

General overview

Corpus	Analytics date	Language
kat_Georgsonl.tsv	9/19/2024	Georgian (ka)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
3,335,164	63,722,098	29,708,949 (46.62 %)	1.6B	24.7 GB	10,095,225,458

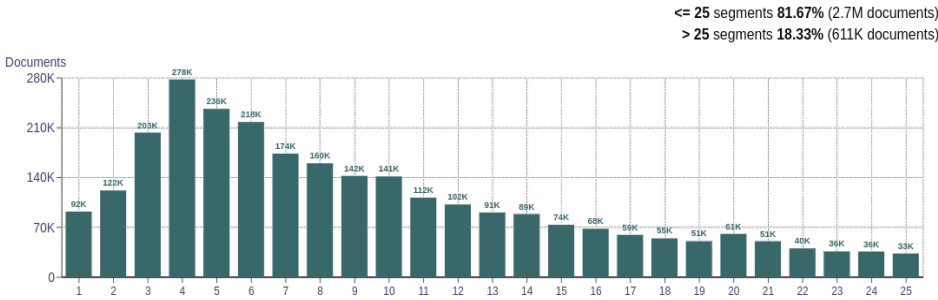
Top 10 domains

Domain	Docs	% of total
wikipedia.org	248K	7.44
radiotavisupleba.ge	141K	4.24
gancxadebebi.ge	96K	2.89
wordpress.com	69K	2.05
netgazeti.ge	34K	1.03
1tv.ge	28K	0.83
on.ge	27K	0.82
sana.ge	26K	0.78
blogspot.com	26K	0.78
ambebi.ge	26K	0.77

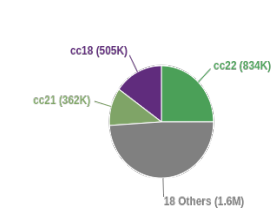
Top 10 TLDs

Domain	Docs	% of total
ge	2.1M	63.33
com	495K	14.85
org	313K	9.38
gov.ge	111K	3.32
net	68K	2.03
edu.ge	51K	1.53
com.ge	34K	1.03
org.ge	21K	0.63
am	16K	0.49
info	15K	0.46

Documents size (in segments)

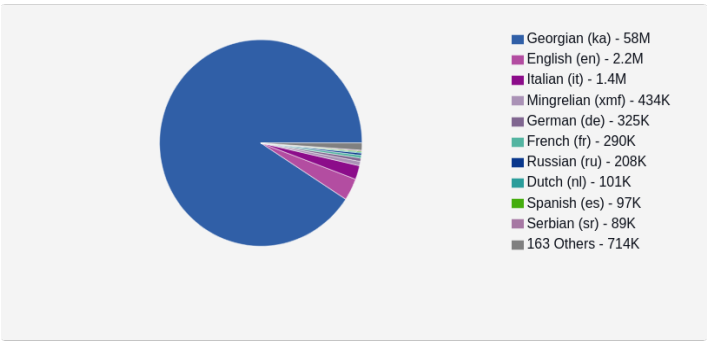


Documents by collection

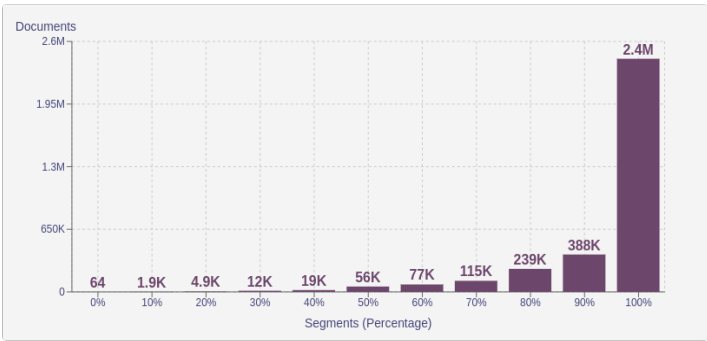


Language Distribution

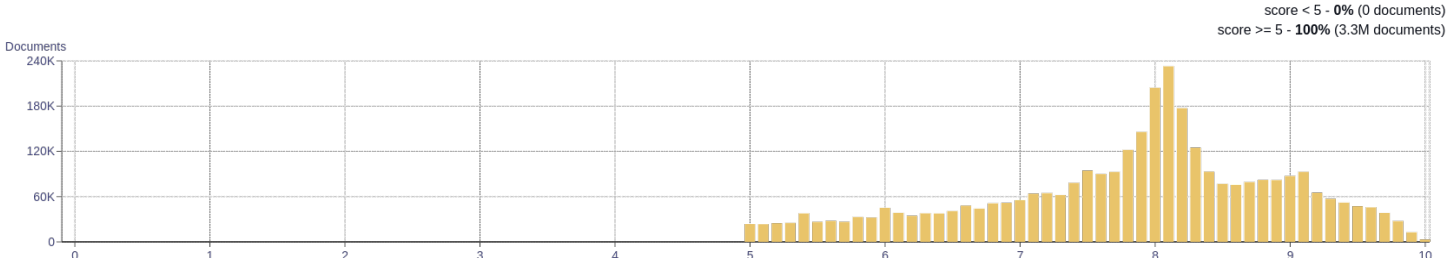
Number of segments



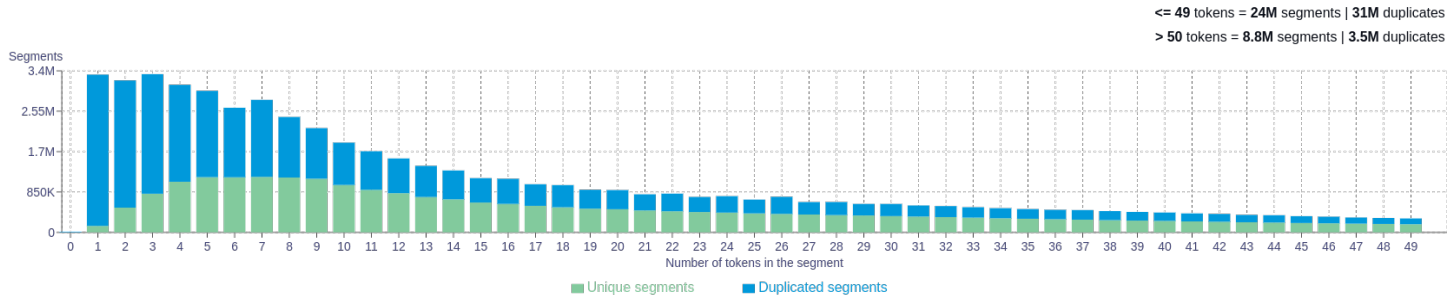
Percentage of segments in Georgian (ka) inside documents



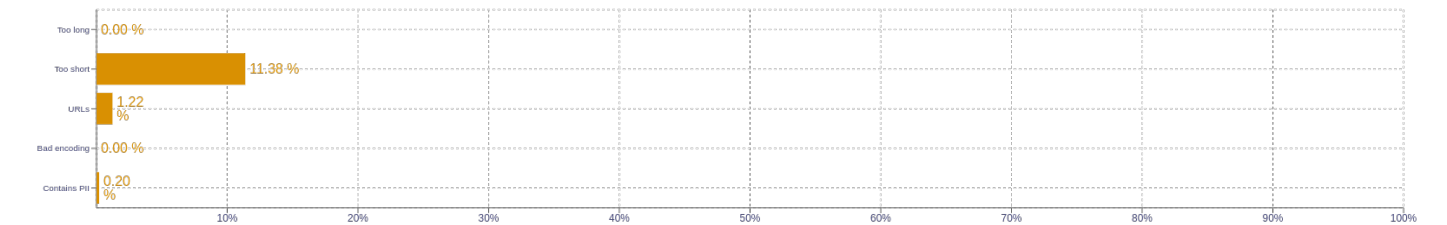
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>ეს 5971888</div> <div>ამ 5686353</div> <div>ის 5999032</div> <div>არის 4570049</div> <div>იყო 3701223</div>
2	<div>ეს არის 543370</div> <div>წყაროს რედაქტირება 355269</div> <div>რა ჟღერა 306579</div> <div>მიუხედავად იმისა 242247</div> <div>წლის განმავლობაში 208765</div>
3	<div>განათლებლისა და მენეჯერების 61610</div> <div>ამა თუ იმ 61200</div> <div>ეს არ არის 60572</div> <div>ფრანგი და ქართველი 56155</div> <div>ბაზარზე იყო ორიენტირებული 56068</div>
4	<div>კინეტის განვითარებისთვის საქართველოში რამდენიმე 56058</div> <div>ძირითადად ფრანგულ ბაზარზე იყო 56058</div> <div>ფრანგულ ბაზარზე იყო ორიენტირებული 56058</div> <div>საქართველოში რამდენიმე არაკომერციული პროექტიც 56058</div> <div>კომპანია ძირითადად ფრანგულ ბაზარზე 56058</div>
5	<div>კინეტის განვითარებისთვის საქართველოში რამდენიმე არაკომერციული 56058</div> <div>ძირითადად ფრანგულ ბაზარზე იყო ორიენტირებული 56058</div> <div>კომპანია ძირითადად ფრანგულ ბაზარზე იყო 56058</div> <div>განვითარებისთვის საქართველოში რამდენიმე არაკომერციული პროექტიც 56058</div> <div>წელს ფრანგი და ქართველი ღამფენებლების 56052</div>

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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>, , , 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>