HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
smo_Latn.jsonl.tsv	12/4/2024	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

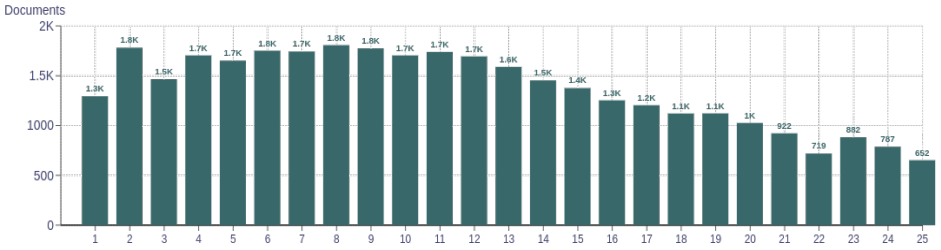
Domain	Docs	% of total
samoanews.com	2.6K	5.76
jw.org	2.2K	4.80
samoaobserver.ws	1.7K	3.70
wikipedia.org	1.2K	2.66
martech.zone	1K	2.28
sobserver.ws	900	1.96
eturbonews.com	828	1.81
samoatimes.co.nz	742	1.62
recursosdeautoayuda.com	734	1.60
lds.org	725	1.58

Top 10 TLDs

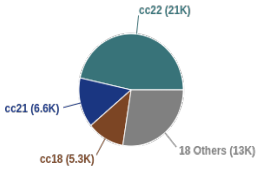
Domain	Docs	% of total
com	30K	65.54
org	6.5K	14.14
ws	2.7K	5.84
co.nz	1.2K	2.68
zone	1K	2.28
net	848	1.85
ru	313	0.68
es	258	0.56
govt.nz	215	0.47
network	181	0.39

Documents size (in segments)

<= 25 segments **74.61%** (34K documents)
> 25 segments **25.39%** (12K documents)

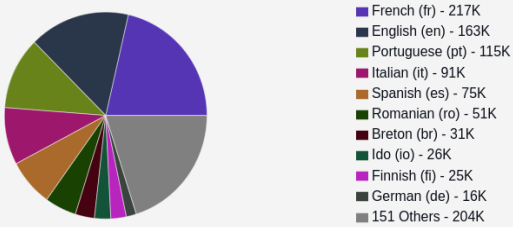


Documents by collection



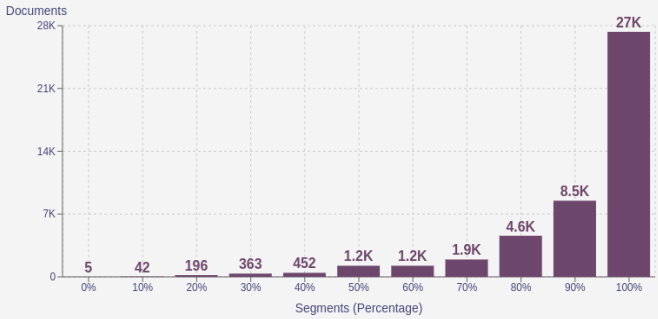
Language Distribution

Number of segments



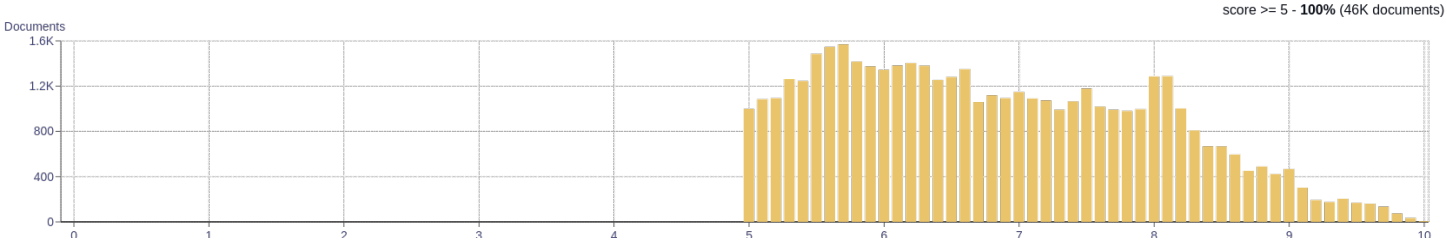
*Samoan (sm) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Samoan (sm) inside documents



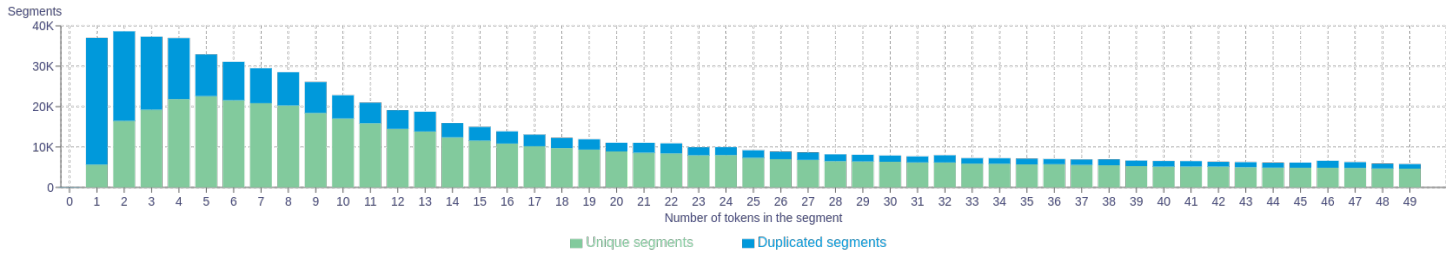
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (46K documents)

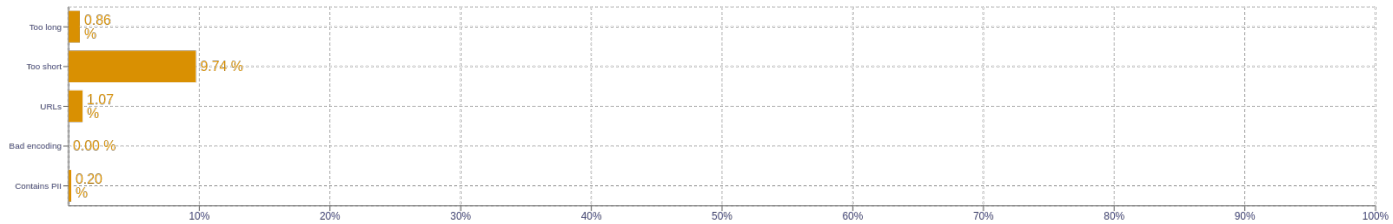


Segment length distribution by token

<= 49 tokens = **476K** segments | **217K** duplicates
> 50 tokens = **320K** segments | **62K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	te | 318553 · mafai | 193269 · tatou | 152542 · pe | 135788 · lelei | 120905
2	ou te | 36659 · matou te | 27474 · te oe | 26962 · nai lo | 20844 · tatou te | 19230
3	afai e te | 13823 · mafai ona tatou | 6941 · pito i luga | 6000 · tatau ona tatou | 5522 · mafai ona maua | 4631
4	fesoasoani ia te oe | 4068 · luga o le initaneti | 2692 · afai e te mana‘o | 2308 · pe afai e te | 2277 · luga o le upega | 2258
5	te fai atu ia te | 1289 · luga o le upega tafa‘ilagi | 1201 · tu‘uina atu ia te oe | 1196 · pe a fai e te | 967 · ta‘u atu ia te oe | 933

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