HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
HPLT-v2-por_Latn.tsv	9/26/2024	Portuguese (pt)

Volumes

Docs	Segments	Characters	Size
237,812,825	6,124,583,396	890,659,363,336	855.25 GB

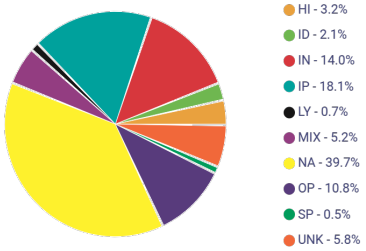
Top 10 domains

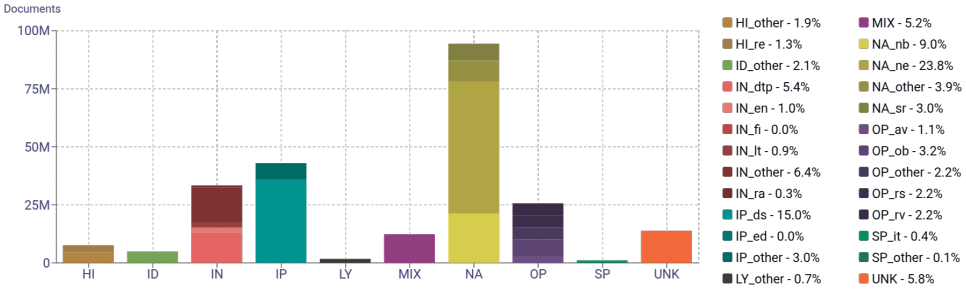
Domain	Docs	% of total
blogspot.com	20M	8.44%
blogspot.com.br	12M	4.84%
blogspot.pt	5.9M	2.49%
uol.com.br	4.3M	1.81%
sapo.pt	4.1M	1.72%
wordpress.com	3.9M	1.63%
globo.com	2.3M	0.98%
wikipedia.org	1.6M	0.69%
estadao.com.br	880K	0.37%
ig.com.br	816K	0.34%

Top 10 TLDs

Domain	Docs	% of total
com.br	96M	40.39%
com	79M	33.30%
pt	21M	8.96%
org	6.6M	2.77%
net	5.7M	2.38%
org.br	5.4M	2.29%
br	2.3M	0.95%
info	868K	0.37%
edu.br	831K	0.35%
blog.br	712K	0.30%

Register labels

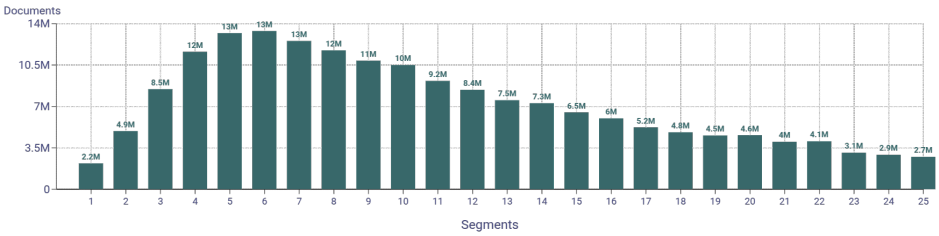




MT:1.9% | 4.5M Documents

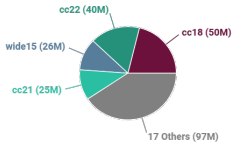
Documents size (in segments)

<= 25 segments **75.79%** (180M documents)
> 25 segments **24.21%** (58M documents)



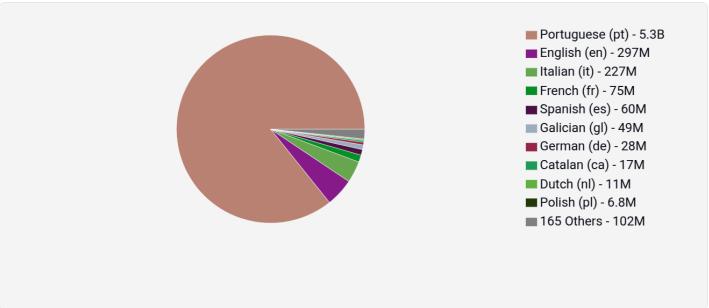
Documents by collection

CC = 60.12%
IA = 39.88%

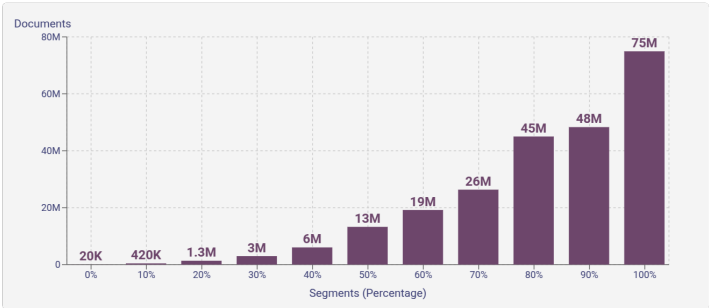


Language Distribution

Number of segments in the Portuguese (pt) corpus

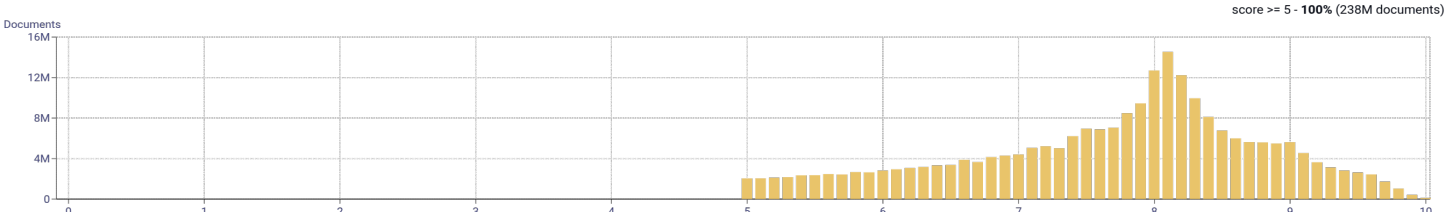


Percentage of segments in Portuguese (pt) inside documents

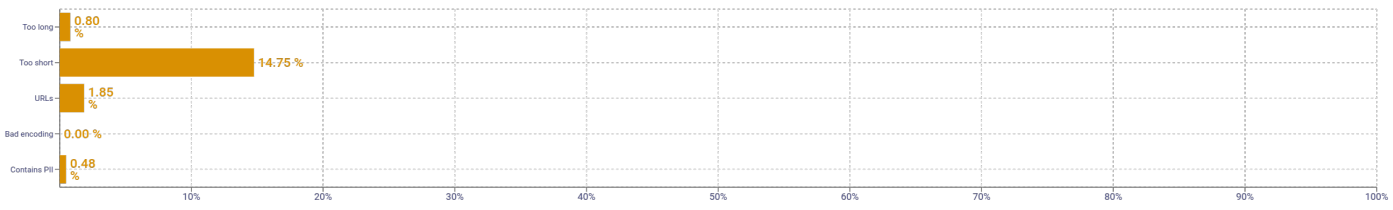


Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (238M documents)



Segment noise distribution



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

Register labels

Name	Abbr.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	**HI**
Recipe	re
Informational persuasion	**IP**
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	**IN**
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	**OP**
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av