HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
so_1.jsonl.tsv	3/17/2024	Somali (so)

Volumes

Docs	Segments	Unique segments	Tokens	Size
283,712	23,606,211	22,713 (0.10 %)	249M	1.32 GB

Type-Token Ratio

Somali (so)
0.01

Top 10 domains

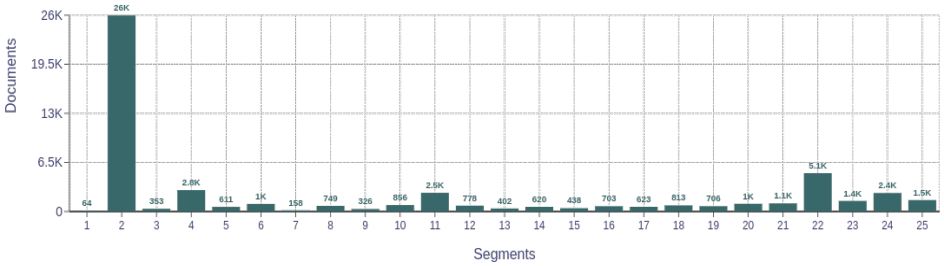
Domain	Docs	% of total
puntlandi.com	8K	2.82
allgalgaduud.com	5.5K	1.94
caasimada.net	4.3K	1.52
somalitalk.com	4K	1.42
almisexpress.com	4K	1.39
cakaaranews.com	3.3K	1.15
radiodalsan.com	3.1K	1.09
allkisima.com	2.9K	1.02
goolfm.net	2.4K	0.84
laashin.com	2.4K	0.83

Top 10 TLDs

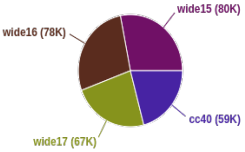
Domain	Docs	% of total
com	218K	76.90
net	40K	14.10
org	7.5K	2.63
so	3.7K	1.30
ca	1.9K	0.66
se	1.3K	0.47
online	1.1K	0.40
no	975	0.34
co.uk	881	0.31
info	676	0.24

Documents size (in segments)

<= 25 segments **18.69%** (53K documents)
> 25 segments **81.31%** (231K documents)

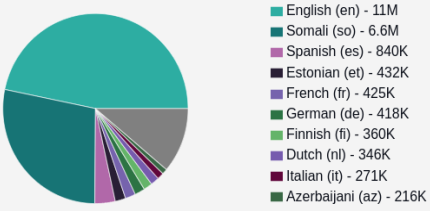


Documents by collection

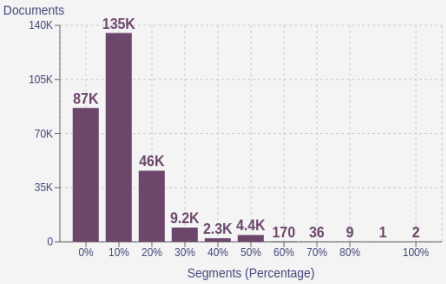


Language Distribution

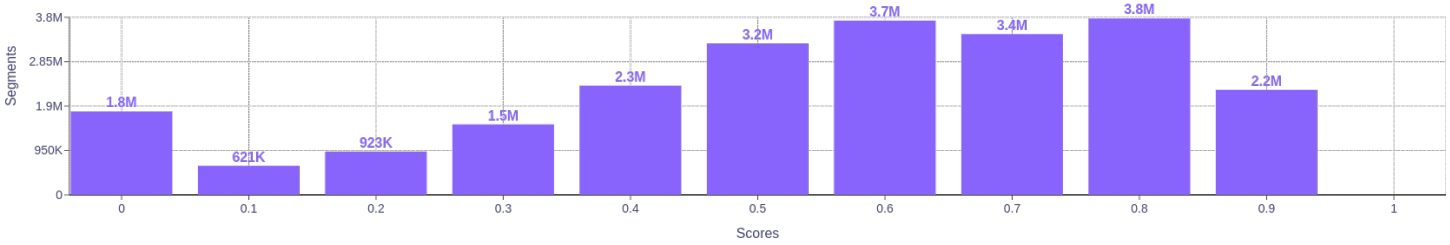
Number of segments



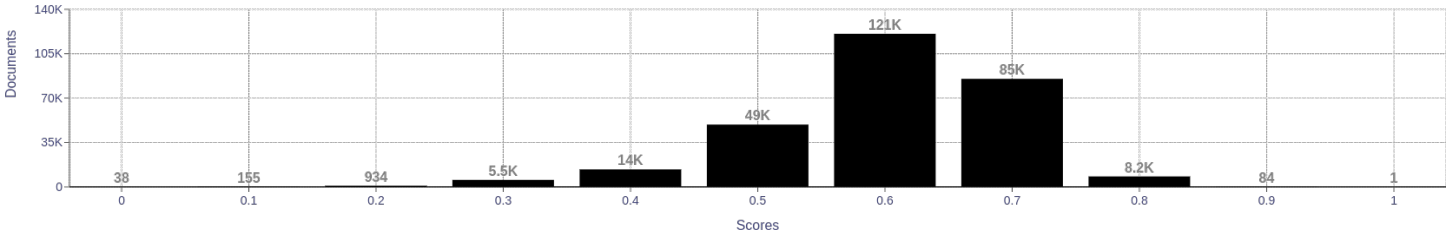
Percentage of segments in Somali (so) inside documents



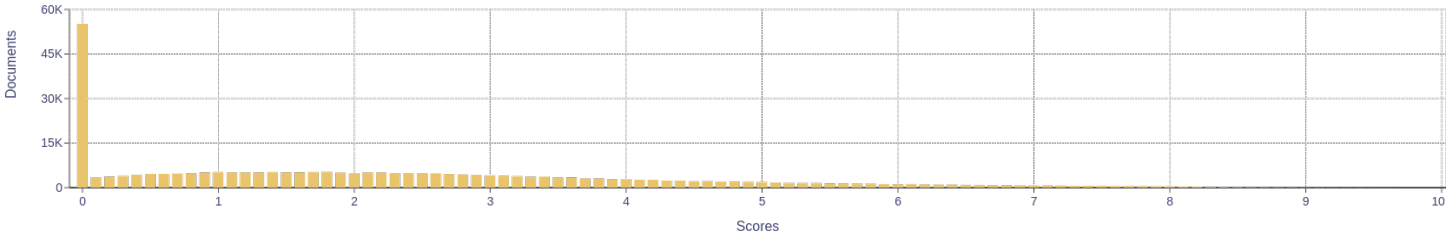
Distribution of segments by fluency score



Distribution of documents by average fluency score

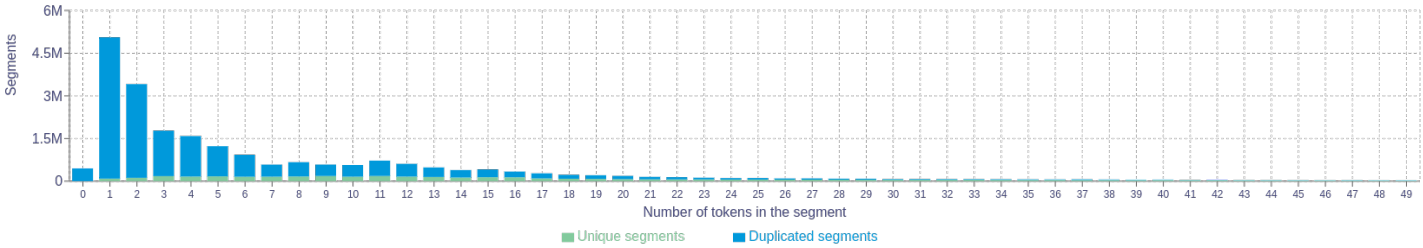


Distribution of documents by document score

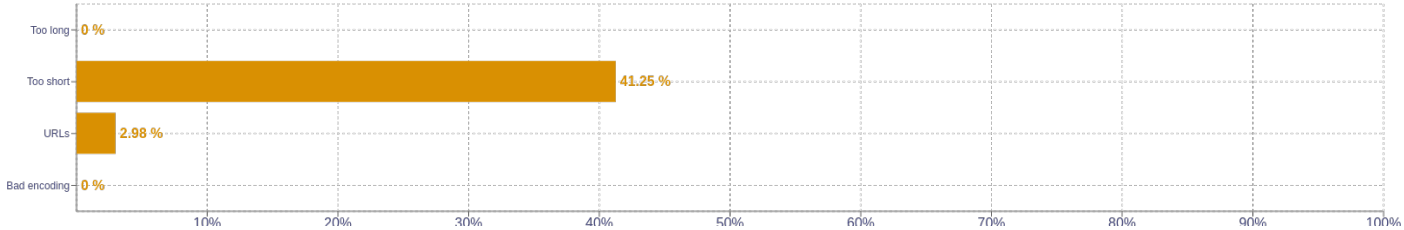


Segment length distribution by token

<= 49 tokens = **4.1M** segments | **19M** duplicates
> 50 tokens = **796K** segments | **183K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	iyo | 2763688 · ee | 2524139 · ah | 2253410 · u | 1953608 · la | 1517808
2	read more | 186480 · ah ee | 180934 · mid ah | 163683 · contact us | 149776 · of the | 149575
3	all rights reserved | 133041 · opens in new | 102073 · click to share | 91413 · to share on | 91405 · news in english | 83799
4	opens in new window | 102059 · click to share on | 91401 · log into your account | 33422 · leave a reply cancel | 30785 · a reply cancel reply | 30576
5	leave a reply cancel reply | 30569 · of new posts by email | 24673 · click to share on twitter | 24433 · notify me of new posts | 24417 · me of new posts by | 24266

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