HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
HPLT-docslite.fr.tsv	7/4/2024	French (fr)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
99,587,633	13,184,580,399			742.32 GB	

Top 10 domains

Domain	Docs	% of total
blogspot.fr	1.7M	1.69
diebuchsuche.com	1.1M	1.08
blogspot.com	684K	0.69
alibaba.com	676K	0.68
dhgate.com	545K	0.55
canalblog.com	539K	0.54
blogspot.be	480K	0.48
over-blog.com	466K	0.47
wordpress.com	448K	0.45
herokuapp.com	426K	0.43

Top 10 TLDs

Domain	Docs	% of total
com	45M	44.70
fr	29M	29.36
org	5.1M	5.10
net	4M	3.98
be	3M	2.98
ca	2.1M	2.09
ch	2.1M	2.09
info	1.2M	1.22
eu	933K	0.94
tk	507K	0.51

Documents size (in segments)

<= 25 segments **10.11%** (10M documents)
> 25 segments **89.89%** (89M documents)
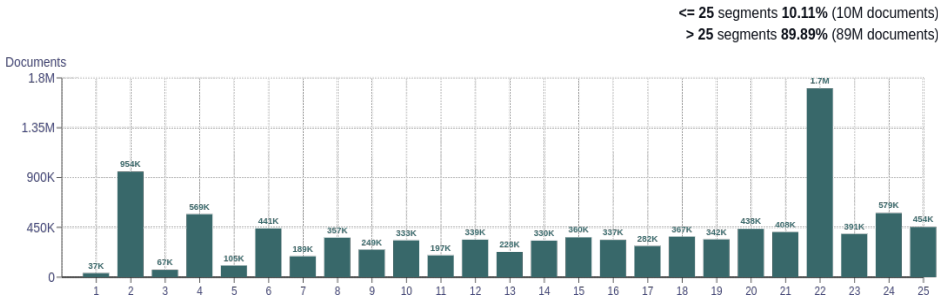
Documents by collection

cc40 (40M)
wide16 (26M)
wide17 (16M)
wide15 (18M)
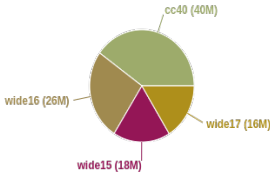
Language Distribution

Number of segments

- French (fr) - 9.5B
- English (en) - 2.4B
- German (de) - 238M
- Spanish (es) - 205M
- Italian (it) - 180M
- Portuguese (pt) - 75M
- Dutch (nl) - 72M
- Catalan (ca) - 49M
- Russian (ru) - 47M
- Czech (cs) - 44M
- 164 Others - 434M
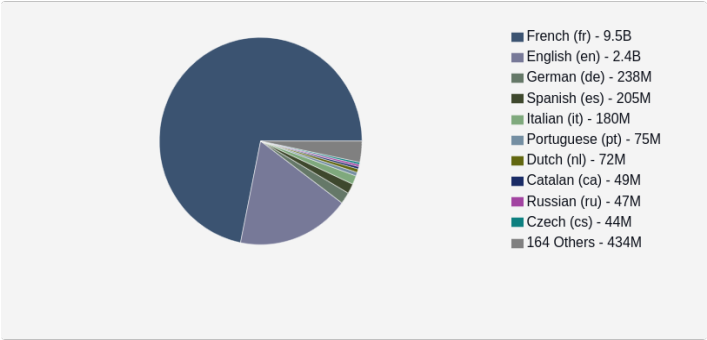
Percentage of segments in French (fr) inside documents
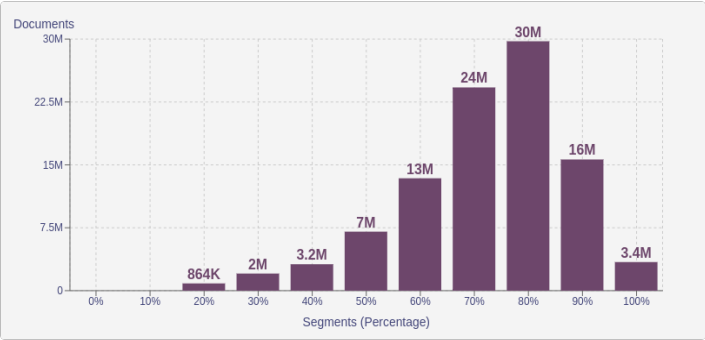
Distribution of documents by document score

score < 5 - **20.83%** (21M documents)
score >= 5 - **79.17%** (79M documents)

Segment noise distribution

Too long	0.28 %
Too short	37.16 %
URLs	1.93 %
Bad encoding	0.00 %
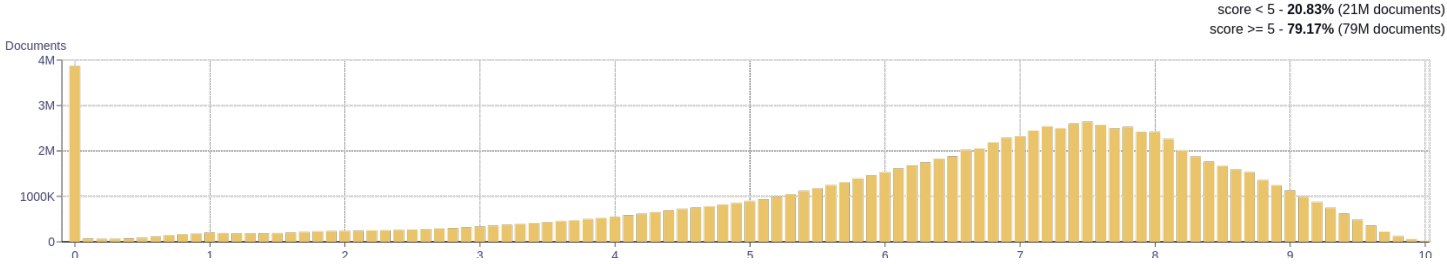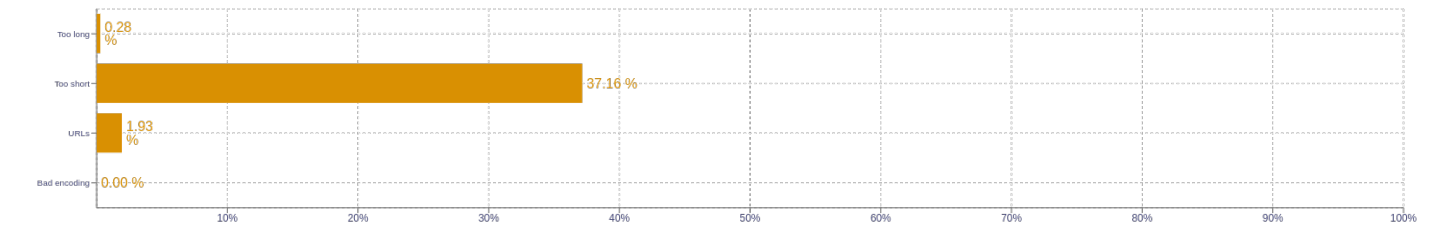
About HPLT Analytics

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