HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
la_1.jsonl.tsv	3/20/2024	Latin (la)

Volumes

Docs	Segments	Unique segments	Tokens	Size
301,702	21,396,940	36,606 (0.17 %)	369M	1.88 GB

Type-Token Ratio

Latin (la)
0.01

Top 10 domains

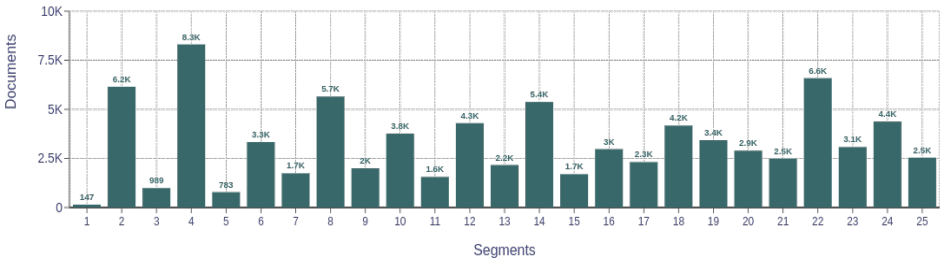
Domain	Docs	% of total
wikipedia.org	10K	3.34
bibliacatolica.com.br	4.2K	1.39
latin.it	2.7K	0.89
blogspot.com	2.2K	0.73
intratext.com	2K	0.68
vatican.va	995	0.33
monacodebacardi.com	758	0.25
wordplanet.org	708	0.23
ucl.ac.be	665	0.22
monumenta.ch	578	0.19

Top 10 TLDs

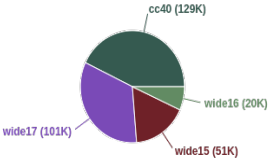
Domain	Docs	% of total
com	153K	50.74
org	25K	8.31
de	10K	3.36
it	9K	2.98
com.br	7.5K	2.47
net	6.8K	2.25
co.uk	4.7K	1.56
nl	4.7K	1.54
fr	4.5K	1.50
pl	3.4K	1.14

Documents size (in segments)

<= 25 segments **27.47%** (83K documents)
> 25 segments **72.53%** (219K documents)

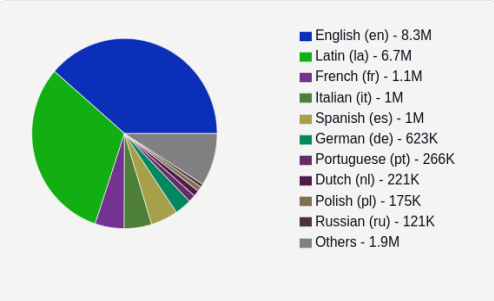


Documents by collection

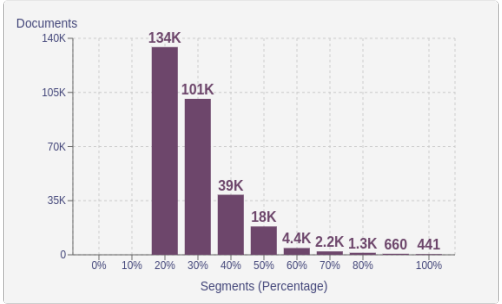


Language Distribution

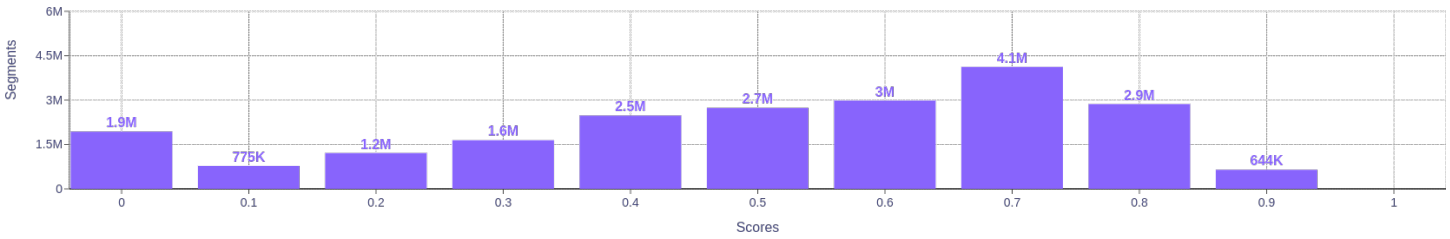
Number of segments



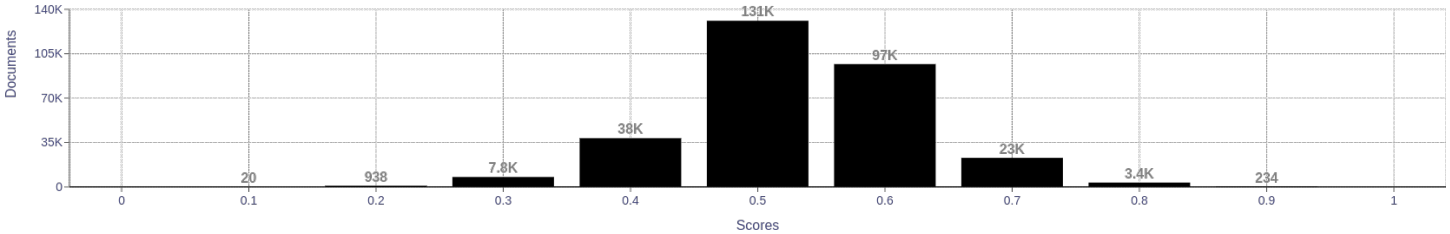
Percentage of segments in Latin (la) inside documents



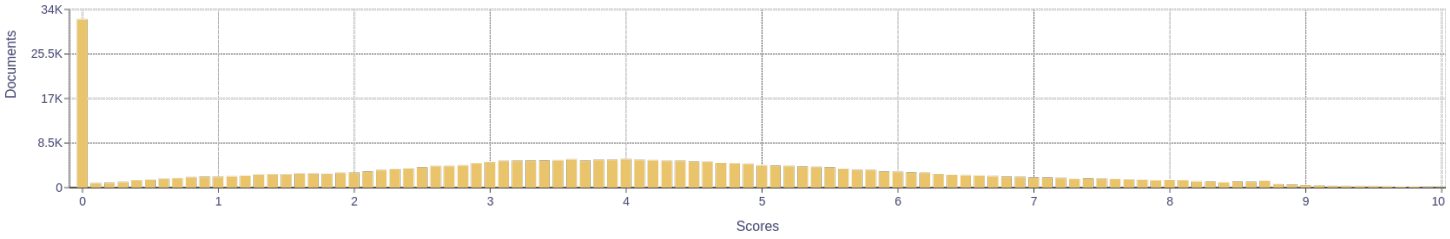
Distribution of segments by fluency score



Distribution of documents by average fluency score

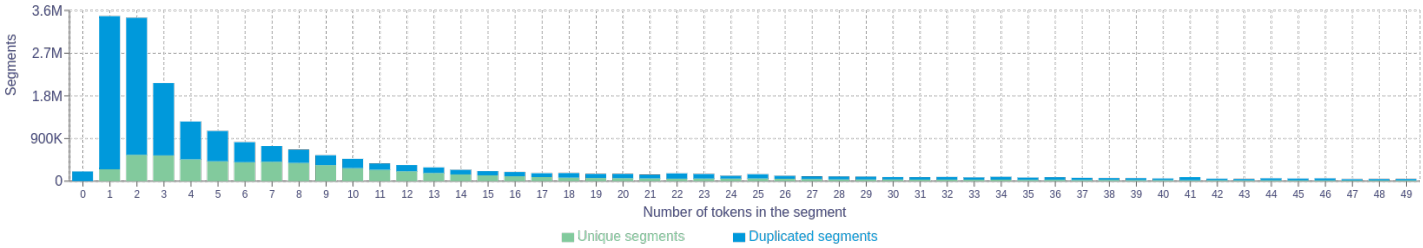


Distribution of documents by document score

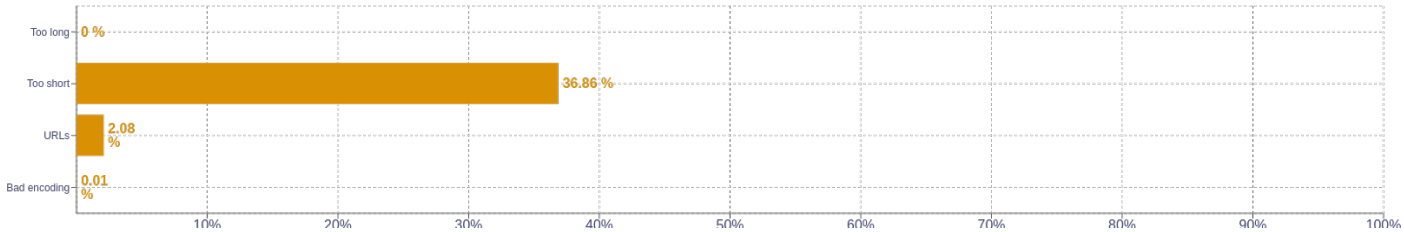


Segment length distribution by token

<= 49 tokens = **6.3M** segments | **13M** duplicates
> 50 tokens = **1.8M** segments | **983K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	sit | 3145132 · amet | 2819745 · dolor | 2672127 · ipsum | 2634268 · lorem | 2322712
2	sit amet | 2633291 · lorem ipsum | 1708690 · dolor sit | 1672241 · ipsum dolor | 1666735 · adipiscing elit | 972811
3	dolor sit amet | 1643934 · ipsum dolor sit | 1602828 · lorem ipsum dolor | 1576412 · consectetur adipiscing elit | 724402 · labore et dolore | 426399
4	ipsum dolor sit amet | 1581118 · lorem ipsum dolor sit | 1524857 · labore et dolore magna | 380833 · do eiusmod tempor incididunt | 330349 · tempor incididunt ut labore | 320449
5	lorem ipsum dolor sit amet | 1504558 · eiusmod tempor incididunt ut labore | 314451 · incididunt ut labore et dolore | 314248 · labore et dolore magna aliqua | 306965 · aliquip ex ea commodo consequat | 191192

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