HPLT Analytics report

HPLTAnalytics

General overview

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

Volumes

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

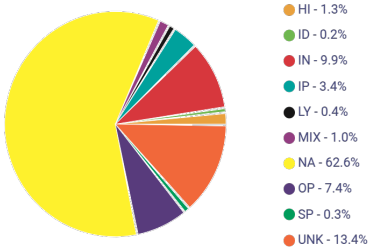
Top 10 domains

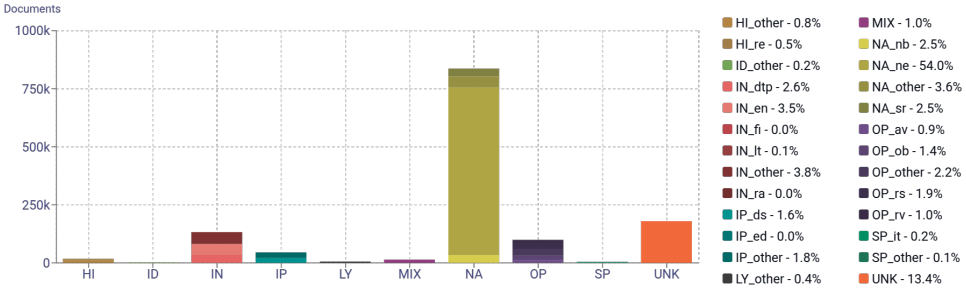
Domain	Docs	% of total
prajavani.net	79K	5.92%
udayavani.com	52K	3.90%
wikipedia.org	52K	3.90%
news18.com	47K	3.52%
blogspot.com	45K	3.37%
filmibeat.com	42K	3.13%
oneindia.com	39K	2.92%
asianetnews.com	33K	2.45%
indiatimes.com	31K	2.29%
gulfkannadiga.com	24K	1.78%

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%

Register labels

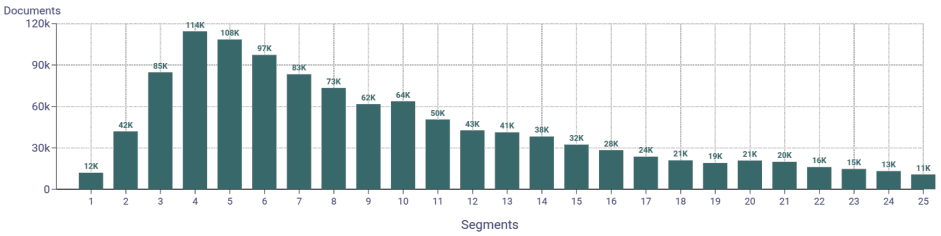




MT:6.5% | 86K Documents

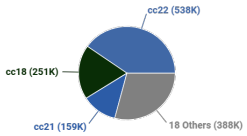
Documents size (in segments)

<= 25 segments **84.75%** (1.1M documents)
> 25 segments **15.25%** (204K documents)



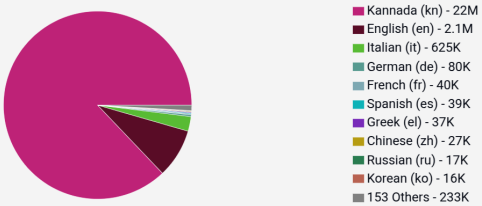
Documents by collection

CC = 79.43%
IA = 20.57%

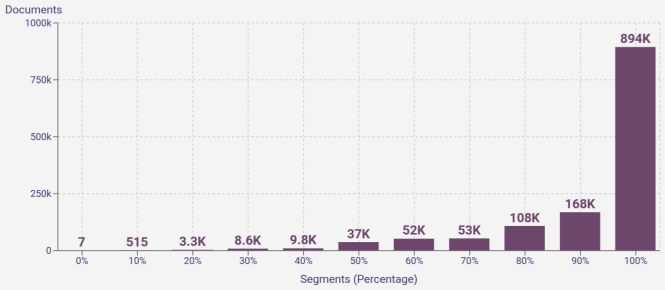


Language Distribution

Number of segments in the Kannada (kn) corpus

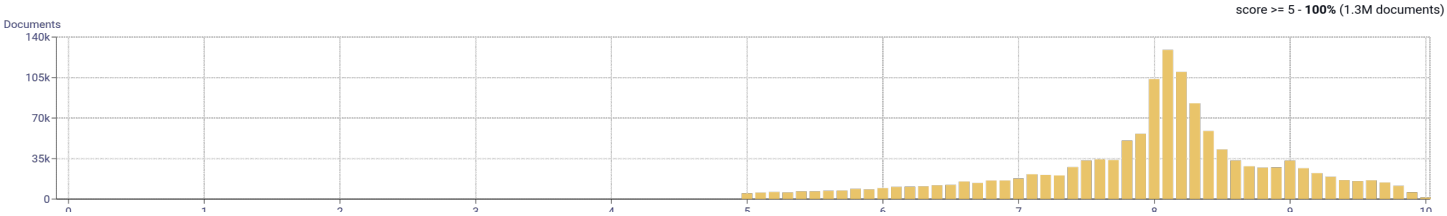


Percentage of segments in Kannada (kn) inside documents



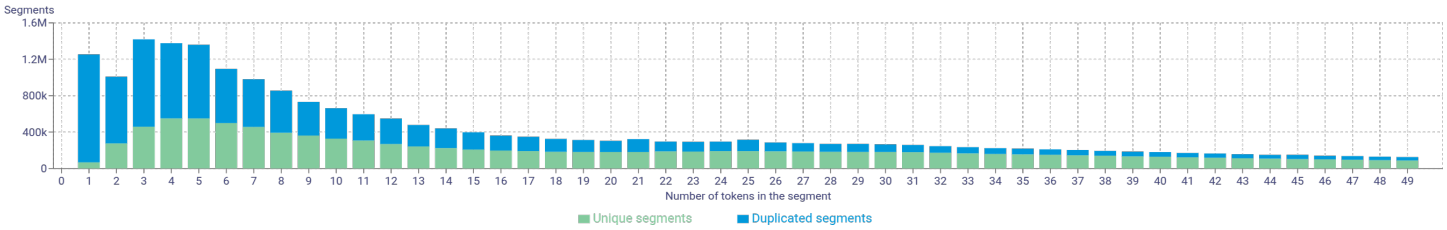
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (1.3M documents)

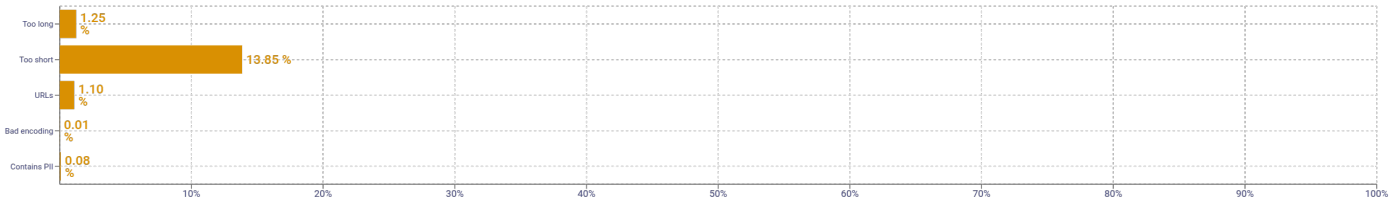


Segment length distribution by token

≤ 49 tokens = **10M** segments | **11M** duplicates
> 50 tokens = **3.7M** segments | **1.4M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	ನಮ್ಮ | 1068630 · ನಿಮ್ಮ | 1018629 · ನೀವು | 922077 · ನಾನು | 912901 · ನನ್ನ | 878025
2	ಇದನ್ನೂ ಓದಿ | 89081 · of the | 78259 · ಕೋಟಿ ರೂ | 64493 · ನರೇಂದ್ರ ಮೋದಿ | 59053 · pm ist | 52952
3	lua error in | 33784 · ಪ್ರಧಾನಿ ನರೇಂದ್ರ ಮೋದಿ | 33331 · from the original | 32458 · archived from the | 32371 · error in ಮಾಡ್ಯೂಲ್ | 31277
4	archived from the original | 32364 · lua error in ಮಾಡ್ಯೂಲ್ | 31250 · from the original on | 30007 · to compare number with | 29362 · compare number with nil | 29362
5	archived from the original on | 29966 · to compare number with nil | 29362 · attempt to compare number with | 29362 · ಸುದ್ದಿಗಳಿಗಾಗಿ ಪ್ರಜಾವಾಣಿ ಆ್ಯಪ್ ಡೌನ್ಲೋಡ್ ಮಾಡಿಕೊಳ್ಳಿ | 24587 · ತಾಜಾ ಸುದ್ದಿಗಳಿಗಾಗಿ ಪ್ರಜಾವಾಣಿ ಆ್ಯಪ್ ಡೌನ್ಲೋಡ್ | 24587

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

Register labels

Name	Abbr.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	**HI**
Recipe	re
Informational persuasion	**IP**
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	**IN**
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	**OP**
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av