HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
lmo_Latn.jsonl.tsv	12/4/2024	Lombard (lmo)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
146,162	2,124,663	841,958 (39.63 %)	76M	340.5 MB	343,386,720

Top 10 domains

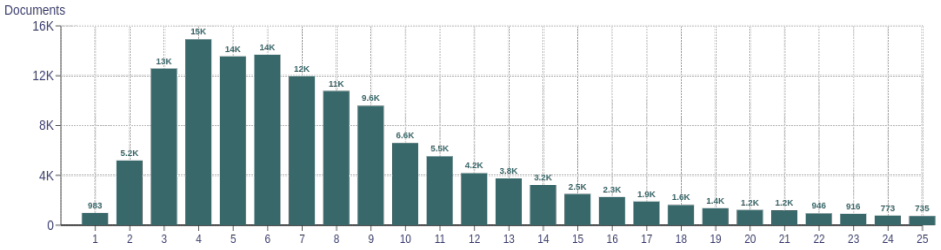
Domain	Docs	% of total
wikipedia.org	62K	42.60
rtr.ch	53K	36.38
gr.ch	2K	1.33
uslaval.it	1.4K	0.93
provinzia.bz.it	940	0.64
admin.ch	714	0.49
wikisource.org	702	0.48
noeles.info	612	0.42
playsuisse.ch	607	0.42
blogspot.com	603	0.41

Top 10 TLDs

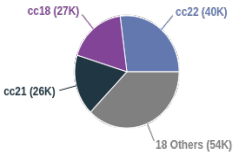
Domain	Docs	% of total
ch	69K	46.99
org	64K	44.02
it	5.1K	3.47
com	3.2K	2.19
bz.it	1.4K	0.99
info	935	0.64
net	606	0.41
is	551	0.38
de	157	0.11
eu	123	0.08

Documents size (in segments)

<= 25 segments **90.3%** (132K documents)
> 25 segments **9.7%** (14K documents)

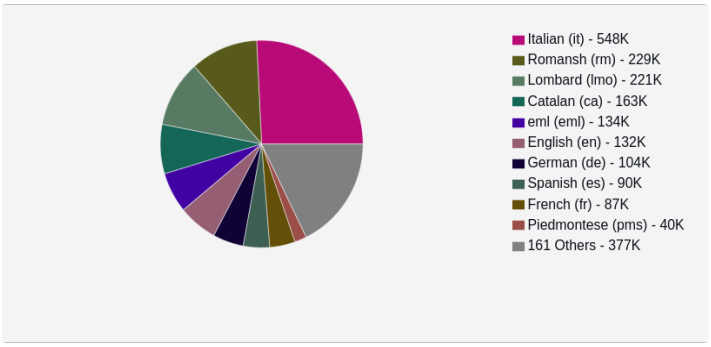


Documents by collection

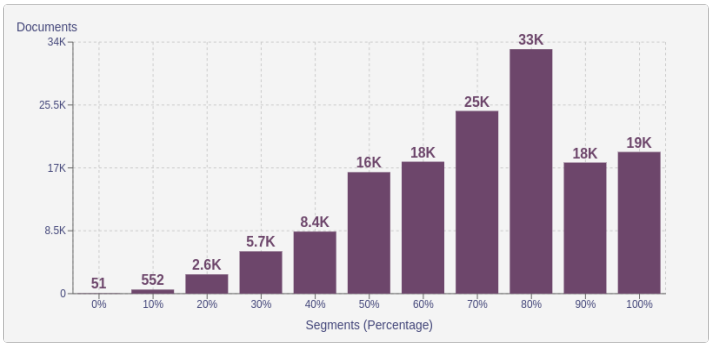


Language Distribution

Number of segments

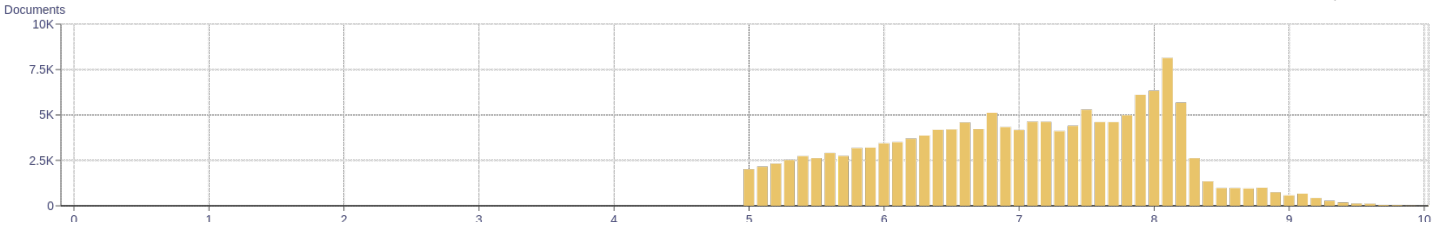


Percentage of segments in Lombard (lmo) inside documents



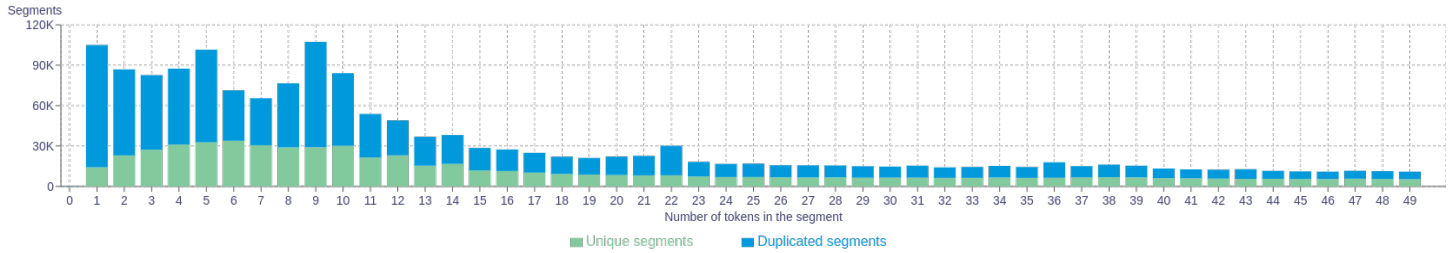
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (146K documents)

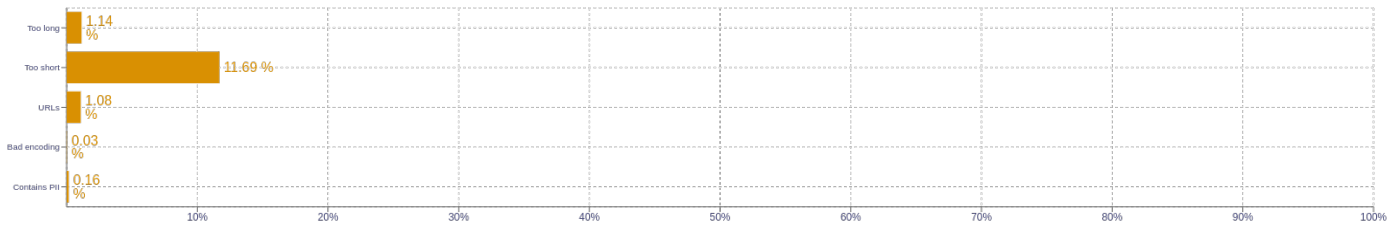


Segment length distribution by token

<= 49 tokens = **605K** segments | **1M** duplicates
> 50 tokens = **494K** segments | **257K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	l | 1801976 · il | 760646 · i | 626798 · en | 576121 · d | 456446
2	en il | 94573 · l sorgènt | 72932 · en l | 47163 · il code | 34275 · modifitgar il | 34059
3	modifitgar il code | 34058 · cunt cunt cunt | 14905 · œ œ œ | 11732 · sö la popolasiù | 8297 · statìstiche demogràfiche istat | 8297
4	cunt cunt cunt cunt | 14597 · œ œ œ œ | 9328 · statìstiche sö la popolasiù | 8297 · popolasiù del istitùto nasiunàl | 8297 · nasiunàl de statìstica relative | 8297
5	cunt cunt cunt cunt cunt | 14369 · sö la popolasiù del istitùto | 8297 · istitùto nasiunàl de statìstica relative | 8297 · œ œ œ œ œ | 7431 · tuca tuca tuca tuca tuca | 5392

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