

General overview

Corpus	Analytics date	Language
kac_Latn.jsonl.tsv	11/26/2024	Kachin (kac)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
7,587	159,416	85,544 (53.66 %)	6.9M	27.31 MB	28,248,636

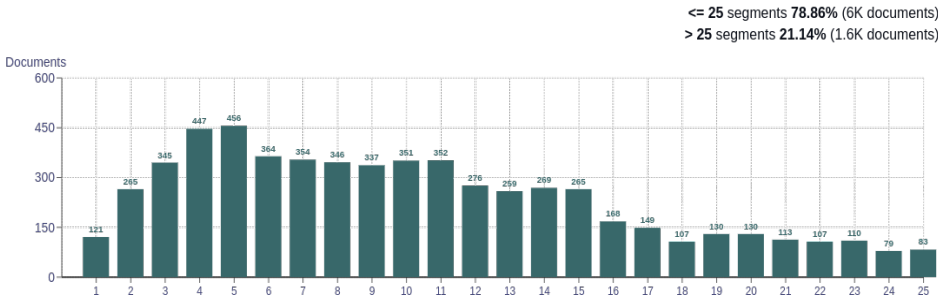
Top 10 domains

Domain	Docs	% of total
blogspot.com	2.4K	31.28
kachinnews.com	1.1K	15.13
dehong.gov.cn	1.1K	14.50
blogspot.sg	371	4.89
jw.org	358	4.72
blogspot.co.nz	342	4.51
kachinnet.net	286	3.77
hkakaborazi.net	139	1.83
rvasia.org	135	1.78
kachinlandnews.com	117	1.54

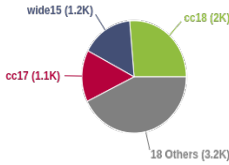
Top 10 TLDs

Domain	Docs	% of total
com	4.2K	55.08
gov.cn	1.1K	14.50
org	837	11.03
net	460	6.06
sg	371	4.89
co.nz	342	4.51
in	72	0.95
ca	56	0.74
se	23	0.30
de	20	0.26

Documents size (in segments)

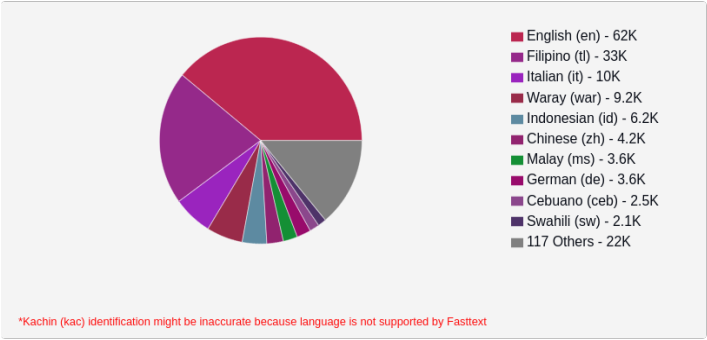


Documents by collection

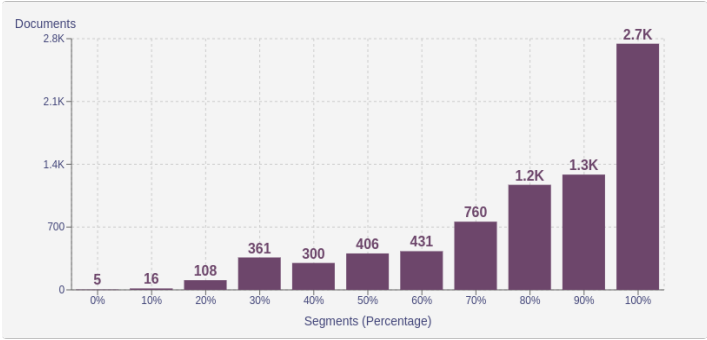


Language Distribution

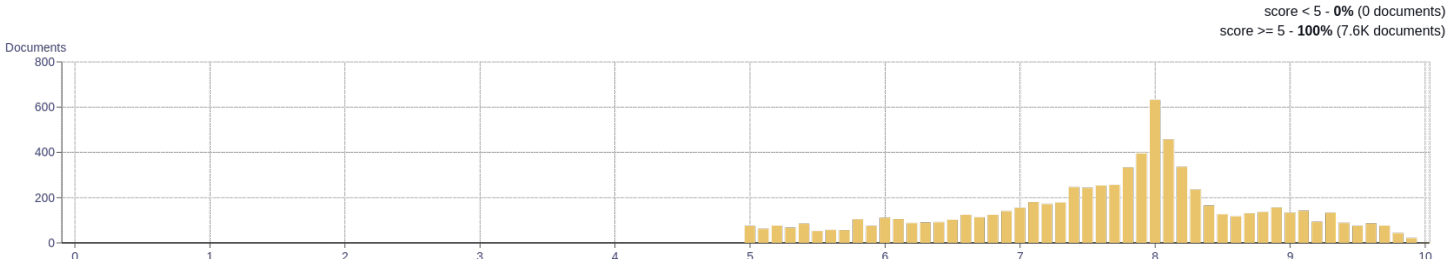
Number of segments



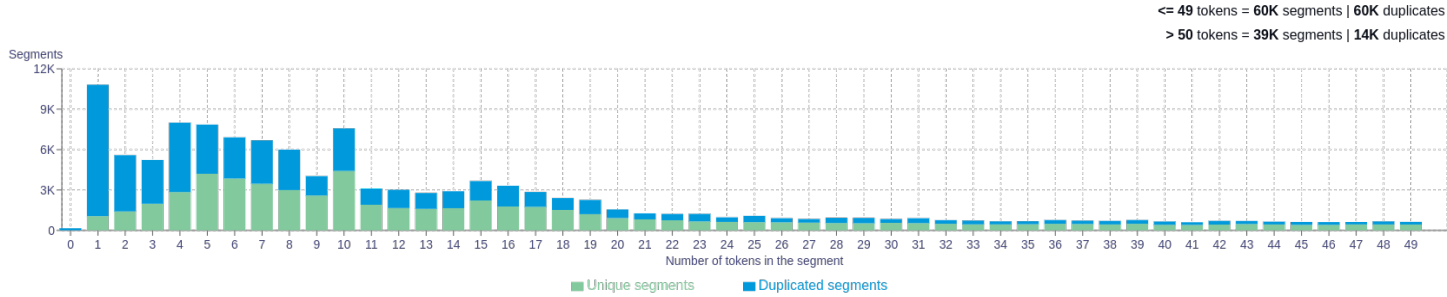
Percentage of segments in Kachin (kac) inside documents



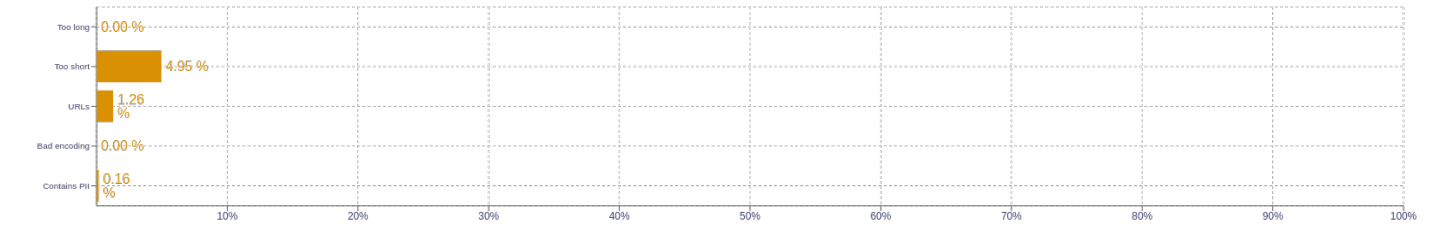
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>ai 428859</div> <div>ni 160002</div> <div>nga 141790</div> <div>hpe 139812</div> <div>gaw 97684</div>
2	<div>nga ai 73792</div> <div>ai lam 36289</div> <div>ni hpe 28241</div> <div>rai nga 26200</div> <div>ra ai 15982</div>
3	<div>rai nga ai 20569</div> <div>yawn yawn yawn 7311</div> <div>chye lu ai 7210</div> <div>nga ma ai 6484</div> <div>ai lam ni 5497</div>
4	<div>yawn yawn yawn yawn 7288</div> <div>lam chye lu ai 2369</div> <div>kaning rai nme law 2159</div> <div>lam rai nga ai 1877</div> <div>ai lam rai nga 1490</div>
5	<div>yawn yawn yawn yawn yawn 7265</div> <div>ai lam rai nga ai 1407</div> <div>ai lam chye lu ai 1363</div> <div>lam na chye lu ai 882</div> <div>shiga na chye lu ai 775</div>

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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>, , , 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>