HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
ka_1.jsonl.tsv	3/26/2024	Georgian (ka)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
533,070	65,524,284	14,489,425 (22.11 %)	769M	10.03 GB	

Top 10 domains

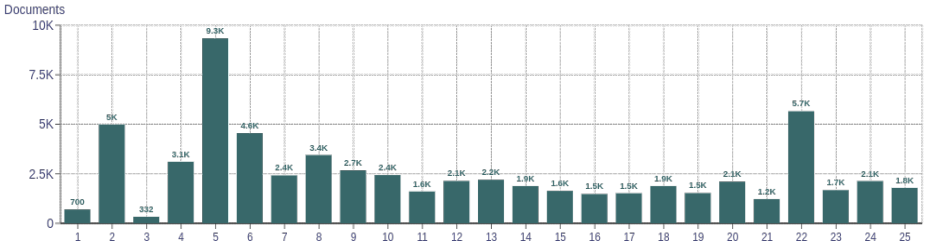
Domain	Docs	% of total
extra.ge	30K	5.67
forum.ge	12K	2.33
wikipedia.org	11K	2.14
chinashop.ge	8.4K	1.58
netgazeti.ge	8.2K	1.53
tabula.ge	7.2K	1.35
interpressnews.ge	5K	0.94
news.ge	4.6K	0.86
wordpress.com	4.5K	0.85
on.ge	4.4K	0.83

Top 10 TLDs

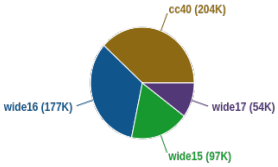
Domain	Docs	% of total
ge	365K	68.52
com	65K	12.19
org	29K	5.53
net	12K	2.19
gov.ge	7.6K	1.43
com.ge	7.1K	1.32
edu.ge	4.2K	0.79
ch	4K	0.75
org.ge	4K	0.75
info	2.8K	0.52

Documents size (in segments)

<= 25 segments **12.13%** (64K documents)
> 25 segments **87.87%** (467K documents)

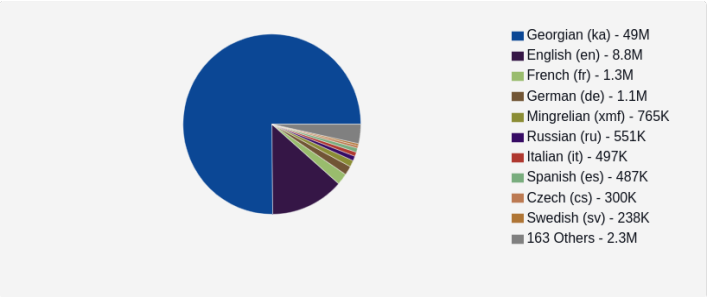


Documents by collection

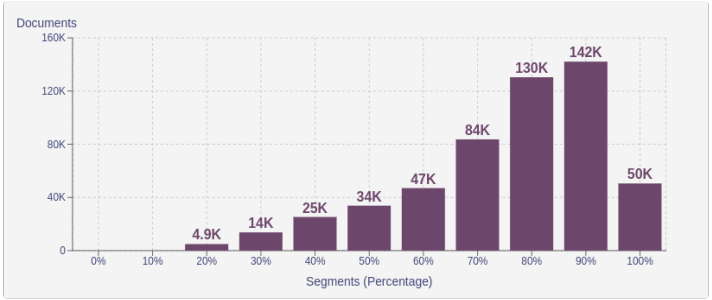


Language Distribution

Number of segments

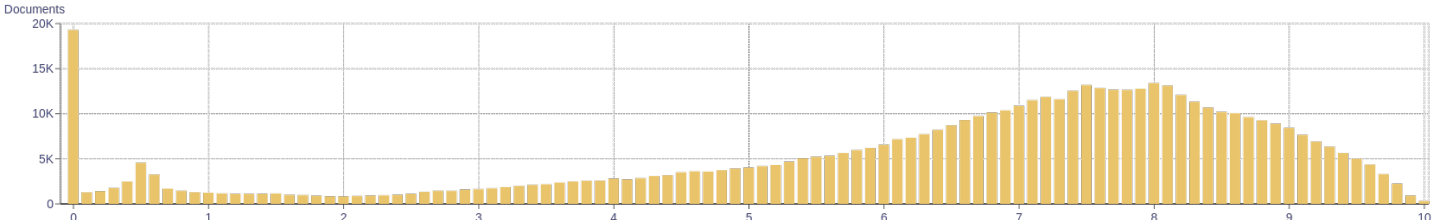


Percentage of segments in Georgian (ka) inside documents



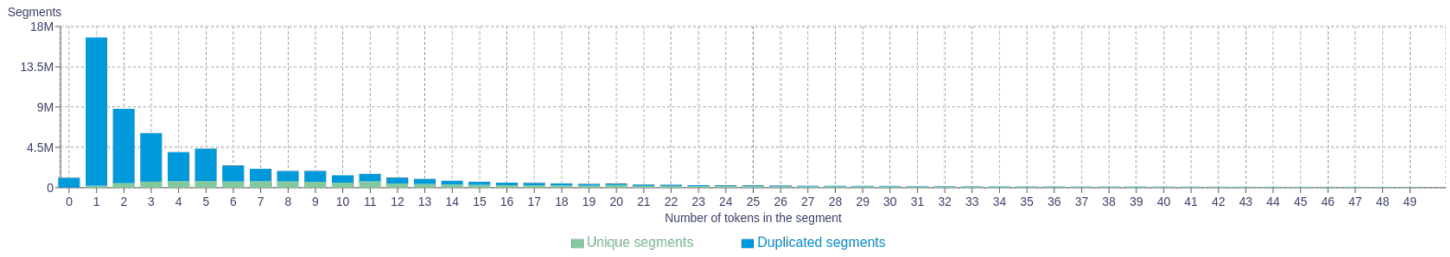
Distribution of documents by document score

score <= 5 - **21.58%** (115K documents)
score > 5 - **78.42%** (418K documents)

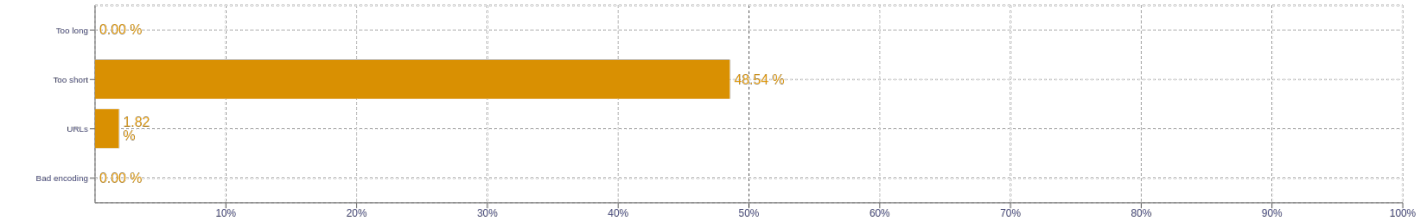


Segment length distribution by token

<= 49 tokens = **13M** segments | **50M** duplicates
> 50 tokens = **2.7M** segments | **913K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	ის | 1799198 · ეს | 1790348 · ამ | 1727935 · არის | 1460367 · the | 1280215
2	posted by | 341878 · span style | 244391 · ახალი ამბები | 211577 · of the | 176848 · ნაღდი ანგარიშსწორების | 158773
3	თბილისის მასშტაბით სრულიად | 97511 · ნაღდი ანგარიშსწორების სურვილის | 96448 · ანგარიშსწორების სურვილის შემთხვევაში | 96448 · შეავსეთ მარტივი ფორმა | 96306 · ღიაკს და შეავსეთ | 96102
4	ნაღდი ანგარიშსწორების სურვილის შემთხვევაში | 96448 · ღიაკს და შეავსეთ მარტივი | 96099 · ჩვენი კურიერი ადგილზე მოგაწვდით | 79409 · კურიერი ადგილზე მოგაწვდით პროდუქციას | 79409 · მიწოდება თბილისის მასშტაბით სრულიად | 69051
5	ღიაკს და შეავსეთ მარტივი ფორმა | 96099 · ჩვენი კურიერი ადგილზე მოგაწვდით პროდუქციას | 79409 · მიწოდება თბილისის მასშტაბით სრულიად უფასოა | 69051 · თამაში და გართობა არასდროს სრულდება | 62522 · დააკლიკე და გადახვალ სათამაშოების საუკეთესო | 62522

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