

General overview

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

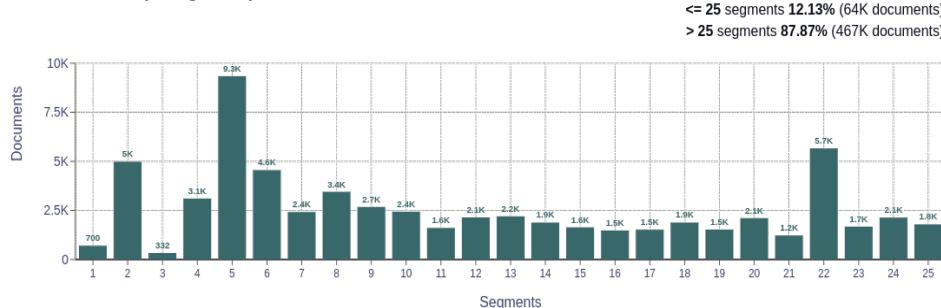
Volumes

Docs	Segments	Unique segments	Tokens	Size
533,070	65,524,284	53,749 (0.08 %)	769M	10.03 GB

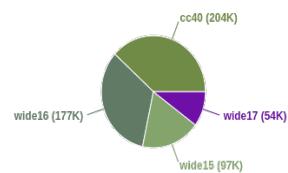
Type-Token Ratio

Georgian (ka)
0.01

Documents size (in segments)

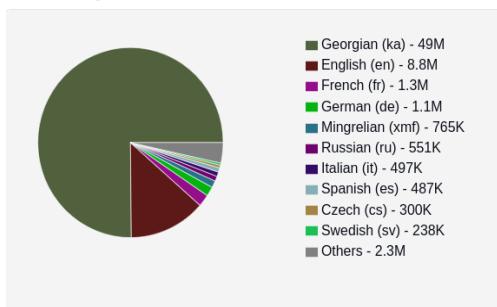


Documents by collection

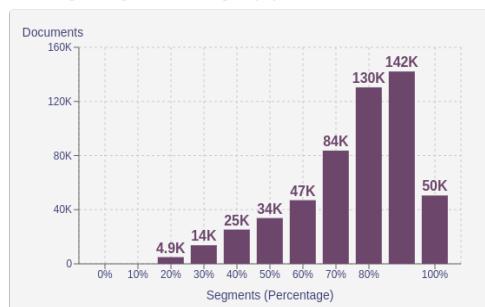


Language Distribution

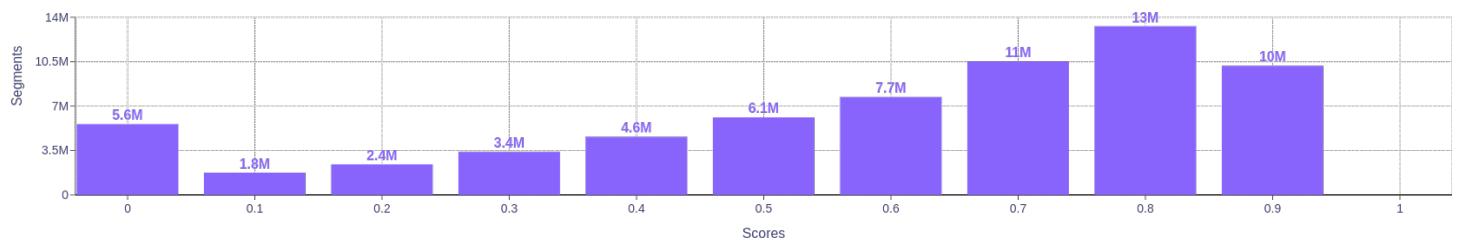
Number of segments



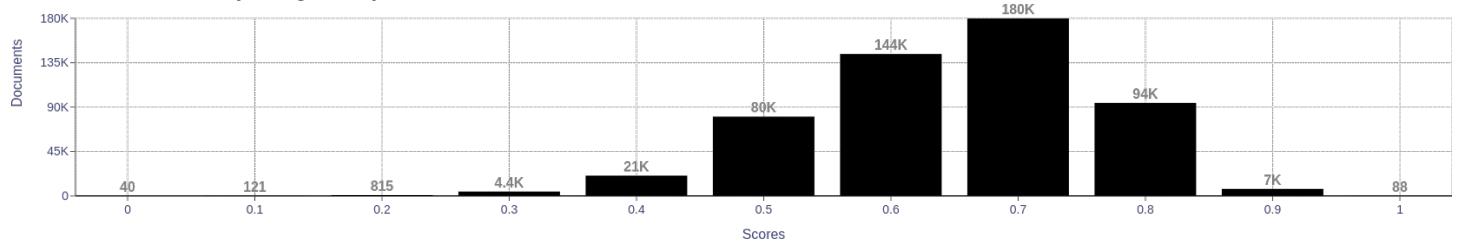
Percentage of segments in Georgian (ka) inside documents



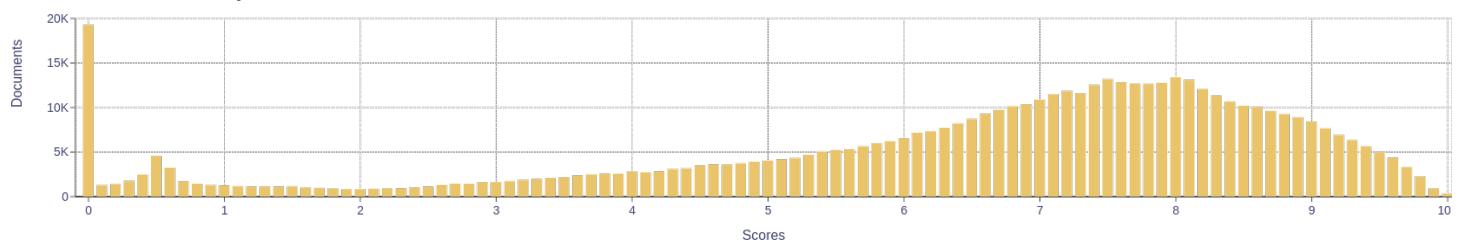
Distribution of segments by fluency score



Distribution of documents by average fluency score

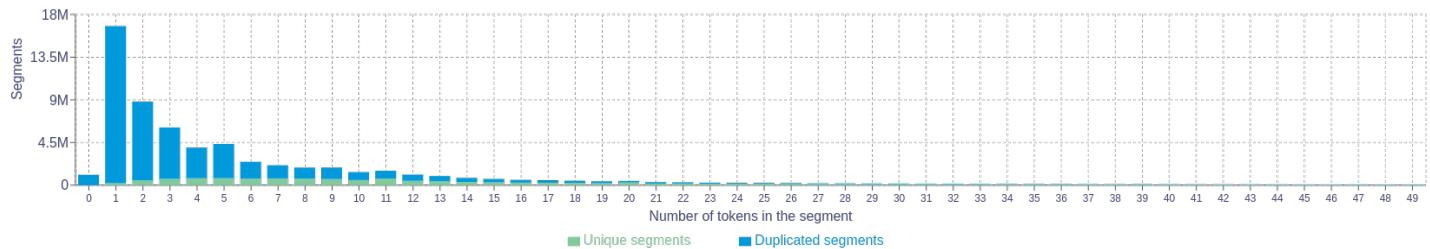


Distribution of documents by document score

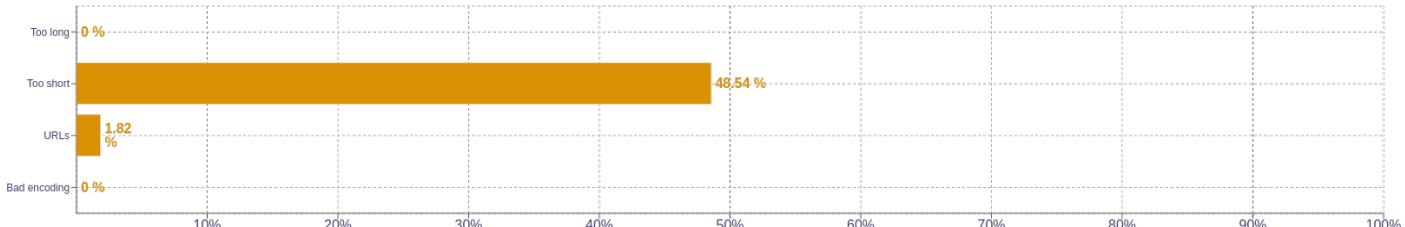


Segment length distribution by token

<= 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 ახადი ამბები 211577 of the 176848 ნაღდი ანგარიშსწორების 158773
3	თბილისი მასტერაბონ სერებრია 97511 ნაღდი ანგარიშსწორების სურვილის 96448 ანგარიშსწორების სურვილის შემთხვევაში 96448 შეავსეთ მარტივი ფორმა 96306 ლილას და შეავსეთ 96102
4	ნაღდი ანგარიშსწორების სურვილის შემთხვევაში 96448 ლილას და შეავსეთ მარტივი 96099 ჩვენი კურიერი აღიიღებ მოგანველი 79409 კურიერი აღიიღებ მოგანველი პროფესიას 79409 მინილება ასიგინს მასტერაბონ სრულია 69051
5	ლილას და შეავსეთ მარტივი ფორმა 96099 ჩვენი კურიერი აღიიღებ მოგანველი პროფესიას 79409 მინილება ასიგინს მასტერაბონ სრულია უფასო 69051 თამაში და გარიბობა არასურის სრულება 62522 დაღარიელ და გაღადახად საათამაშიერის საკუთხევლი 62522

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 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>`, ``, ``, etc.) replaced by newlines.

Language distribution

Language distribution

Distribution of segments by fluency score

Distribution of segments by fluency score

Obtained with Monocle4er (<https://github.com/bt>)

Distribution of documents by average fluency score

Obtained with Monocleanner (<https://github.com/monocleanner>)

Distribution of documents by document score

Obtained with Web Docs Sc

Segment length distribution by token

Tokenized with http

Segment noise distribution