HPLT Analytics report

General overview

Corpus	Date	Language
som_Latn.jsonl.tsv	9/23/2024	Somali (so)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
966,507	16,384,689	8,400,746 (51.27 %)	434M	2,549,211,220	2.39 GB

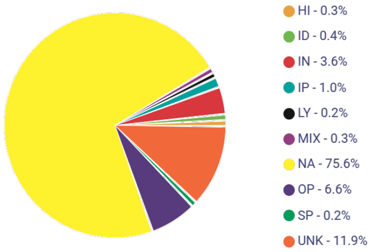
Top 10 domains

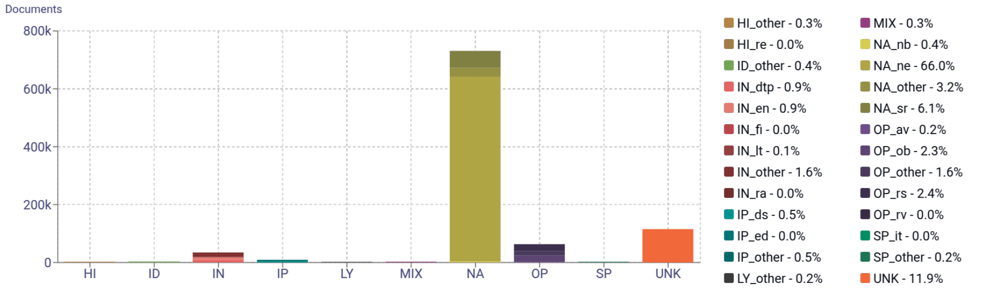
Domain	Docs	% of total
somalitalk.com	36K	3.70%
voasomali.com	17K	1.78%
caasimada.net	14K	1.41%
dunidaonline.com	13K	1.39%
somaliland.org	13K	1.34%
puntlandi.com	13K	1.31%
goobjoog.com	12K	1.27%
wikipedia.org	12K	1.21%
wordpress.com	11K	1.18%
goolfm.net	11K	1.13%

Top 10 TLDs

Domain	Docs	% of total
com	699K	72.28%
net	154K	15.93%
org	52K	5.34%
so	16K	1.70%
se	6.2K	0.65%
online	5.5K	0.57%
ca	4.2K	0.43%
info	3.4K	0.35%
is	2.8K	0.29%
fi	2.2K	0.23%

Register labels

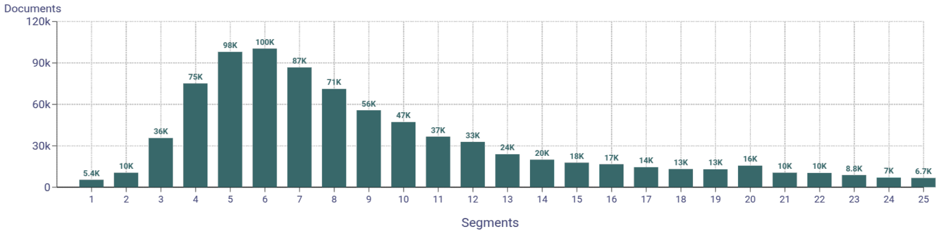




🤖 **MT**:6.0% | 58K Documents

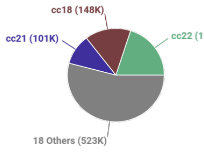
Documents size (in segments)

<= 25 segments **86.18%** (833K documents)
> 25 segments **13.82%** (134K documents)



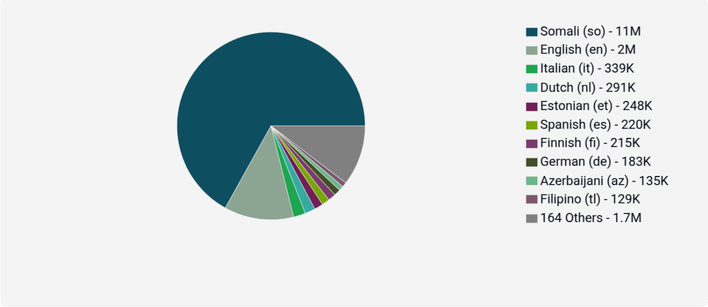
Documents by collection

CC = 56.73%
IA = 43.27%

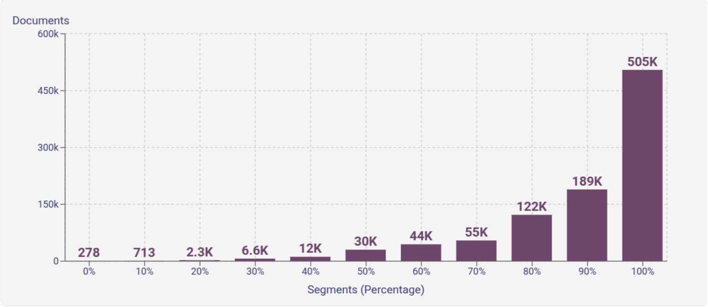


Language Distribution

Number of segments in the Somali (so) corpus

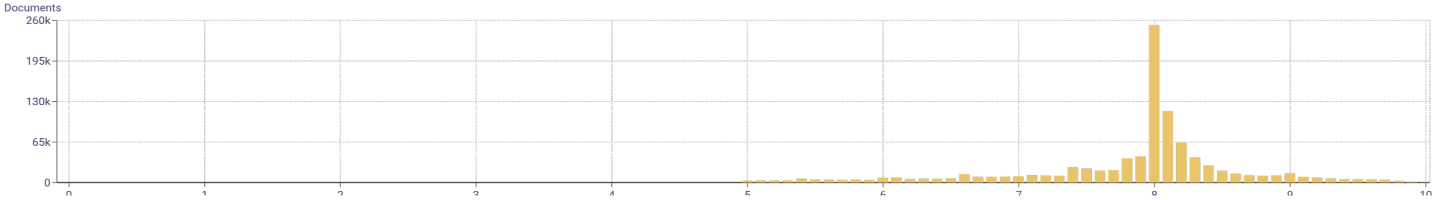


Percentage of segments in Somali (so) inside documents



Distribution of documents by document score

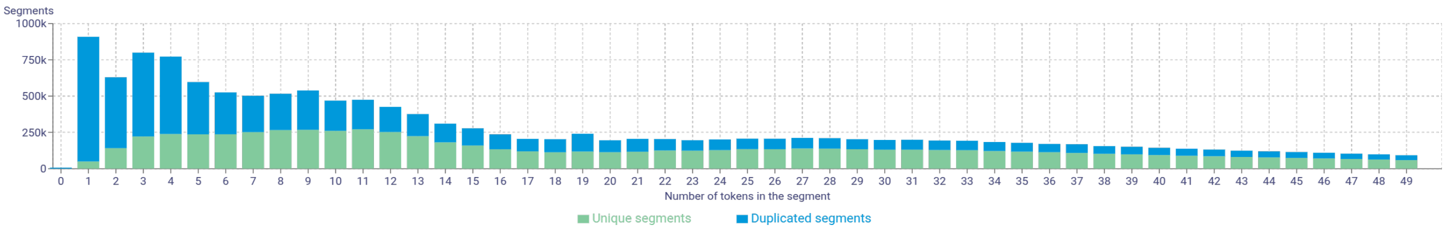
score < 5 - **0%** (0 documents)
score >= 5 - **100%** (967K documents)



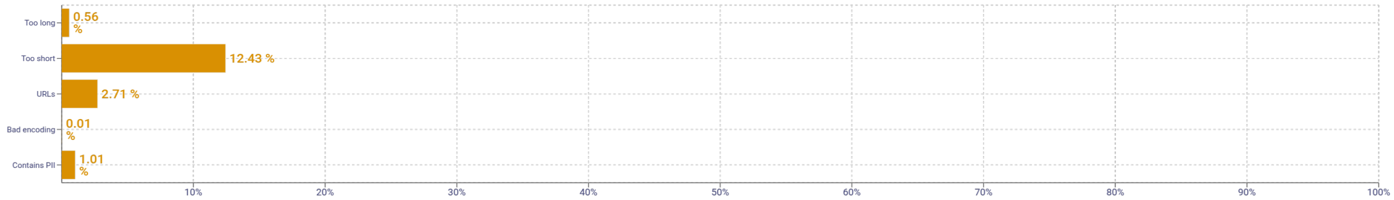
Segment length distribution by token

≤ 49 tokens = **7M** segments | **7M** duplicates

> 50 tokens = **2.4M** segments | **940K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	iyo | 7052255 ah | 6346633 ee | 6345824 u | 5569995 la | 4116900
2	mid ah | 577476 ah ee | 546674 kala duwan | 254500 magaalada muqdisho | 206322 ee dalka | 197778
3	qaar ka mid | 93434 mid ka mid | 88474 waxaa ka mid | 63976 kala duwan ee | 62056 ee magaalada muqdisho | 46958
4	qaar ka mid ah | 88647 mid ka mid ah | 82206 waxaa ka mid ah | 45806 reer binu israa 'iil | 24524 ah oo ku saabsan | 17948
5	badan oo ka mid ah | 11966 madaxweynaha jamhuuriyadda federaalka soomaaliya mudane | 8169 soomaaliya mudane xasan sheekh maxamuud | 7434 madaxweynaha soomaaliya xasan sheekh maxamuud | 6727 sida uu hadalka u dhigay | 6500

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

Register labels

Name	Abbr.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	**HI**
Recipe	re
Informational persuasion	**IP**
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	**IN**
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	**OP**
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av