HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
srd_Latn.jsonl.tsv	11/27/2024	Sardinian (sc)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
53,815	917,090	444,889 (48.51 %)	30M	147,885,449	143.62 MB

Top 10 domains

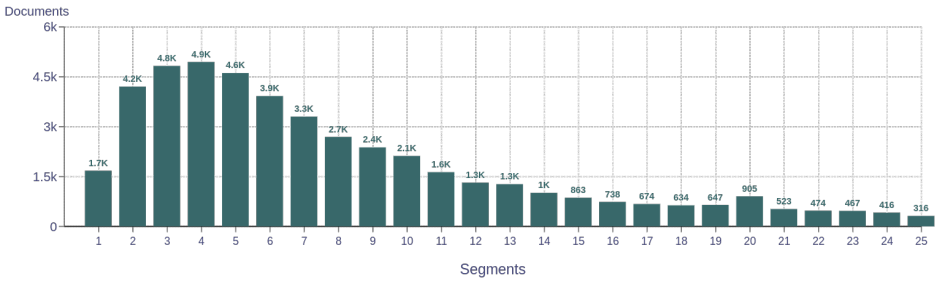
Domain	Docs	% of total
wikipedia.org	19K	34.38
sardegnacultura.it	2.6K	4.89
ilminuto.info	2.1K	3.84
nor-web.eu	1.7K	3.23
sagazeta.info	1.1K	2.03
blogspot.com	989	1.84
wordpress.com	788	1.46
reisar.eu	777	1.44
anthonymuroni.it	600	1.11
istorias.it	567	1.05

Top 10 TLDs

Domain	Docs	% of total
org	20K	37.54
it	13K	23.63
com	8.8K	16.27
info	3.5K	6.54
eu	3K	5.56
net	1.8K	3.41
or.it	497	0.92
ru	484	0.90
de	193	0.36
com.br	186	0.35

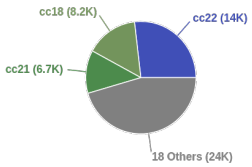
Documents size (in segments)

<= 25 segments **86.52%** (47K documents)
> 25 segments **13.48%** (7.3K documents)



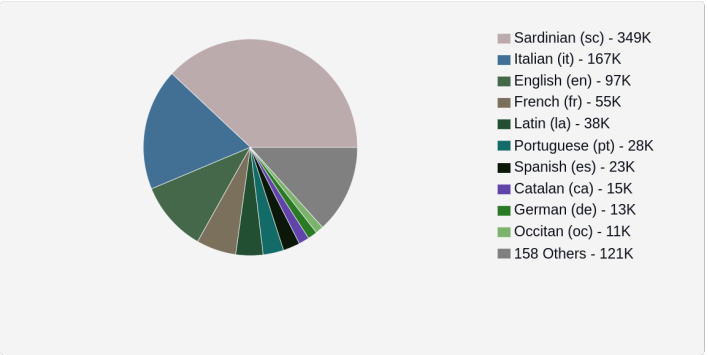
Documents by collection

CC = 68.64%
IA = 31.36%

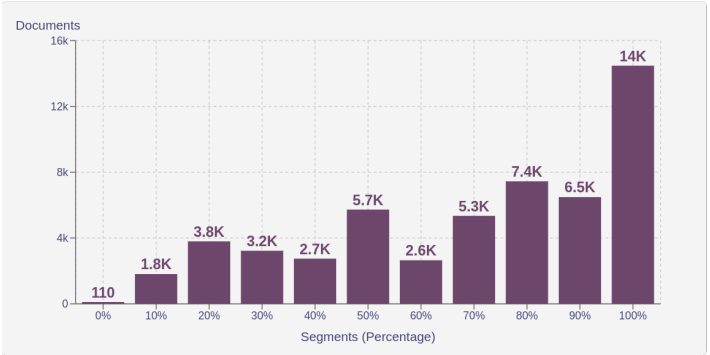


Language Distribution

Number of segments in the Sardinian (sc) corpus

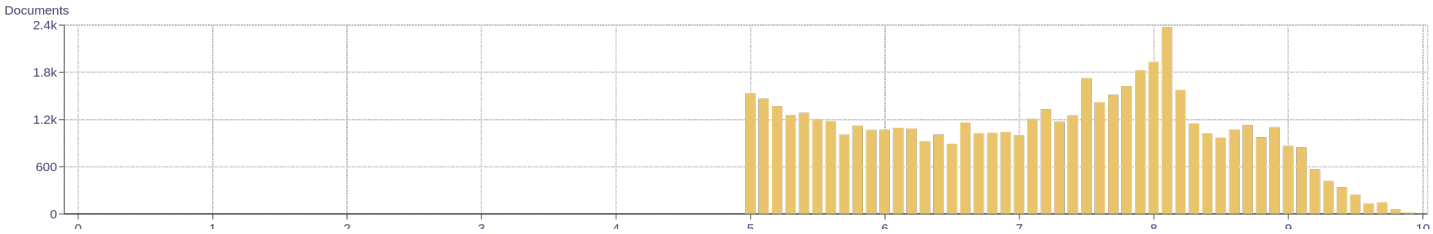


Percentage of segments in Sardinian (sc) inside documents



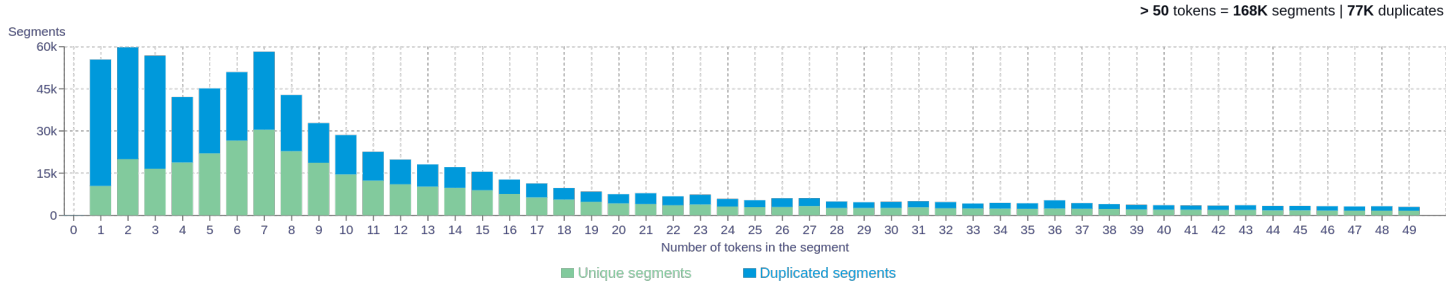
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (54K documents)

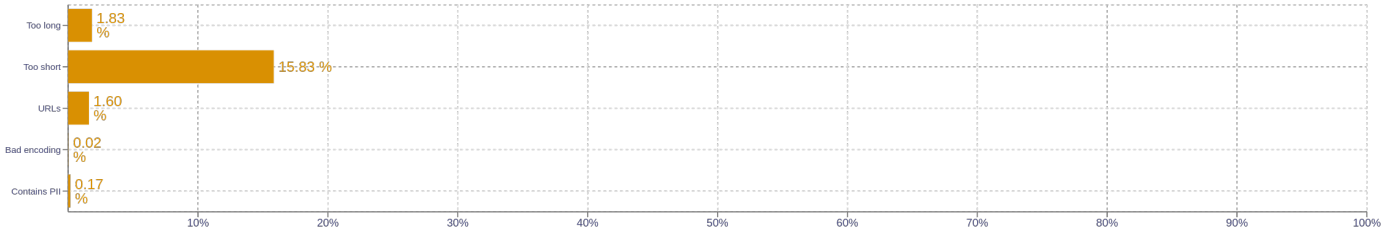


Segment length distribution by token

≤ 49 tokens = **354K** segments | **395K** duplicates
> 50 tokens = **168K** segments | **77K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	sa | 544830 · s | 392867 · est | 259401 · sos | 168252 · si | 162404
2	cun sa | 17141 · sa limba | 16984 · dae sa | 12222 · est sa | 11485 · est unu | 11280
3	còdighe de orìgine | 8703 · modìfica su còdighe | 8651 · sinònimos e contràrios | 8422 · scanu valerio scanu | 7468 · valerio scanu valerio | 7435
4	scanu valerio scanu valerio | 7428 · valerio scanu valerio scanu | 7361 · milano milano milano milano | 6866 · carmelo lisciotto carmelo lisciotto | 3505 · lisciotto carmelo lisciotto carmelo | 3460
5	modìfica su còdighe de orìgine | 8651 · scanu valerio scanu valerio scanu | 7354 · valerio scanu valerio scanu valerio | 7321 · milano milano milano milano milano | 6850 · lisciotto carmelo lisciotto carmelo lisciotto | 3460

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