HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
HPLT-docslite.it.tsv	7/2/2024	Italian (it)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
53,525,580	7,467,332,489			476.72 GB	

Top 10 domains

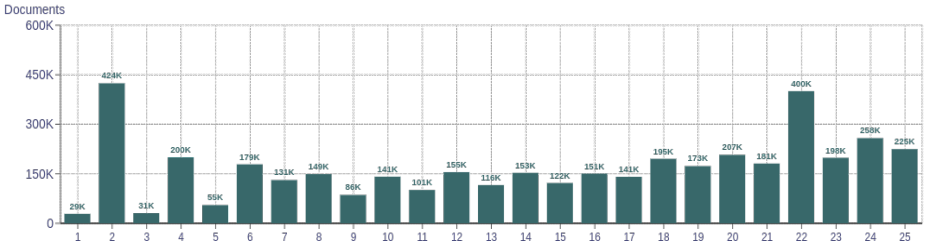
Domain	Docs	% of total
blogspot.it	2.4M	4.47
blogspot.com	769K	1.44
dhgate.com	532K	0.99
alibaba.com	521K	0.97
ibiza2015.it	492K	0.92
wordpress.com	354K	0.66
docplayer.it	240K	0.45
aliexpress.com	207K	0.39
blogspot.ch	180K	0.34
repubblica.it	160K	0.30

Top 10 TLDs

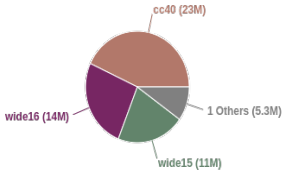
Domain	Docs	% of total
it	28M	53.16
com	16M	29.01
org	2.1M	3.83
net	1.7M	3.19
eu	832K	1.56
ch	612K	1.14
info	603K	1.13
de	175K	0.33
biz	159K	0.30
tv	154K	0.29

Documents size (in segments)

<= 25 segments **7.85%** (4.2M documents)
> 25 segments **92.15%** (49M documents)

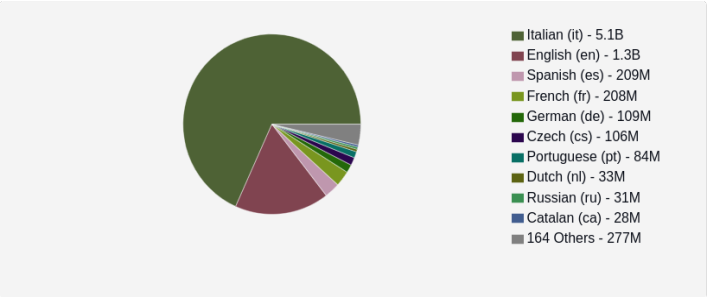


Documents by collection

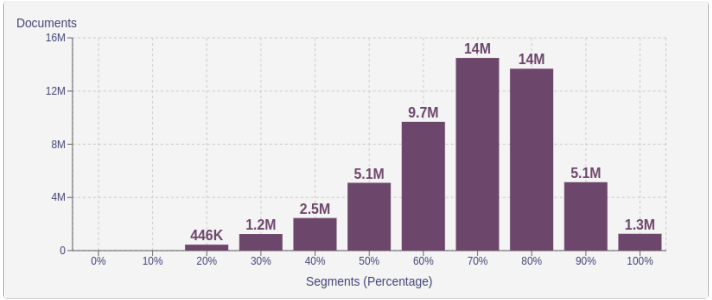


Language Distribution

Number of segments

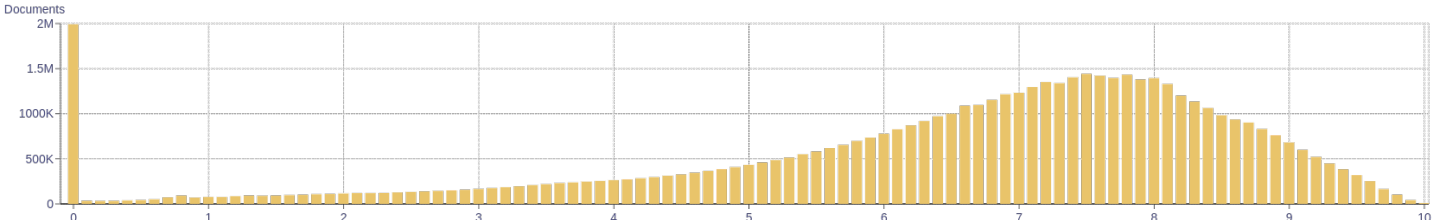


Percentage of segments in Italian (it) inside documents



Distribution of documents by document score

score <= 5 - **18.89%** (10M documents)
score > 5 - **81.11%** (43M documents)



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