HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
sat_Olck.jsonl.tsv	11/28/2024	Santali (sat)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
2,566	45,801	26,205 (57.21 %)	1.3M	14.52 MB	6,222,595

Top 10 domains

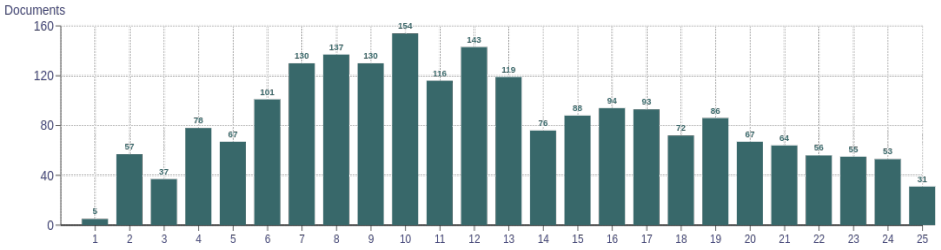
Domain	Docs	% of total
wikipedia.org	2.3K	89.44
vikaspedia.in	178	6.94
globalvoices.org	47	1.83
raharahla.com	15	0.58
wikimedia.org	8	0.31
wikiplanet.click	5	0.19
wikiversity.org	4	0.16
know.cf	3	0.12
santalinews.com	3	0.12
mediawiki.org	2	0.08

Top 10 TLDs

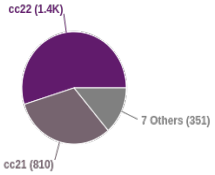
Domain	Docs	% of total
org	2.4K	91.89
in	178	6.94
com	21	0.82
click	5	0.19
cf	3	0.12
cn	1	0.04

Documents size (in segments)

<= 25 segments **82.19%** (2.1K documents)
> 25 segments **17.81%** (457 documents)

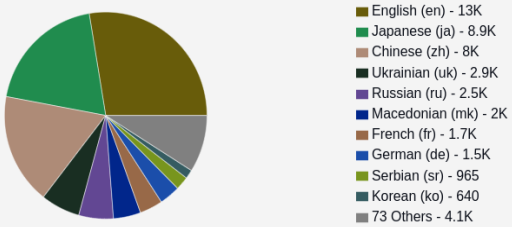


Documents by collection



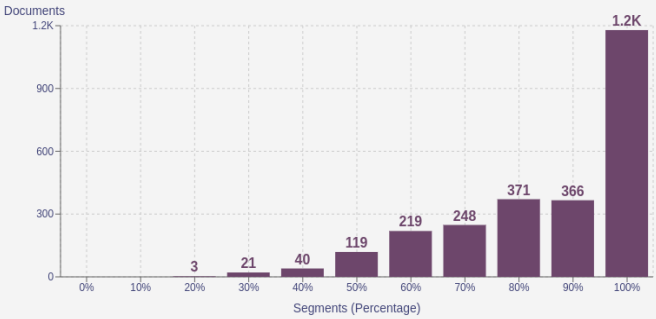
Language Distribution

Number of segments



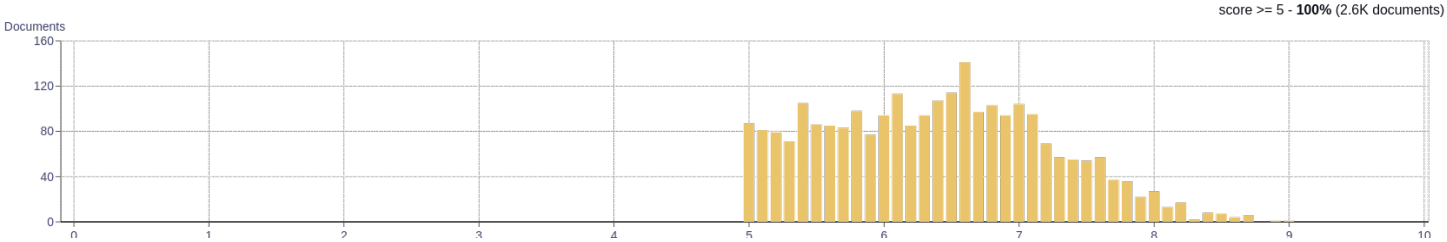
*Santali (sat) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Santali (sat) inside documents



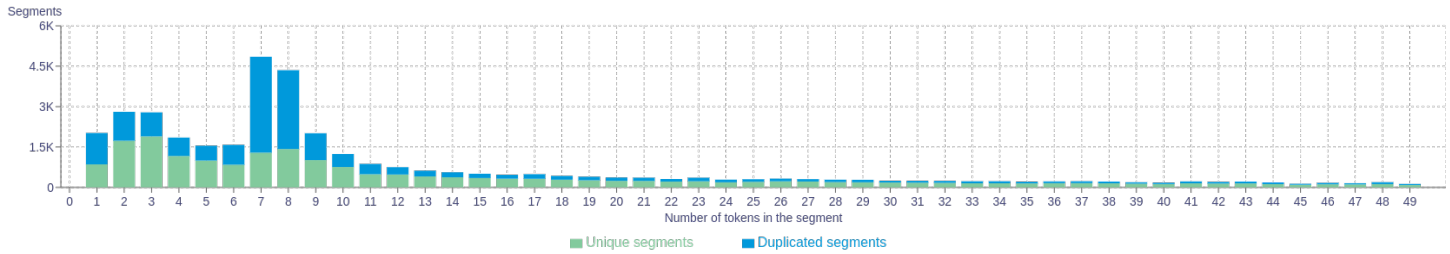
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (2.6K documents)

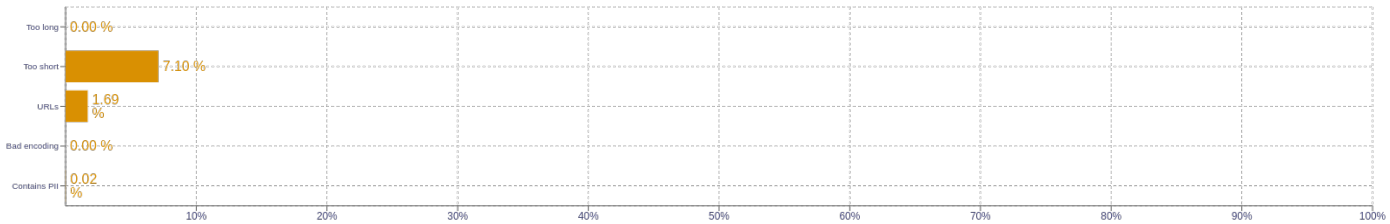


Segment length distribution by token

<= 49 tokens = **20K** segments | **17K** duplicates
> 50 tokens = **8.3K** segments | **2.4K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	ᏦᏁᏆᏃᏭ | 17999 · ᏆᎪᏫᏫᏁᎤ | 8934 · ᏬᎴᎮᎤ | 7799 · ᏮᏫ | 6596 · ᏬᏫ | 6556
2	ᏆᎪᏫᏫᏁᎤ ᏦᏁᏆᏃᏭ | 8849 · ᏬᏫ ᏢᏣᎪᏭ | 1134 · ᏧᏕᏃᎶ ᏰᏯᏭ | 922 · ᏢᏫ ᎤᏔᏬᏕᏫ | 705 · ᏢᏕᏫ ᎤᏔᏬᏕᏫ | 665
3	wikipedia articles with | 394 · from the original | 382 · archived from the | 378 · ᏬᏫ ᏢᏣᎪᏭ ᏢᏣᏭ | 368 · ᏦᏃᏬᏭ ᏓᏫᏃᎶ ᏬᏫ | 324
4	archived from the original | 378 · ᏦᏃᏬᏭ ᏓᏫᏃᎶ ᏬᏫ ᏢᏣᎪᏭ | 322 · from the original on | 315 · ᎮᎴᏃᎶ ᏬᎴᎮᎤ ᏰᎴ ᏨᏮᏮ | 218 · ᏆᏔᏫᎶᏭ ᎤᏕ ᏁᏆᏆᏃᏭ ᏧᏦᏫ | 181
5	archived from the original on | 312 · ᏦᏃᏬᏭ ᏓᏫᏃᎶ ᏬᏫ ᏢᏣᎪᏭ ᏢᏣᏭ | 277 · ᏆᏔᏫᎶᏭ ᎤᏕ ᏁᏆᏆᏃᏭ ᏧᏦᏫ ᏰᏣᎴᏫᏭ | 177 · ᏬᎴᎮᎤ ᏓᏫᏃᎶ ᏢᏣᎪᏭ ᎴᎴ ᏢᏣᎪᏭ | 134 · ᏧᏍᏫᎶ ᏰᏯᏭᎶ ᏬᏭᏯ ᎴᏣᏭᎶ | 126

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