

General overview

Corpus	Analytics date	Language
HPLT-v2-fra_Latn.tsv	10/3/2024	French (fr)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
401,831,660	10,557,090,126			1.37 TB	1,447,270,554,770

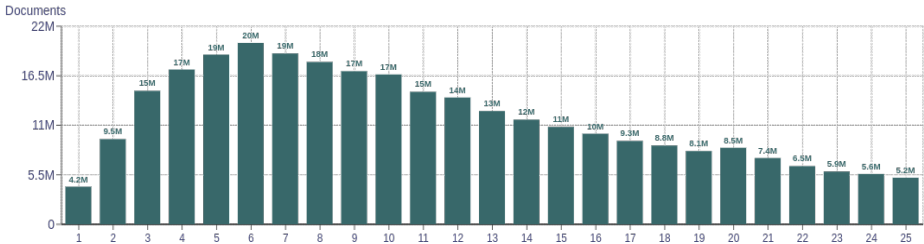
Top 10 domains

Domain	Docs	% of total
canalblog.com	5.8M	1.44
blogspot.com	5.5M	1.36
wikipedia.org	3.2M	0.81
wordpress.com	3.1M	0.78
over-blog.com	3.1M	0.76
blogspot.fr	2.6M	0.66
lefigaro.fr	1.2M	0.31
lexpress.fr	970K	0.24
herokuapp.com	901K	0.22
skyrock.com	899K	0.22

Top 10 TLDs

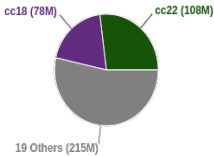
Domain	Docs	% of total
com	182M	45.32
fr	120M	29.87
org	24M	6.00
net	16M	3.98
be	11M	2.66
ca	8.3M	2.06
ch	7M	1.75
info	5.4M	1.34
eu	3.6M	0.90
co	1.4M	0.35

Documents size (in segments)



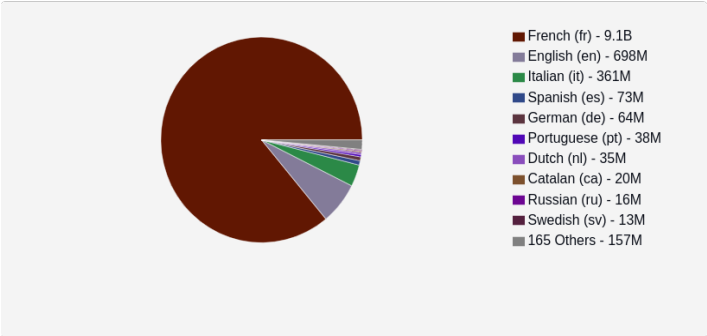
<= 25 segments 73.25% (294M documents)
> 25 segments 26.75% (107M documents)

Documents by collection

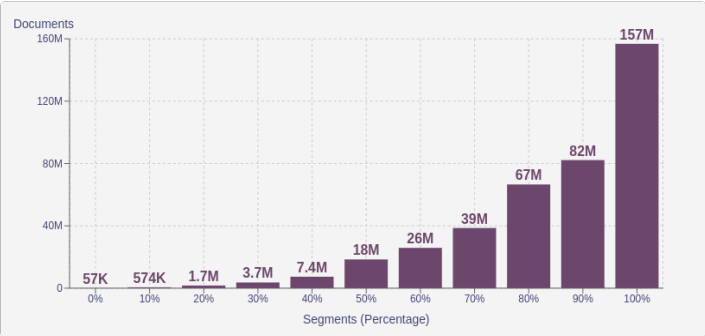


Language Distribution

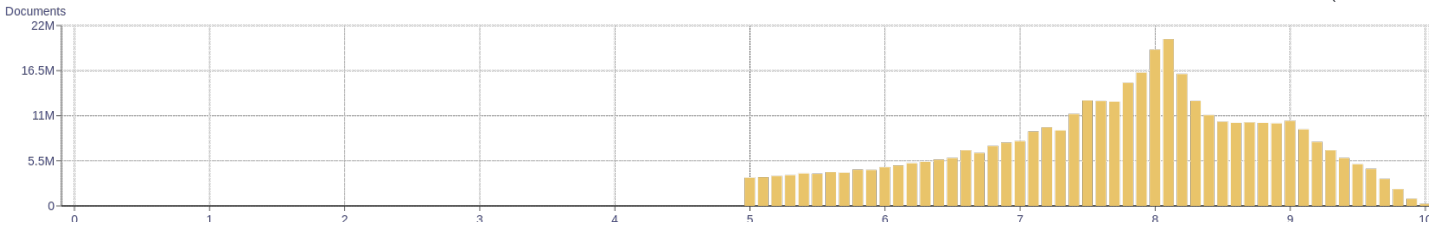
Number of segments



Percentage of segments in French (fr) inside documents

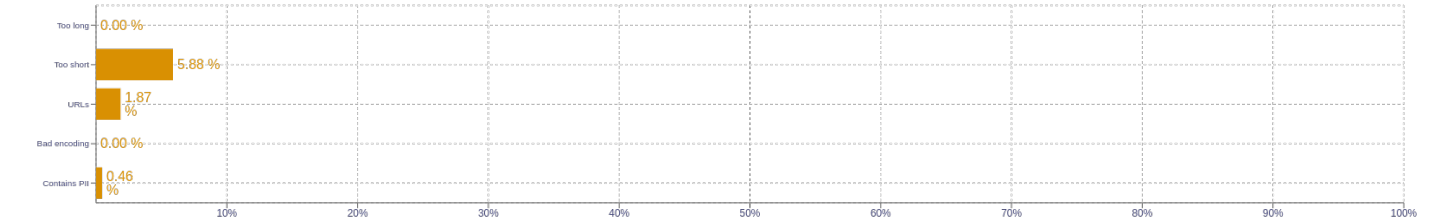


Distribution of documents by document score



score < 5 - 0% (0 documents)
score >= 5 - 100% (402M documents)

Segment noise distribution



About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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 of 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>, , , 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>