HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
als_Latn.jsonl.tsv	9/23/2024	Albanian (als)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
5,385,262	95,101,632	48,574,292 (51.08 %)	3.2B	16.0 GB	16,005,838,206

Top 10 domains

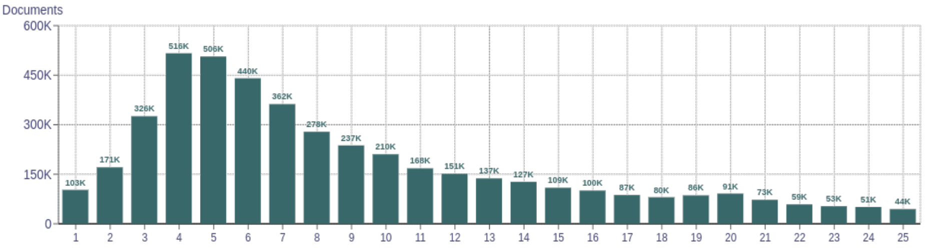
Domain	Docs	% of total
evropaelire.org	134K	2.48
zeriamerikes.com	110K	2.05
wikipedia.org	100K	1.86
botasot.info	71K	1.32
albeu.com	48K	0.90
blogspot.com	43K	0.80
teksteshqip.com	42K	0.78
telegrafi.com	41K	0.77
koha.net	37K	0.69
shqiperia.com	33K	0.61

Top 10 TLDs

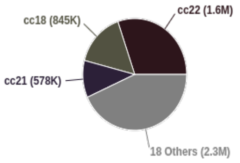
Domain	Docs	% of total
com	2.3M	41.79
al	1.1M	20.44
org	474K	8.81
net	411K	7.62
info	305K	5.66
mk	162K	3.02
tv	153K	2.84
gov.al	62K	1.15
com.al	54K	1.01
ch	48K	0.90

Documents size (in segments)

<= 25 segments **84.81%** (4.6M documents)
> 25 segments **15.19%** (818K documents)

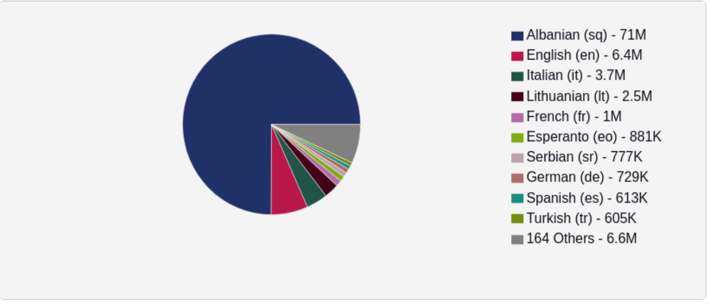


Documents by collection

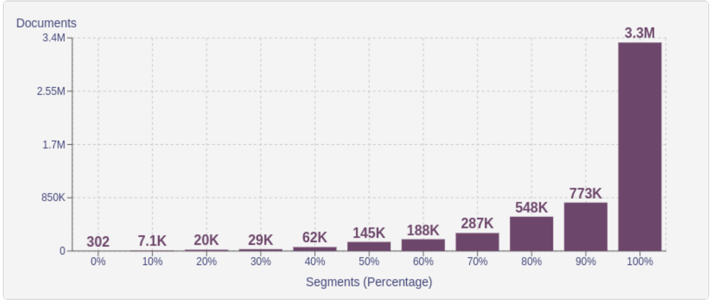


Language Distribution

Number of segments

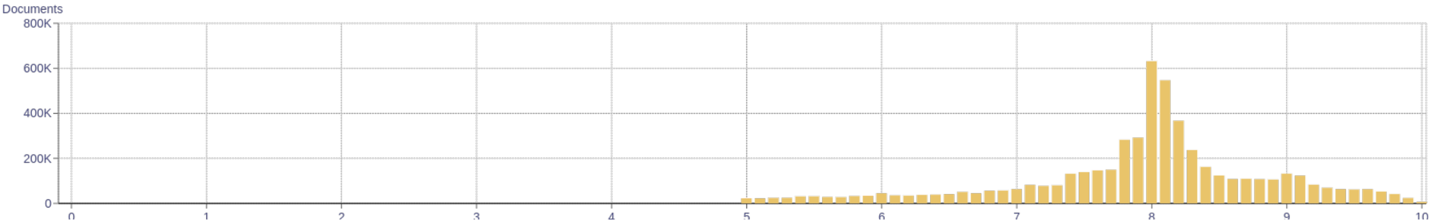


Percentage of segments in Albanian (als) inside documents



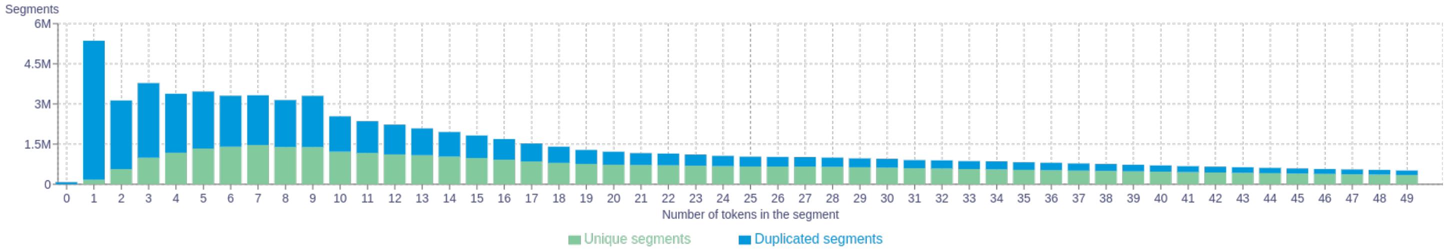
Distribution of documents by document score

score <= 5 - **99.87%** (5.4M documents)
score > 5 - **0.13%** (6.8K documents)

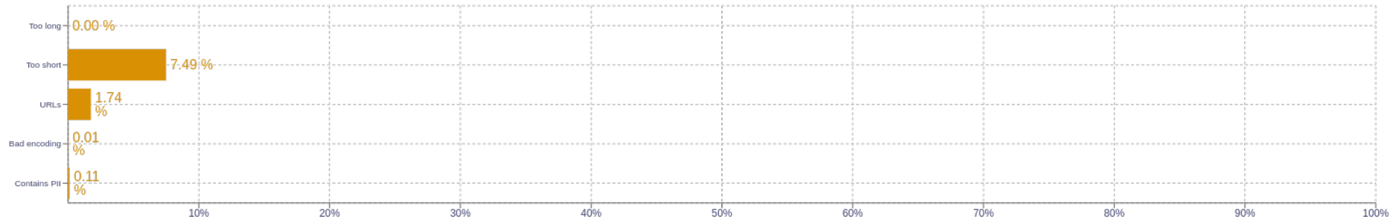


Segment length distribution by token

<= 49 tokens = **36M** segments | **40M** duplicates
> 50 tokens = **19M** segments | **6.5M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	është | 21759174 shumë | 8812607 kanë | 6473161 tė | 6048064 duhet | 5137758
2	është shumë | 404479 është bërë | 375825 kanë qenë | 364525 read more | 318634 edi rama | 293209
3	herë të parë | 274560 shtetet e bashkuara | 260659 redakto tekstin burimor | 237717 duhet të jetë | 225196 republikës së kosovës | 189623
4	gjithnjë e më shumë | 59240 luftës së dytë botërore | 57162 luaj online flash lojë | 47112 është shumë e rëndësishme | 38496 sot e kësaj dite | 32286
5	miqtë tuaj më të mirë | 68169 ndajnë këtë lojë me miqtë | 67888 harroni të vlerësoni këtë game | 55687 shtetet e bashkuara të amerikës | 54702 shteteve të bashkuara të amerikës | 45287

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