HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
lij_Latn.jsonl.tsv	9/25/2024	Ligurian (lij)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
8,371	157,715	76,776 (48.68 %)	7.4M	33.67 MB	31,311,759

Top 10 domains

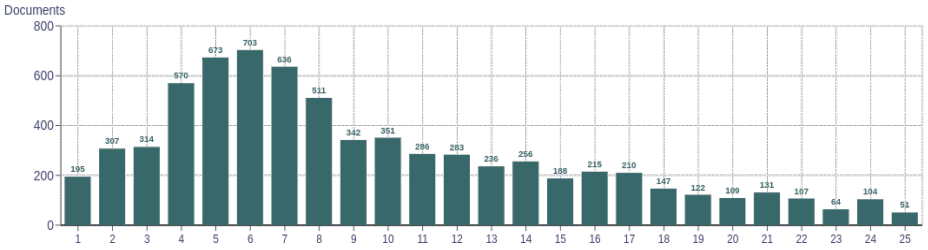
Domain	Docs	% of total
wikipedia.org	6.7K	80.48
wikisource.org	126	1.51
ilsecoloxix.it	91	1.09
bible.is	78	0.93
primocanale.it	49	0.59
blogspot.com	36	0.43
eodishasamachar.com	36	0.43
in-links.eu	36	0.43
genovapress.com	34	0.41
bywiki.com	29	0.35

Top 10 TLDs

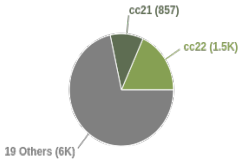
Domain	Docs	% of total
org	7K	83.97
com	532	6.36
it	356	4.25
net	86	1.03
is	78	0.93
eu	50	0.60
si	20	0.24
com.ar	17	0.20
ru	14	0.17
com.br	14	0.17

Documents size (in segments)

<= 25 segments **84.95%** (7.1K documents)
> 25 segments **15.05%** (1.3K documents)

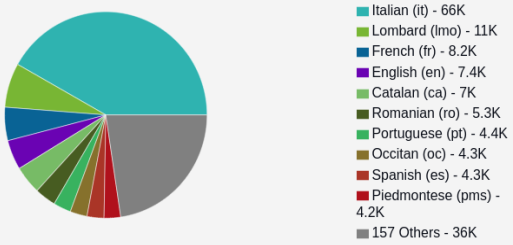


Documents by collection



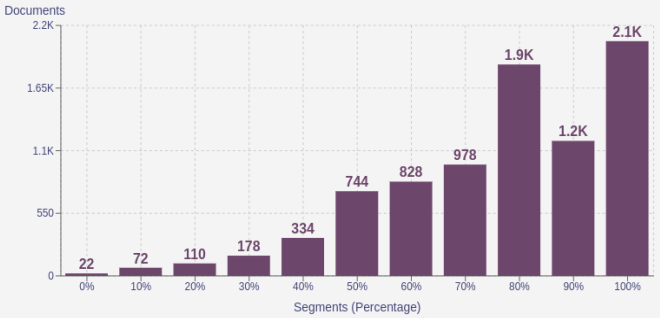
Language Distribution

Number of segments



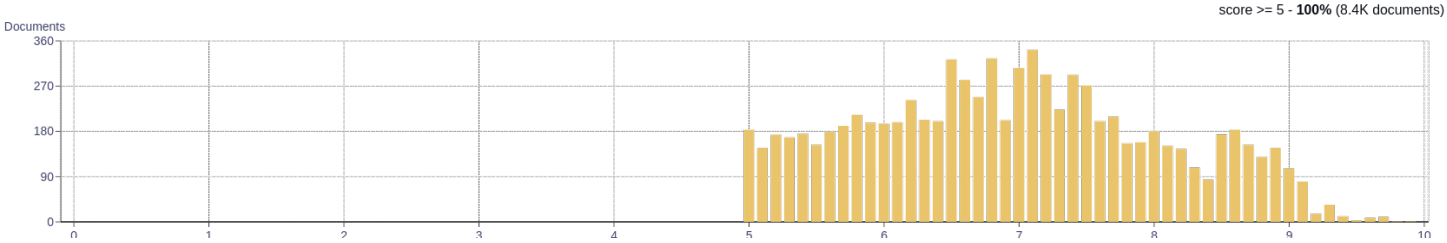
*Ligurian (lij) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Ligurian (lij) inside documents



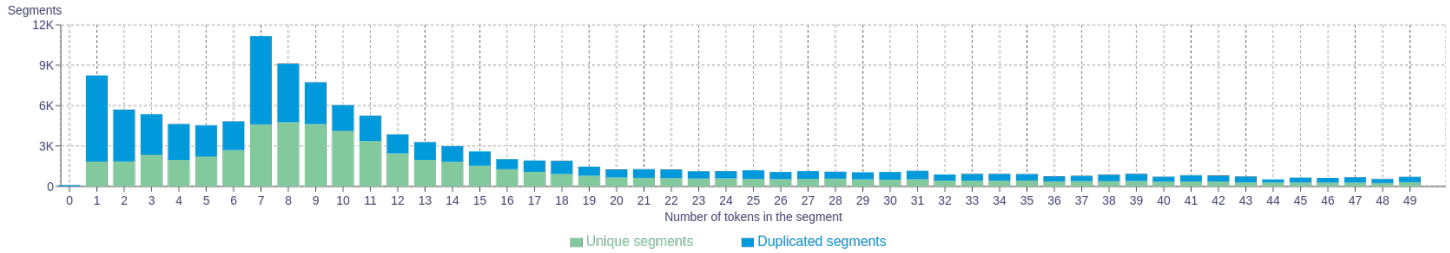
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (8.4K documents)

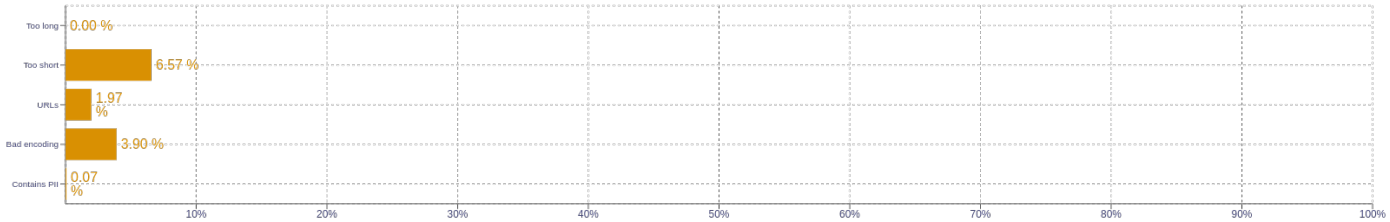


Segment length distribution by token

<= 49 tokens = **60K** segments | **61K** duplicates
> 50 tokens = **37K** segments | **20K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	l | 167155 · d | 45088 · n | 37951 · è | 34758 · u | 32327
2	modifica wikitèsto | 8239 · modifica wikitesto | 4938 · u l | 3034 · url consultòu | 2327 · u n | 2272
3	ö ö ö | 1852 · commons a contêgne | 1733 · p e r | 1148 · c o n | 1092 · z io n | 966
4	ö ö ö ö | 1848 · wikimedia commons a contêgne | 1733 · inmàgine ò di âtri | 760 · zeauriniâle zeauriniâle zeauriniâle zeauriniâle | 602 · miou miou miou miou | 434
5	ö ö ö ö ö | 1844 · commons a contêgne di files | 934 · inmàgine ò di âtri files | 760 · commons a contêgne di inmàgine | 740 · zeauriniâle zeauriniâle zeauriniâle zeauriniâle zeauriniâle | 592

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