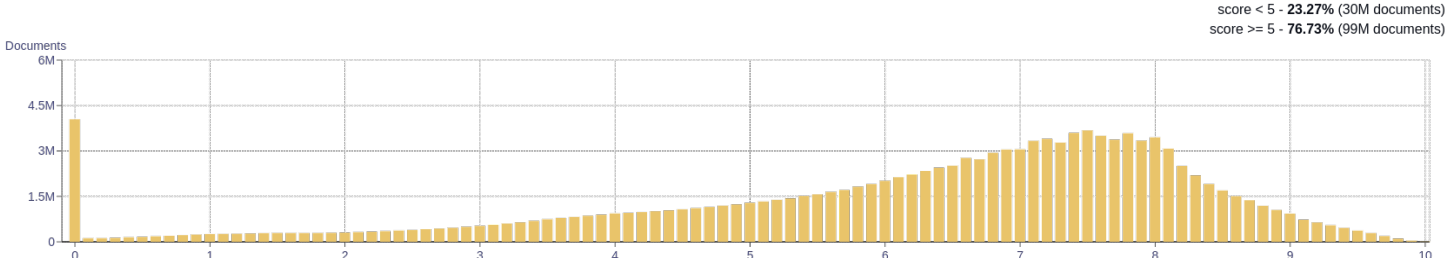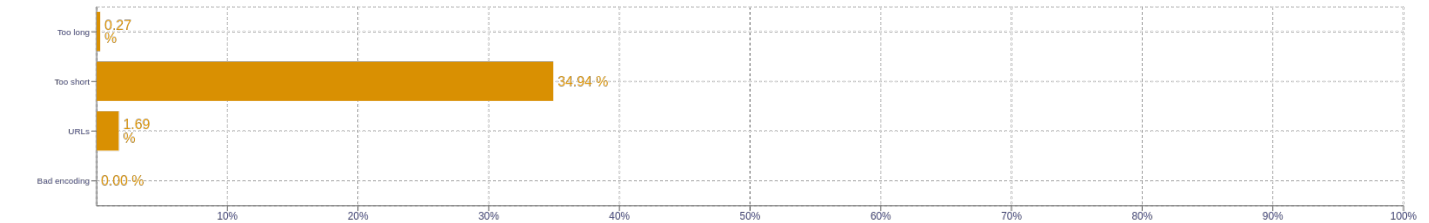
About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, <ul>, <ol>, etc.) replaced by newlines.

Volumes - Tokens

Tokenized with https://github.com/hplt-project/data-analytics-tool/blob/main/tokenizers-info.md

Type-Token Ratio

Lexical variety computed as *number or types (uniques)/number of tokens*, after removing punctuation (https://www.sltinfo.com/wp-content/uploads/2014/01/type-token-ratio.pdf).

Document size (in segments)

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, <ul>, <ol>, etc.) replaced by newlines.

Language distribution

Language identified with FastSpell (https://github.com/mbanon/fastspell).

Distribution of segments by fluency score

Obtained with Monocleaner (https://github.com/bitextor/monocleaner).

Distribution of documents by average fluency score

Obtained with Monocleaner (https://github.com/bitextor/monocleaner).

Distribution of documents by document score

Obtained with Web Docs Scorer (https://github.com/pablop16n/web-docs-scorer/).

Segment length distribution by token

Tokenized with https://github.com/hplt-project/data-analytics-tool/blob/main/tokenizers-info.md

Segment noise distribution

Obtained with Bicleaner Hardrules (https://github.com/bitextor/bicleaner-hardrules/).

Frequent n-grams

Tokenized with https://github.com/hplt-project/data-analytics-tool/blob/main/tokenizers-info.md, after removing n-grams starting or ending in a stopword. Stopwords from https://github.com/hplt-project/data-analytics-tool/blob/main/scripts/resources/README.txt