HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
bjn_Arab.jsonl.tsv	12/5/2024	Banjar (bjn)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
1,112	19,529	11,025 (56.45 %)	810K	5.63 MB	3,298,215

Top 10 domains

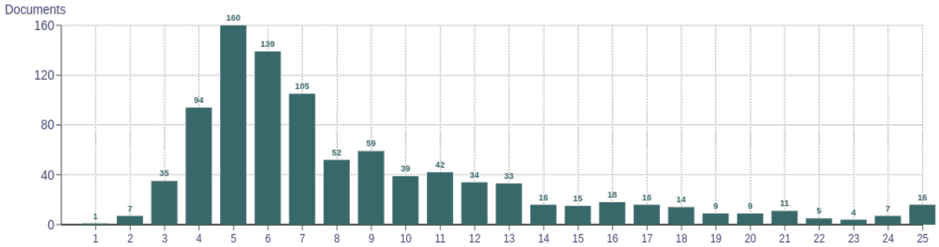
Domain	Docs	% of total
utusanmelayu.com.my	514	46.22
wordpress.com	283	25.45
blogspot.com	131	11.78
blogspot.my	18	1.62
utusantv.com	14	1.26
wikisource.org	13	1.17
urusniaga.my	9	0.81
unicode.org	8	0.72
wikimedia.org	8	0.72
tronik.org	7	0.63

Top 10 TLDs

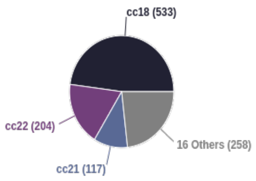
Domain	Docs	% of total
com.my	516	46.40
com	478	42.99
org	43	3.87
my	27	2.43
net	5	0.45
sg	5	0.45
eu	4	0.36
web.id	4	0.36
org.my	3	0.27
moe	3	0.27

Documents size (in segments)

<= 25 segments **84.53%** (940 documents)
> 25 segments **15.47%** (172 documents)

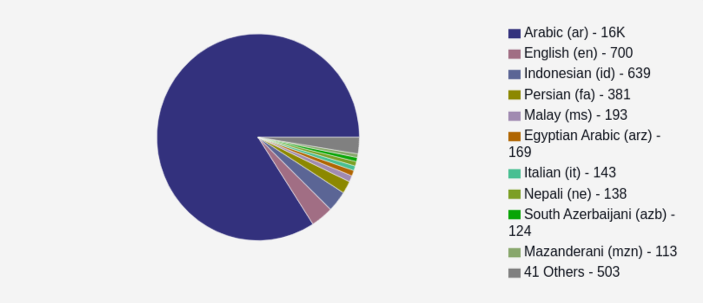


Documents by collection



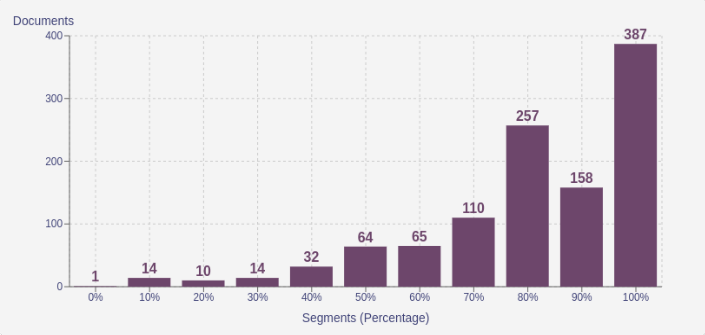
Language Distribution

Number of segments



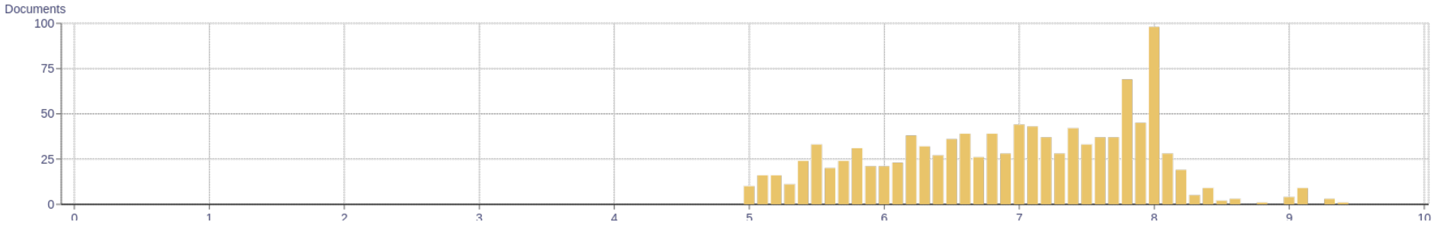
*Banjar (bjn) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Banjar (bjn) inside documents



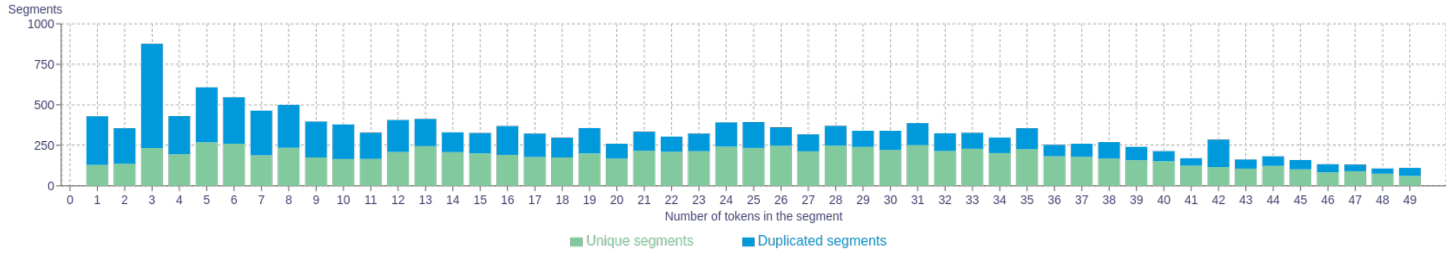
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (1.1K documents)

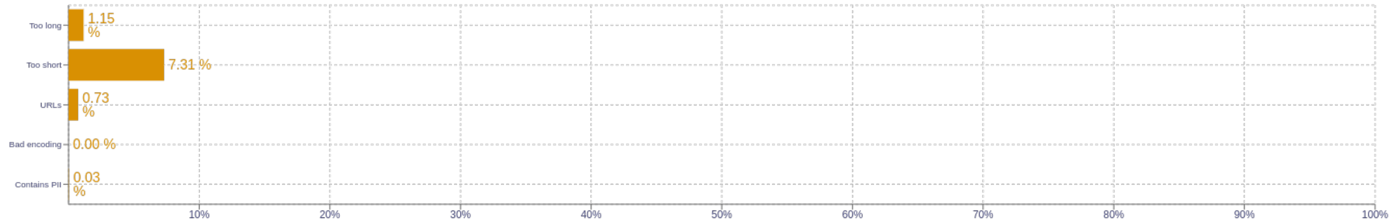


Segment length distribution by token

<= 49 tokens = **9.1K** segments | **7.2K** duplicates
> 50 tokens = **3.3K** segments | **1.3K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	16456 | يڠ 9389 | م 8403 | و 8149 | ن 7257 | ي
2	602 | اورڠ يڠ 515 | هاري اين 478 | اڤ يڠ 461 | يڠ تله 454 | ڤردان منتري
3	436 | دركن دركن دركن 364 | يڠ دڤرتوان اݢوڠ 290 | الله عليه وسلم 289 | صلى الله عليه 239 | الله سبحانه وتعالى
4	434 | دركن دركن دركن دركن 283 | صلى الله عليه وسلم 168 | ماسيس چت ماسيس چت 154 | چت ماسيس چت ماسيس 136 | ستياڤ اورڠ عداله برحق
5	432 | دركن دركن دركن دركن دركن 154 | چت ماسيس چت ماسيس چت 154 | ماسيس چت ماسيس چت ماسيس 128 | ستياڤ اورڠ عداله برحق كڤد 98 | رسول الله صلى الله عليه

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