

General overview

Corpus	Analytics date	Language
ace_Latn.jsonl.tsv	10/2/2024	Acehnese (ace)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
12,930	206,187	107,160 (51.97 %)	9.7M	49.23 MB	50,639,739

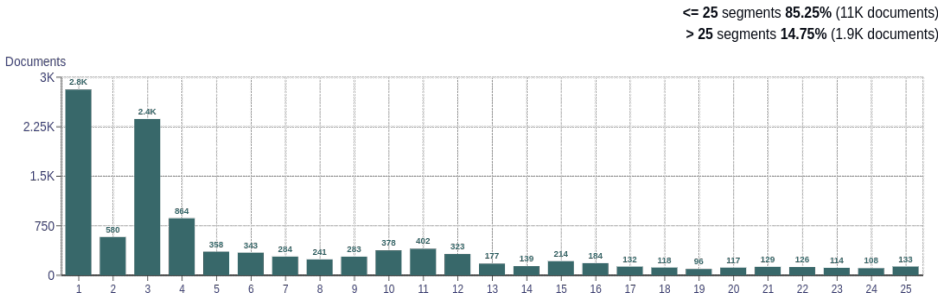
Top 10 domains

Domain	Docs	% of total
bible.is	4.3K	33.40
wikipedia.org	2.8K	21.99
wordproject.org	1.2K	9.15
petalokasi.org	362	2.80
blogspot.com	338	2.61
wordpress.com	209	1.62
azlyricdb.com	180	1.39
kodeposindo.xyz	166	1.28
fanskipop.com	148	1.14
jerseysu.com	85	0.66

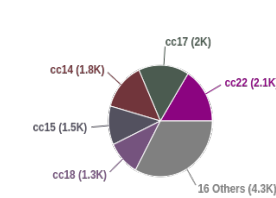
Top 10 TLDs

Domain	Docs	% of total
org	4.6K	35.52
is	4.3K	33.40
com	2.6K	19.81
xyz	172	1.33
net	166	1.28
com.br	93	0.72
ru	87	0.67
mobi	71	0.55
co.id	65	0.50
tv	50	0.39

Documents size (in segments)

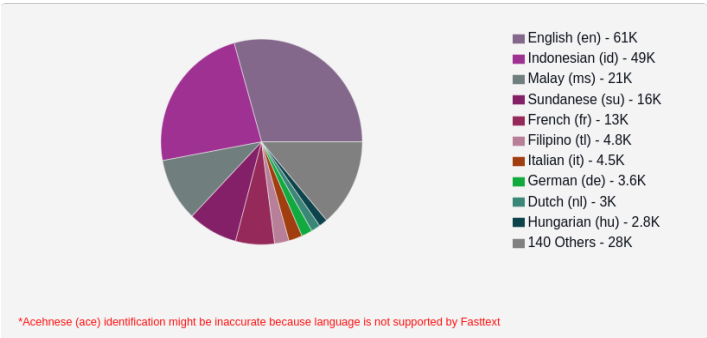


Documents by collection

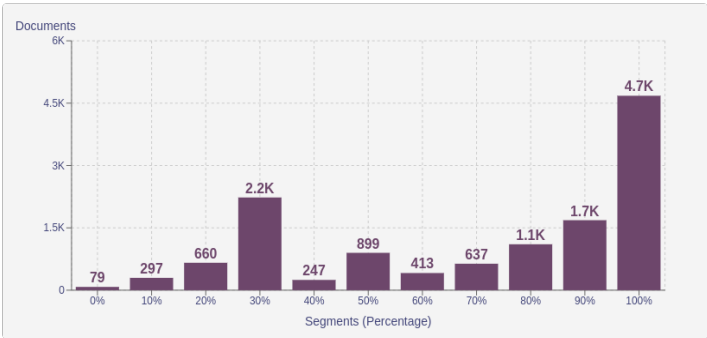


Language Distribution

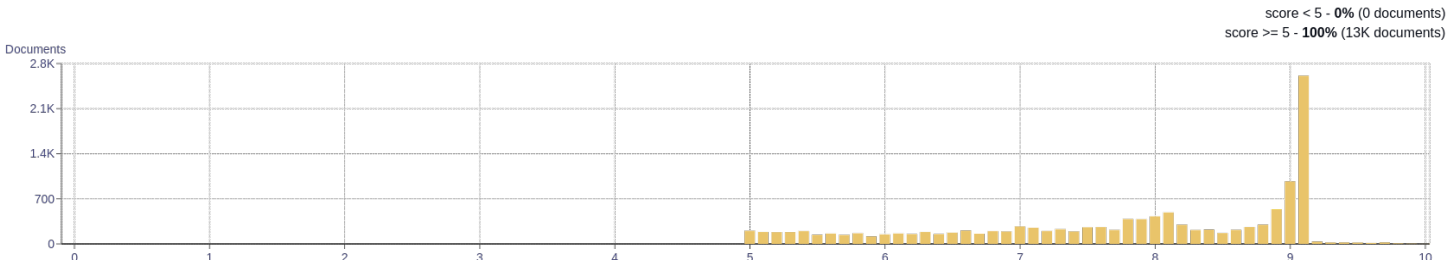
Number of segments



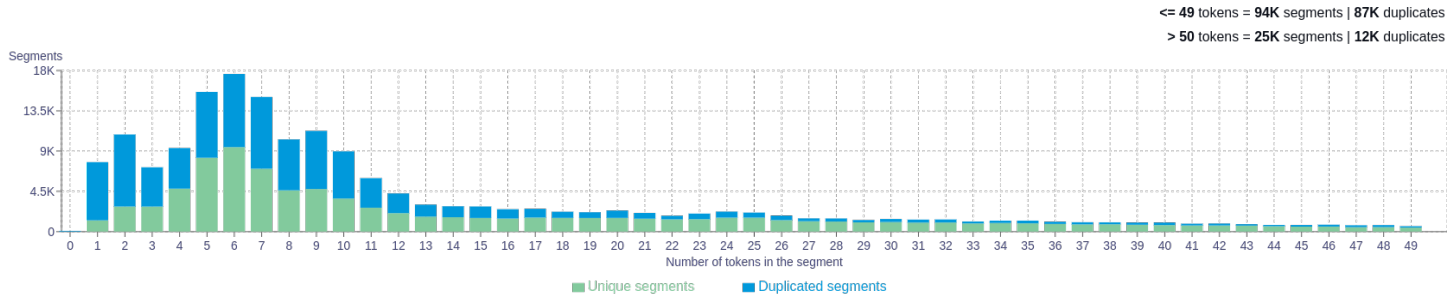
Percentage of segments in Acehnese (ace) inside documents



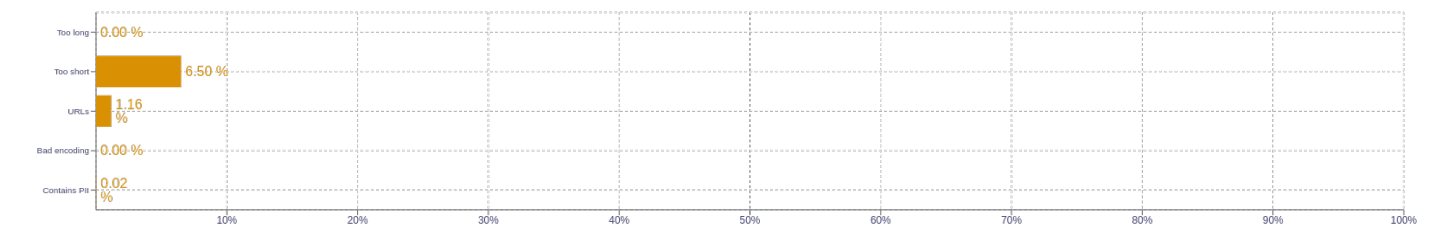
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>blah 290258</div> <div>bak 100221</div> <div>ureuëng 74691</div> <div>gata 72273</div> <div>ulôn 63060</div>
2	<div>blah blah 258752</div> <div>teu allah 10208</div> <div>zis gas 9104</div> <div>meunan cit 8720</div> <div>blahblah blah 8433</div>
3	<div>blah blah blah 236064</div> <div>blah blahblah blah 7693</div> <div>blahblah blah blah 6509</div> <div>blah blah blahblah 6088</div> <div>taiq zis gas 5663</div>
4	<div>blah blah blah blah 217503</div> <div>blahblah blah blah blah 6054</div> <div>blah blah blahblah blah 5964</div> <div>blah blahblah blah blah 5021</div> <div>blah blah blah blahblah 5639</div>
5	<div>blah blah blah blah blah 204731</div> <div>blah blah blahblah blah blah 5730</div> <div>blah blah blah blahblah blah 5533</div> <div>blah blahblah blah blah blah 5371</div> <div>blahblah blah blah blah blah 4115</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>