

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

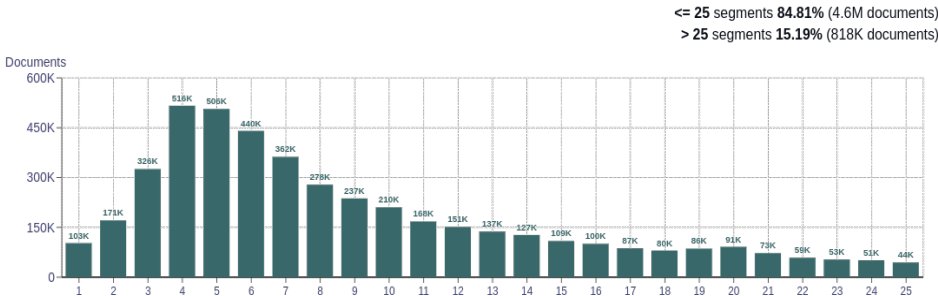
Top 10 domains

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

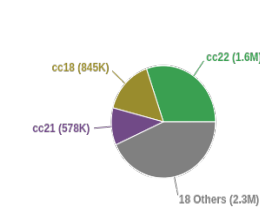
Top 10 TLDs

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)

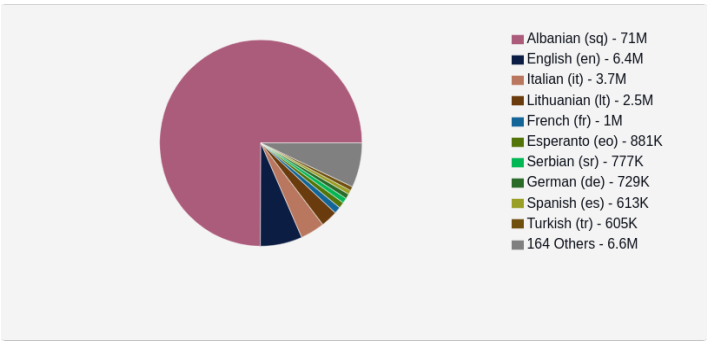


Documents by collection

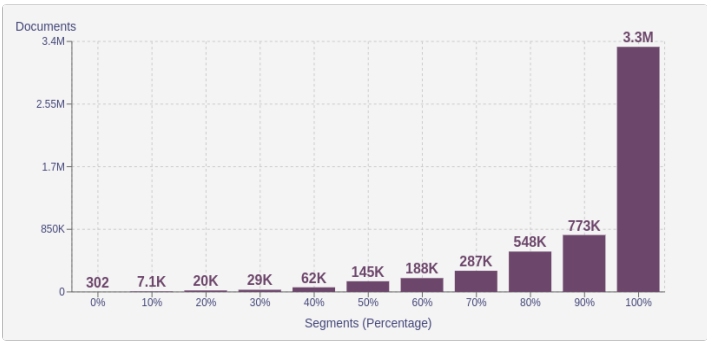


Language Distribution

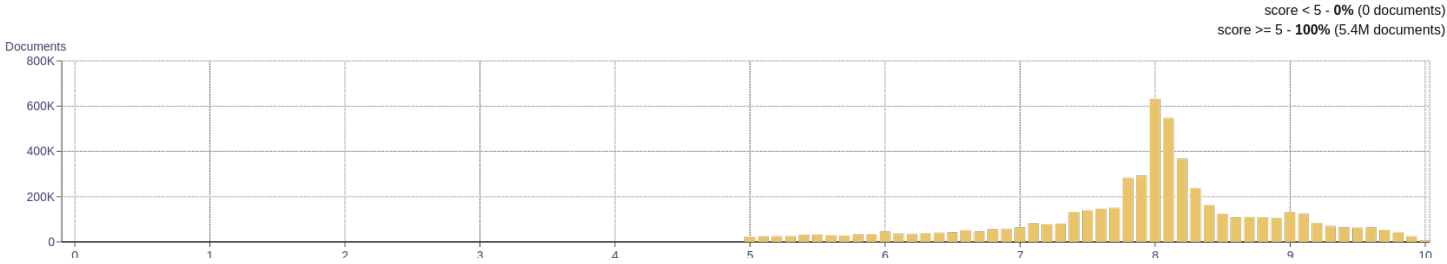
Number of segments



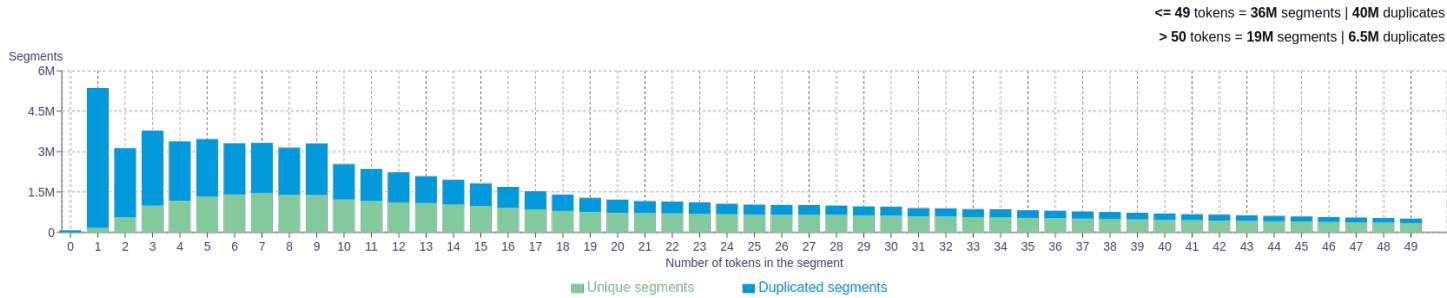
Percentage of segments in Albanian (als) inside documents



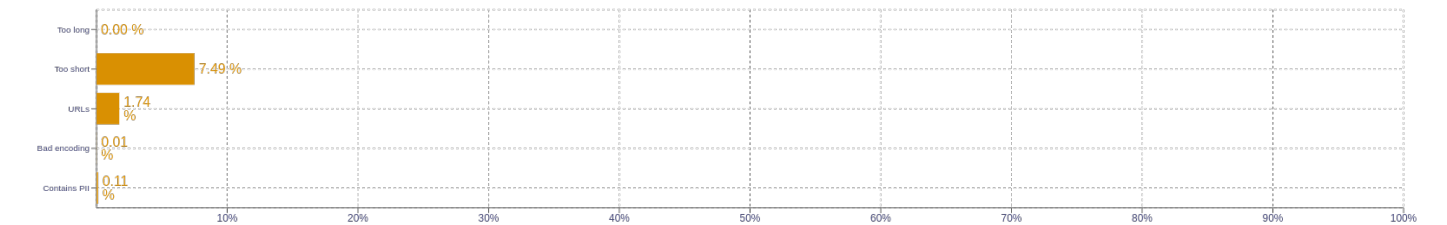
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

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

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 of 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>