HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
knc_Latn.jsonl.tsv	12/13/2024	Kanuri (knc)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
2,472	10,521	5,151 (48.96 %)	3M	13.01 MB	11,941,756

Top 10 domains

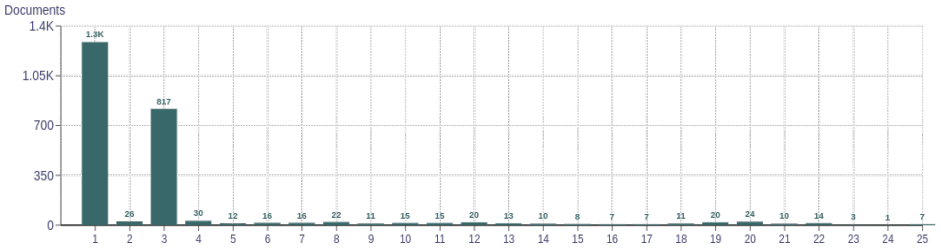
Domain	Docs	% of total
bible.is	2.1K	85.15
bibles.org	78	3.16
dandalkura.com	70	2.83
fivecowries.online	41	1.66
ngbible.com	38	1.54
ebible.org	32	1.29
wikimedia.org	8	0.32
mp3songspk.info	6	0.24
unicode.org	5	0.20
blogspot.com	5	0.20

Top 10 TLDs

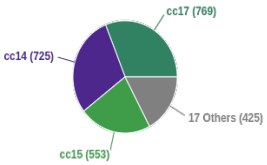
Domain	Docs	% of total
is	2.1K	85.15
com	151	6.11
org	134	5.42
online	41	1.66
ru	7	0.28
info	7	0.28
in	5	0.20
net	5	0.20
org.in	3	0.12
co.uk	2	0.08

Documents size (in segments)

<= 25 segments **97.94%** (2.4K documents)
> 25 segments **2.06%** (51 documents)

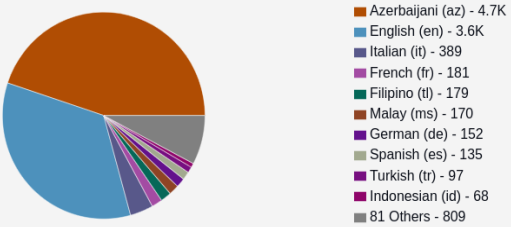


Documents by collection



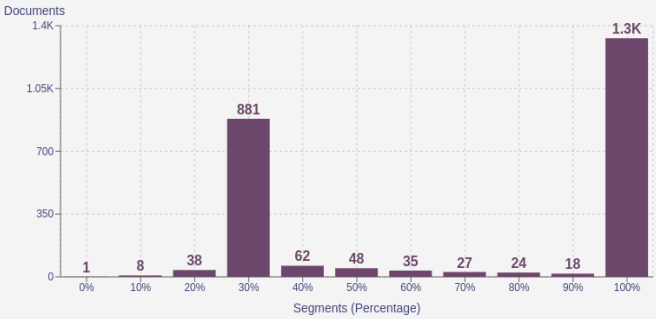
Language Distribution

Number of segments



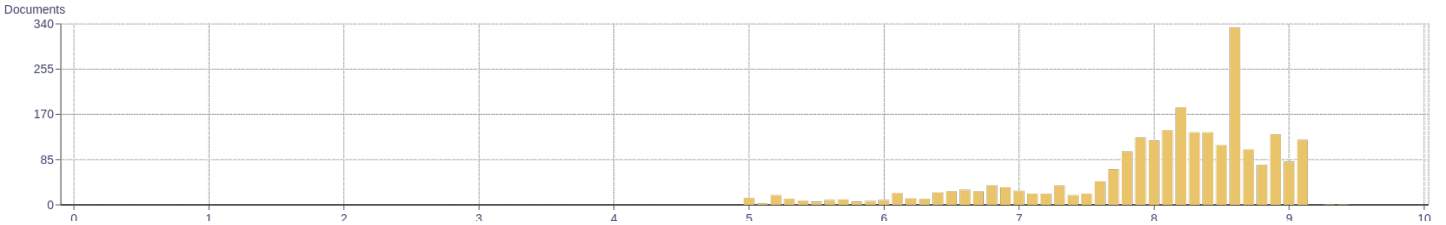
*Kanuri (knc) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Kanuri (knc) inside documents



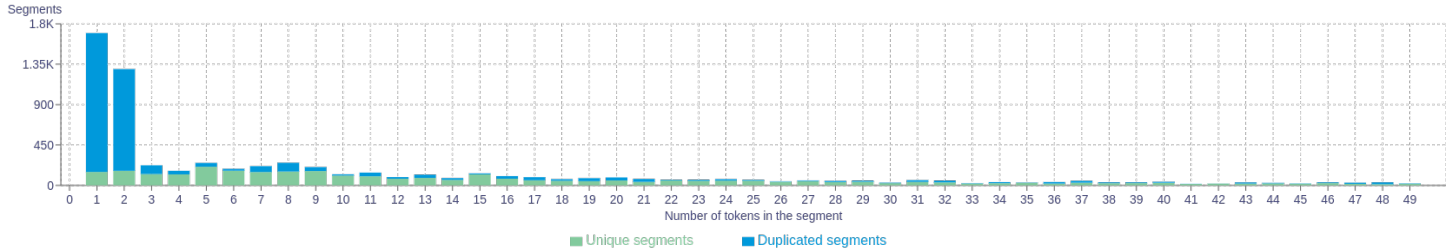
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (2.5K documents)

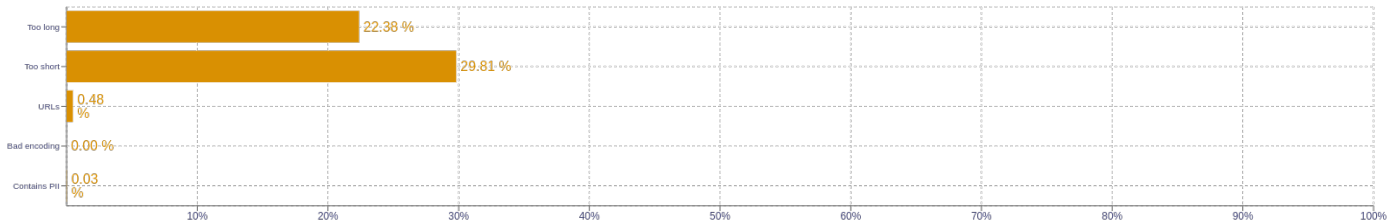


Segment length distribution by token

<= 49 tokens = **3.1K** segments | **3.7K** duplicates
> 50 tokens = **3.7K** segments | **1.7K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	ka | 59234 · kɨ | 56711 · na | 50316 · tɨ | 38743 · tə | 33073
2	tɨ kɨ | 5600 · ə nə | 4937 · je kɨ | 4896 · aye na | 4564 · ndo hay | 3851
3	ane tuk na | 1837 · andza niye na | 1194 · bazlam i mbələm | 1141 · njé ndo je | 972 · de ə nə | 899
4	kate t kate t | 896 · t kate t kate | 894 · kɨ ta luwə tɨ | 559 · ndo məpe mədzal gər | 549 · məpe mədzal gər hay | 518
5	t kate t kate t | 894 · kate t kate t kate | 893 · ndo məpe mədzal gər hay | 495 · məpe mədzal gər hay ka | 489 · mədzal gər hay ka yesu | 479

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