HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
HPLT-docslite.ca.tsv	6/8/2024	Catalan (ca)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
4,543,154	623,010,179	113,244,981 (18.18 %)	7.3B	33.42 GB	

Top 10 domains

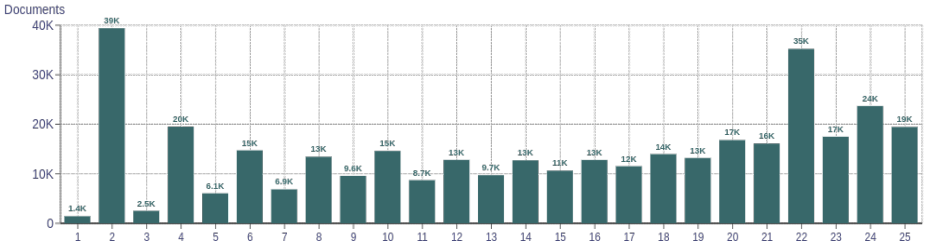
Domain	Docs	% of total
blogspot.com.es	528K	11.62
blogspot.com	244K	5.37
wordpress.com	61K	1.35
wikipedia.org	54K	1.18
vivados.es	37K	0.82
gencat.cat	35K	0.76
ara.cat	28K	0.62
cugat.cat	22K	0.49
agoda.com	21K	0.46
uab.cat	19K	0.41

Top 10 TLDs

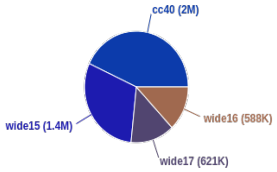
Domain	Docs	% of total
cat	1.5M	33.95
com	1.4M	31.20
com.es	529K	11.65
org	343K	7.56
es	252K	5.55
net	111K	2.44
edu	43K	0.95
info	42K	0.93
ad	20K	0.44
eu	19K	0.42

Documents size (in segments)

<= 25 segments **8%** (363K documents)
> 25 segments **92%** (4.2M documents)

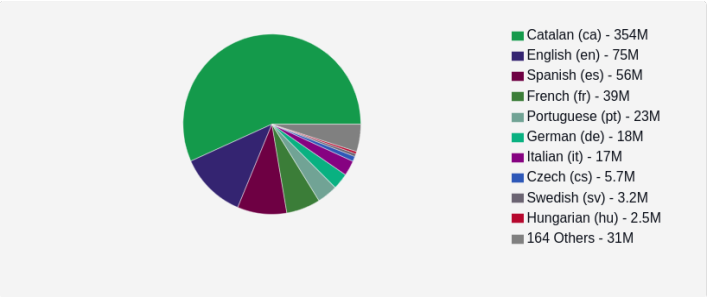


Documents by collection

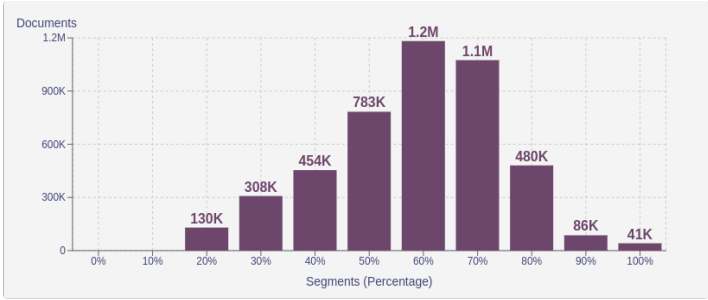


Language Distribution

Number of segments

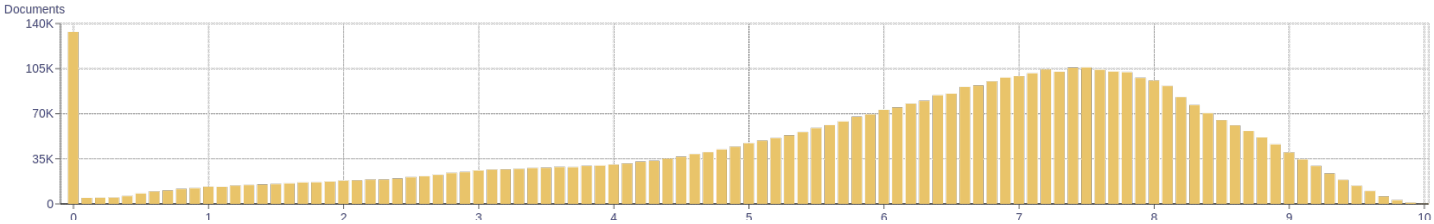


Percentage of segments in Catalan (ca) inside documents



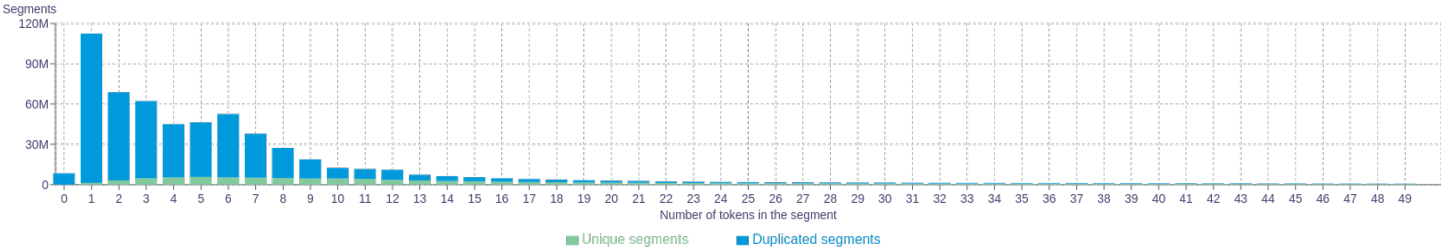
Distribution of documents by document score

score <= 5 - **26.62%** (1.2M documents)
score > 5 - **73.38%** (3.3M documents)



Segment length distribution by token

<= 49 tokens = **92M** segments | **500M** duplicates
> 50 tokens = **31M** segments | **9.6M** duplicates



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