

General overview

Corpus	Analytics date	Language
smo_Latn.jsonl.tsv	11/5/2024	Samoan (sm)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
45,856	1,012,457	733,868 (72.48 %)	44M	180.02 MB	185,180,583

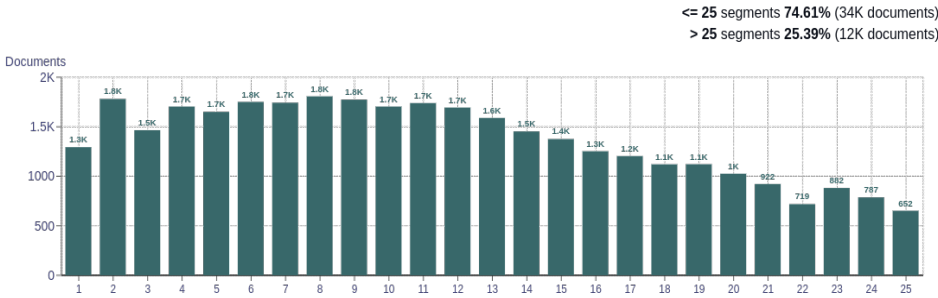
Top 10 domains

Domain	Docs	% of total
samoanews.com	2.6K	5.76
jw.org	2.2K	4.80
samoabserver.ws	1.7K	3.70
wikipedia.org	1.2K	2.66
martech.zone	1K	2.28
sobserver.ws	900	1.96
eturbonews.com	828	1.81
samoatimes.co.nz	742	1.62
recursosdeautoayuda.com	734	1.60
lds.org	725	1.58

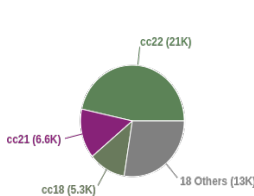
Top 10 TLDs

Domain	Docs	% of total
com	30K	65.54
org	6.5K	14.14
ws	2.7K	5.84
co.nz	1.2K	2.68
zone	1K	2.28
net	848	1.85
ru	313	0.68
es	258	0.56
govt.nz	215	0.47
network	181	0.39

Documents size (in segments)

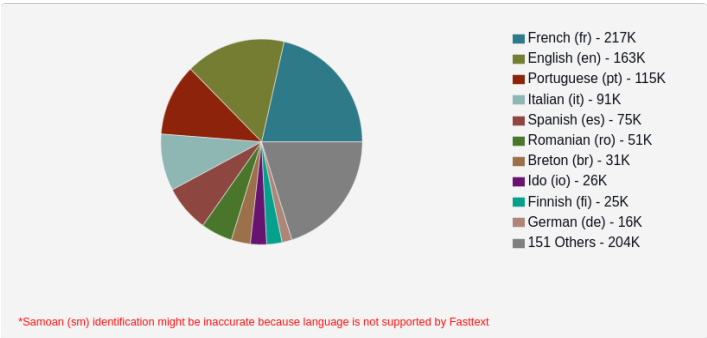


Documents by collection

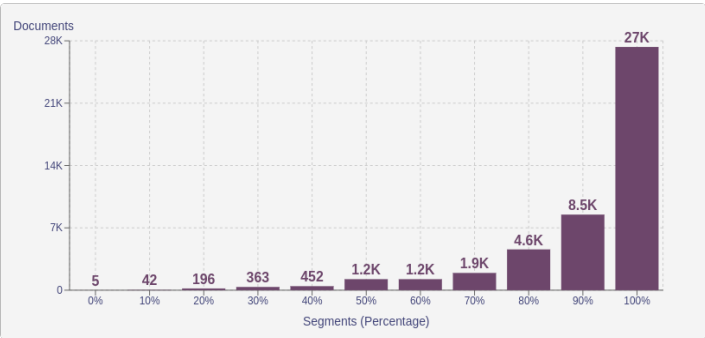


Language Distribution

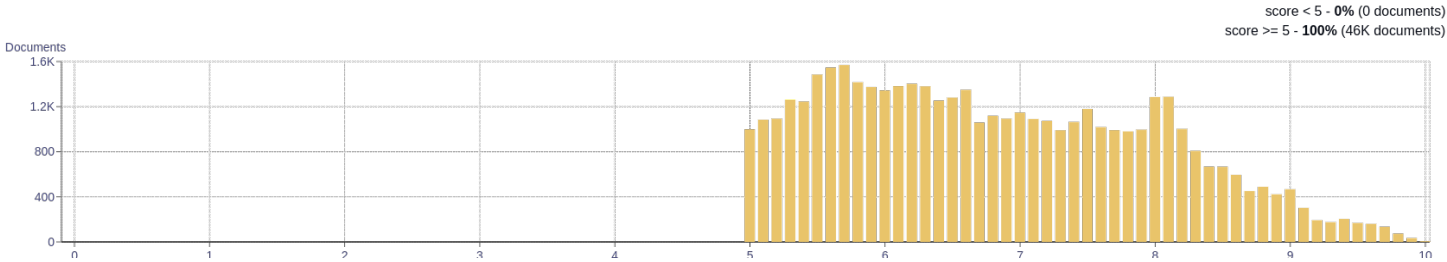
Number of segments



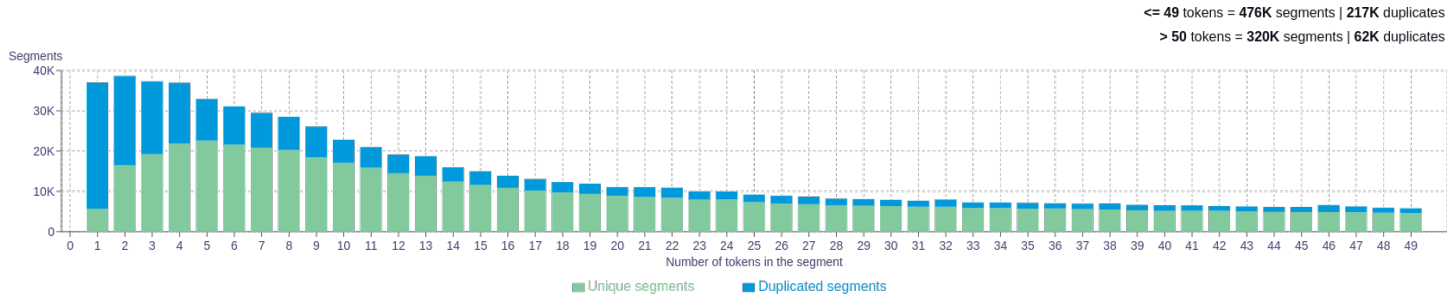
Percentage of segments in Samoan (sm) inside documents



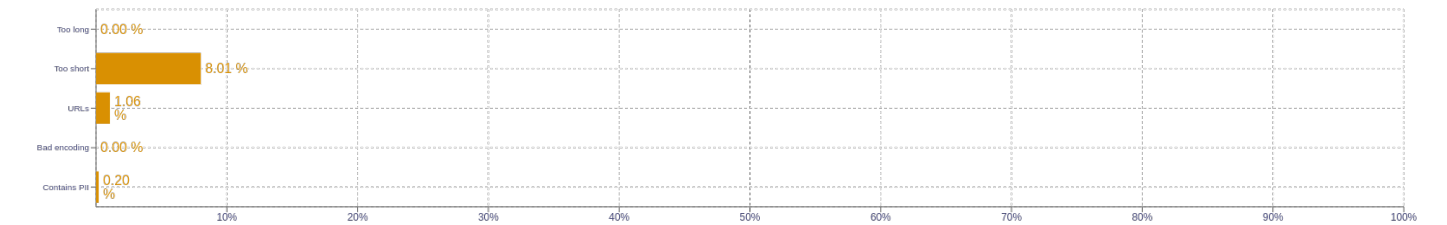
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



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 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>, <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>