HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
kn_1.jsonl.tsv	3/19/2024	Kannada (kn)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
228,215	29,241,332	6,052,194 (20.70 %)	301M	3.8 GB	

Top 10 domains

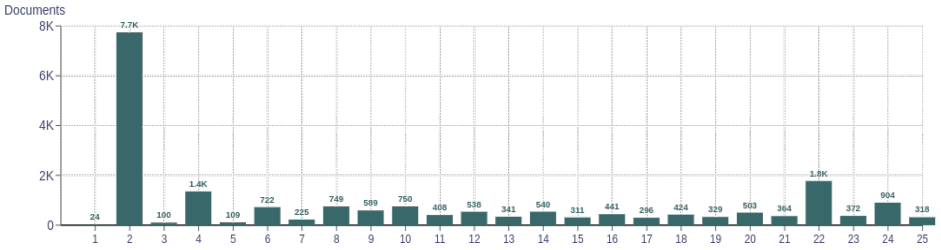
Domain	Docs	% of total
blogspot.in	15K	6.53
indiatimes.com	9.9K	4.35
news18.com	7.8K	3.43
varthabharati.in	7.7K	3.36
kannadadunia.com	6.5K	2.85
asianetnews.com	5.8K	2.55
wikipedia.org	4.6K	2.01
blogspot.com	4.5K	1.96
newskannada.com	4.1K	1.79
prajavani.net	4K	1.75

Top 10 TLDs

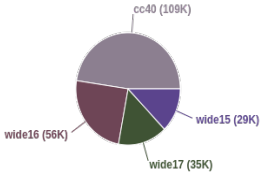
Domain	Docs	% of total
com	145K	63.39
in	44K	19.40
org	17K	7.25
net	11K	4.77
news	2.7K	1.18
co.in	1.2K	0.51
gov.in	689	0.30
pt	488	0.21
ae	452	0.20
today	438	0.19

Documents size (in segments)

<= 25 segments **8.86%** (20K documents)
> 25 segments **91.14%** (208K documents)

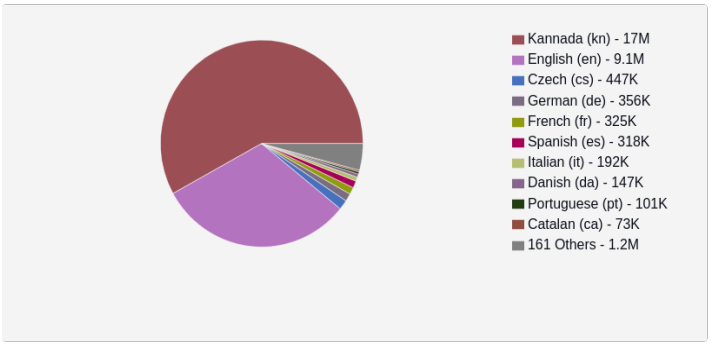


Documents by collection

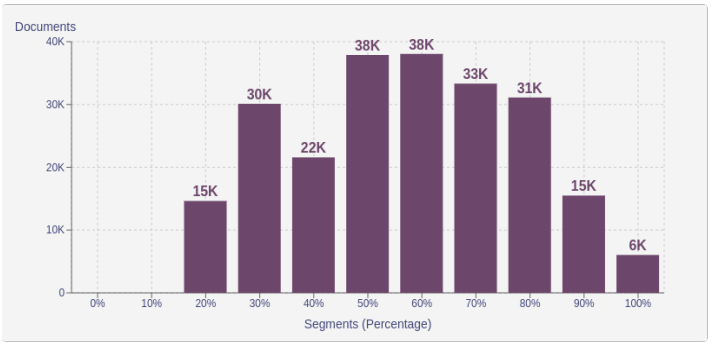


Language Distribution

Number of segments

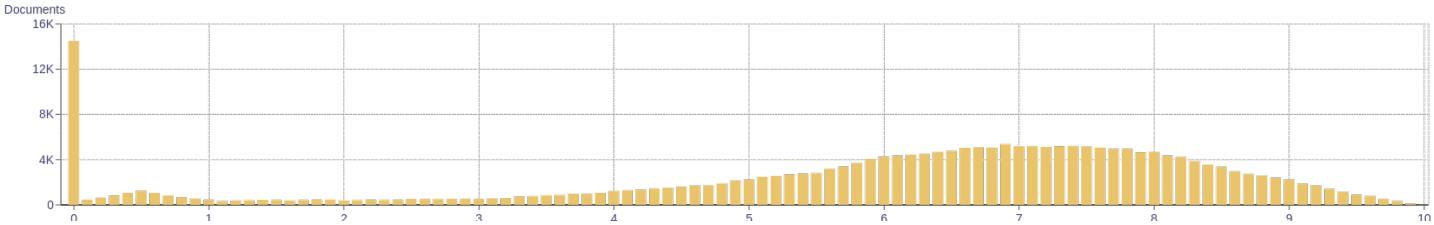


Percentage of segments in Kannada (kn) inside documents



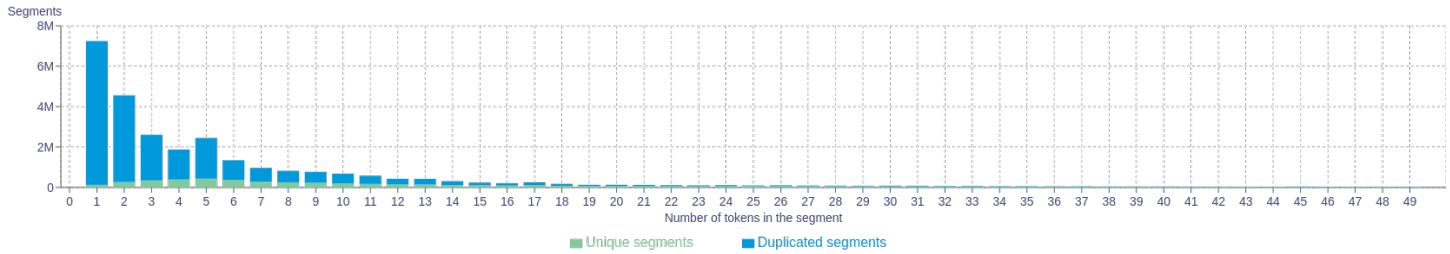
Distribution of documents by document score

score < 5 - **23.86%** (54K documents)
score >= 5 - **76.14%** (174K documents)

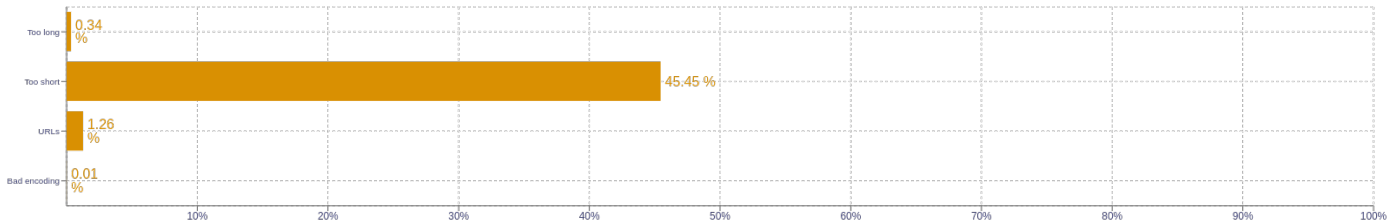


Segment length distribution by token

<= 49 tokens = **5.2M** segments | **23M** duplicates
> 50 tokens = **1.1M** segments | **315K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	the | 1080755 · to | 1014534 · in | 823065 · news | 801921 · of | 674903
2	of the | 120948 · span style | 115762 · rights reserved | 112002 · all rights | 104013 · no comments | 96691
3	all rights reserved | 103766 · opens in new | 53064 · in new window | 53044 · to twittershare to | 50021 · share to twittershare | 50021
4	opens in new window | 53041 · share to twittershare to | 50021 · twittershare to facebookshare to | 45919 · to twittershare to facebookshare | 45919 · to facebookshare to pinterest | 45919
5	twittershare to facebookshare to pinterest | 45919 · to twittershare to facebookshare to | 45919 · share to twittershare to facebookshare | 45919 · leave a reply cancel reply | 31648 · address will not be published | 27715

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