

General overview

Corpus	Analytics date	Language
kan_Knda.jsonl.tsv	9/18/2024	Kannada (kn)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
1,335,847	24,929,282	12,772,248 (51.23 %)	653M	10.46 GB	4,274,156,104

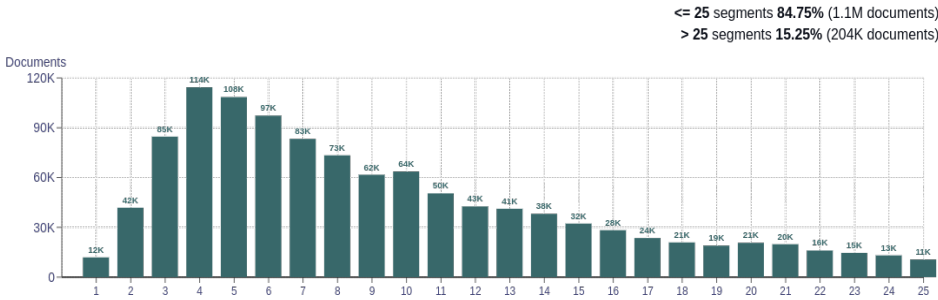
Top 10 domains

Domain	Docs	% of total
prajavani.net	79K	5.92
udayavani.com	52K	3.90
wikipedia.org	52K	3.90
news18.com	47K	3.52
blogspot.com	45K	3.37
filmibeat.com	42K	3.13
oneindia.com	39K	2.92
asianetnews.com	33K	2.45
indiatimes.com	31K	2.29
gulfkannadiga.com	24K	1.78

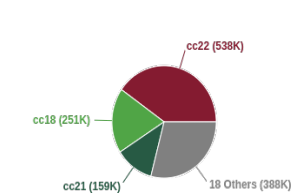
Top 10 TLDs

Domain	Docs	% of total
com	905K	67.75
in	165K	12.33
net	116K	8.70
org	90K	6.75
news	24K	1.78
co.in	3.4K	0.25
live	3.4K	0.25
gov.in	3.1K	0.23
today	2.2K	0.17
online	2K	0.15

Documents size (in segments)

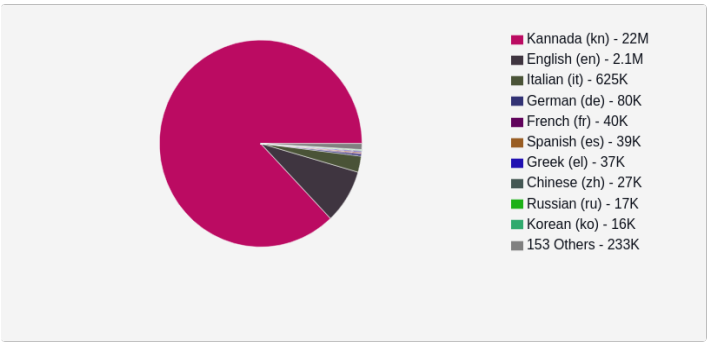


Documents by collection

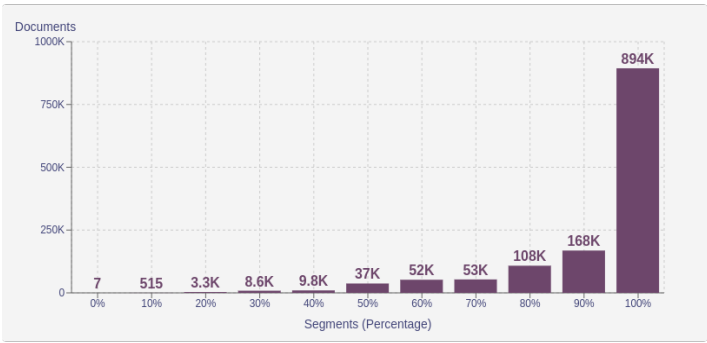


Language Distribution

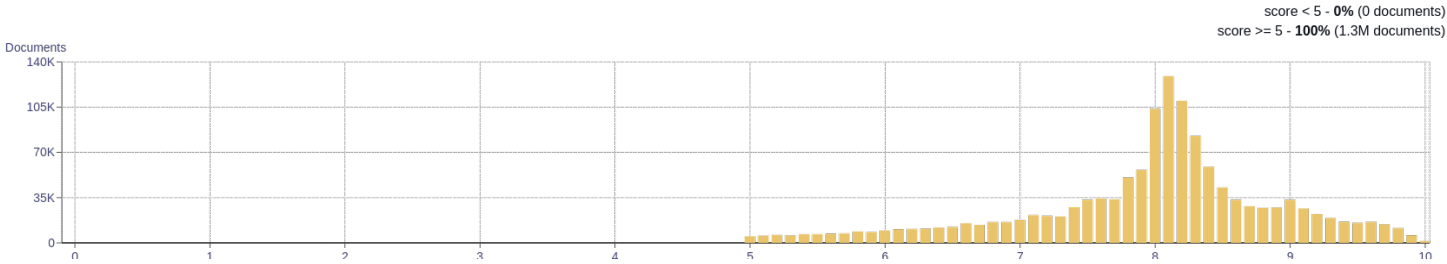
Number of segments



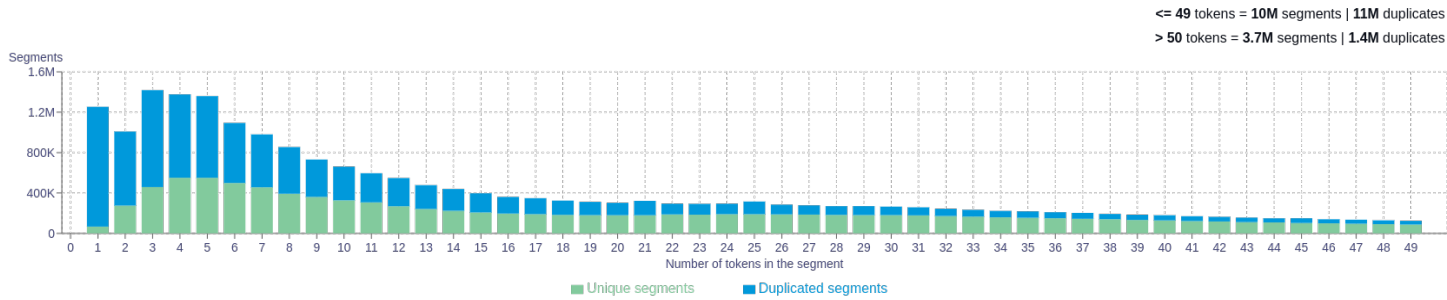
Percentage of segments in Kannada (kn) inside documents



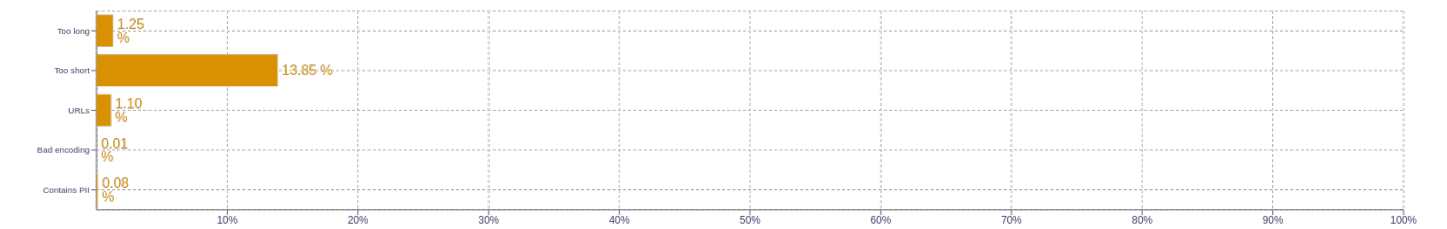
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>ಸವ್ವು 1068630</div> <div>ನಿವ್ವು 1018629</div> <div>ನೀವು 922077</div> <div>ನಾನು 912901</div> <div>ನನ್ನ 878025</div>
2	<div>ಇನ್ನೂ ಓದಿ 89081</div> <div>of the 78259</div> <div>ಕೋಟಿ ರೂ 64493</div> <div>ಸರೇಂದ್ರ ಮೊದಿ 59053</div> <div>pm ist 52952</div>
3	<div>lua error in 33784</div> <div>ಪ್ರಧಾನಿ ಸರೇಂದ್ರ ಮೊದಿ 33331</div> <div>from the original 32458</div> <div>archived from the 32371</div> <div>error in ಮಾರ್ಗೋಲ್ 31277</div>
4	<div>archived from the original 32364</div> <div>lua error in ಮಾರ್ಗೋಲ್ 31250</div> <div>from the original on 30007</div> <div>to compare number with 29362</div> <div>compare number with nil 29362</div>
5	<div>archived from the original on 29966</div> <div>to compare number with nil 29362</div> <div>attempt to compare number with 29362</div> <div>ಸುದ್ದಿಗಳಾಗಿ ಪ್ರಜಾವಾಣಿ ಆಫ್ ಡೌನ್‌ಲೋಡ್ ಮಾಡಿಕೊಳ್ಳಿ 24587</div> <div>ತಾಜಾ ಸುದ್ದಿಗಳಾಗಿ ಪ್ರಜಾವಾಣಿ ಆಫ್ ಡೌನ್‌ಲೋಡ್ 24587</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.stinfo.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>