HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
hy_1.jsonl.tsv	3/21/2024	Armenian (hy)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
621,465	67,013,428	13,749,789 (20.52 %)	794M	7.21 GB	

Top 10 domains

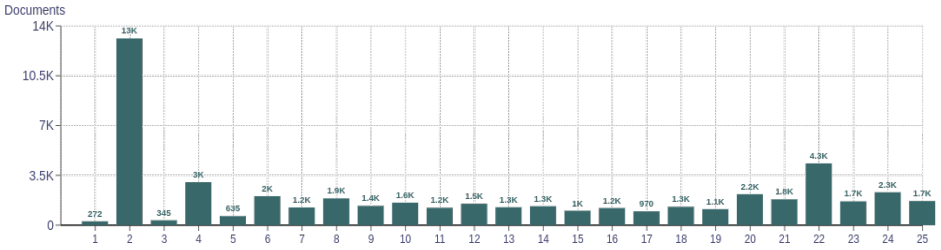
Domain	Docs	% of total
epress.am	49K	7.89
armlur.am	28K	4.46
aravot.am	21K	3.39
wikipedia.org	21K	3.38
blognews.am	13K	2.03
realnews.am	6.5K	1.05
mediamall.am	5.7K	0.92
haynews.am	4.8K	0.77
ingablog.ru	4.5K	0.73
armradio.am	4.2K	0.67

Top 10 TLDs

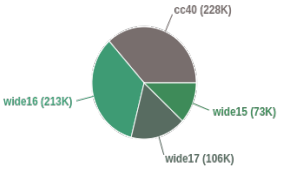
Domain	Docs	% of total
am	415K	66.84
com	82K	13.13
org	36K	5.80
ru	33K	5.39
info	16K	2.52
net	9.1K	1.47
blog	2.6K	0.42
news	2.6K	0.41
online	2.3K	0.37
today	1.3K	0.21

Documents size (in segments)

<= 25 segments **8.13%** (50K documents)
> 25 segments **91.87%** (570K documents)

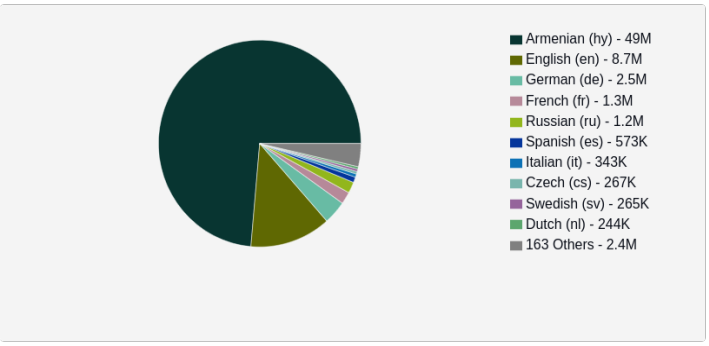


Documents by collection

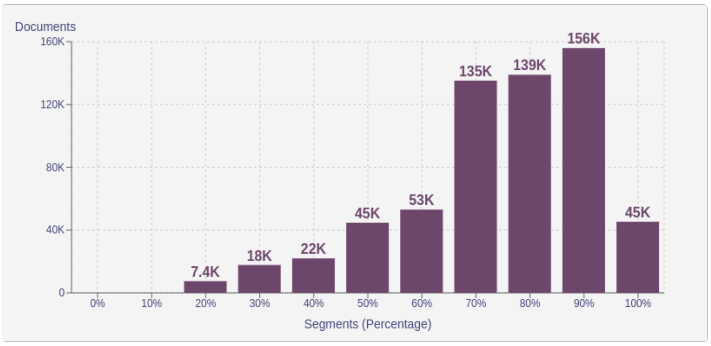


Language Distribution

Number of segments

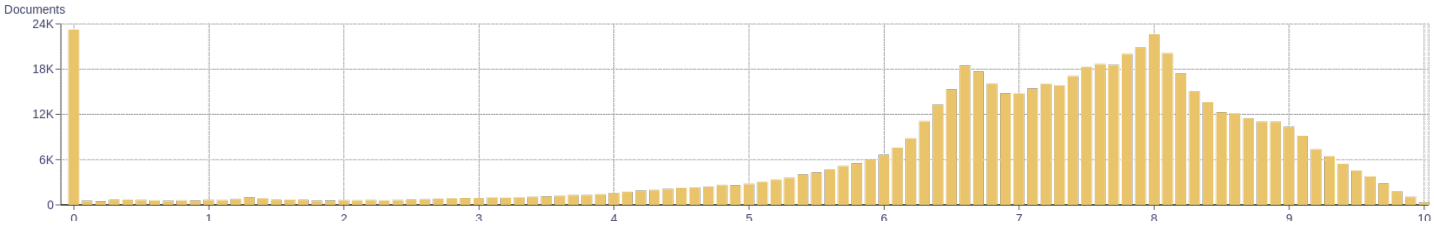


Percentage of segments in Armenian (hy) inside documents



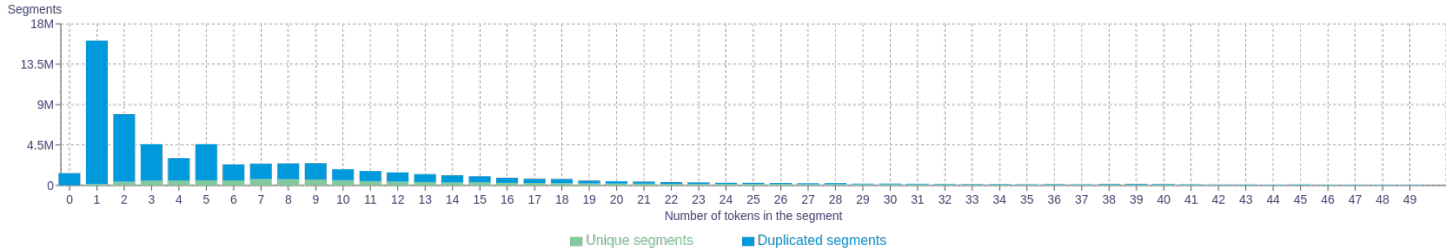
Distribution of documents by document score

score < 5 - **12.02%** (75K documents)
score >= 5 - **87.98%** (547K documents)

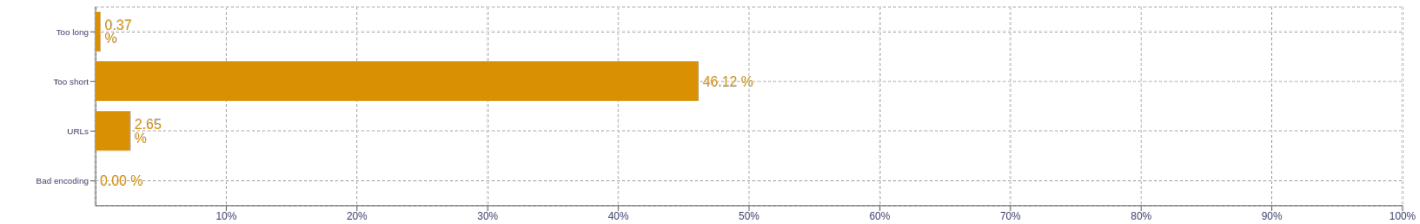


Segment length distribution by token

<= 49 tokens = **12M** segments | **53M** duplicates
> 50 tokens = **2.9M** segments | **750K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	մասին | 2076616 · եւ | 2055212 · մի | 1673183 · am | 1560627 · չի | 1435923
2	մի քանի | 246348 · մեր մասին | 224034 · հայաստանի հանրապետության | 220727 · իրավունքները պաշտպանված | 203535 · բոլոր իրավունքները | 200671
3	բոլոր իրավունքները պաշտպանված | 194283 · պատասխանատվություն չի կրում | 113511 · skip to content | 85556 · նրանց համար չեկավ | 83803 · զոհվել է ժ | 83790
4	նրանց համար չեկավ գարուն | 83803 · հովհաննիսյանը զոհվել է ժ | 83789 · am բոլոր իրավունքները պաշտպանված | 56777 · կայքը պատասխանատվություն չի կրում | 54338 · նյութերի ամբողջական կամ մասնակի | 50988
5	զաքար հովհաննիսյանը զոհվել է ժ | 83789 · կայքի նյութերի ամբողջական կամ մասնակի | 49039 · նյութերի ամբողջական կամ մասնակի օգտագործման | 48870 · ամբողջական կամ մասնակի օգտագործման դեպքում | 48870 · պատասխանատվություն չի կրում կայքում արտահայտված | 48778

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