HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
tel_Telu.jsonl.tsv	9/17/2024	Telugu (te)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
2,058,193	39,190,209	19,411,118 (49.53 %)	1B	15.69 GB	6,468,174,327

Top 10 domains

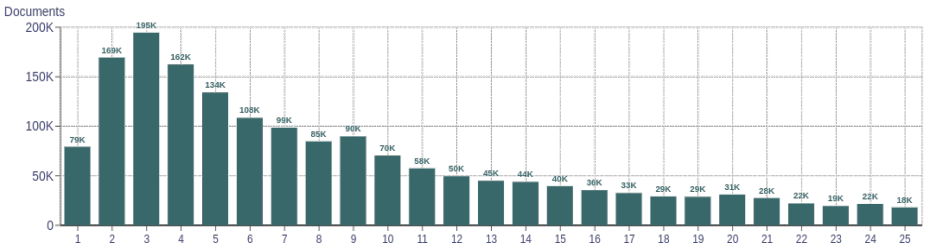
Domain	Docs	% of total
wikipedia.org	92K	4.48
blogspot.com	81K	3.92
andhrajyothy.com	62K	3.00
eenadu.net	58K	2.80
sakshi.com	53K	2.55
blogspot.in	47K	2.28
news18.com	46K	2.26
filmibeat.com	45K	2.21
asianetnews.com	35K	1.72
oneindia.com	30K	1.44

Top 10 TLDs

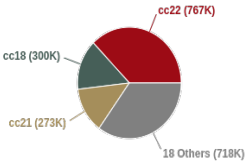
Domain	Docs	% of total
com	1.4M	69.77
in	187K	9.08
org	163K	7.92
net	138K	6.71
co.in	15K	0.71
info	13K	0.61
gov.in	9.6K	0.47
news	9.2K	0.45
page	8.6K	0.42
tv	8.6K	0.42

Documents size (in segments)

<= 25 segments **82.26%** (1.7M documents)
> 25 segments **17.74%** (365K documents)

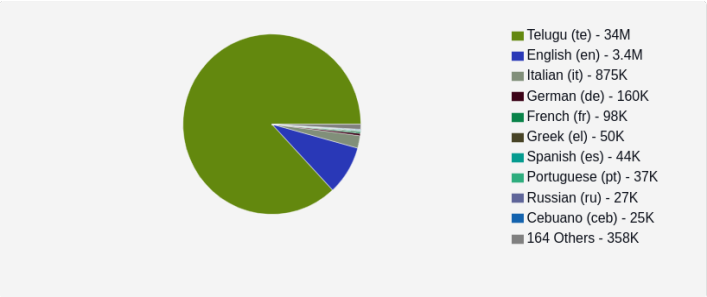


Documents by collection

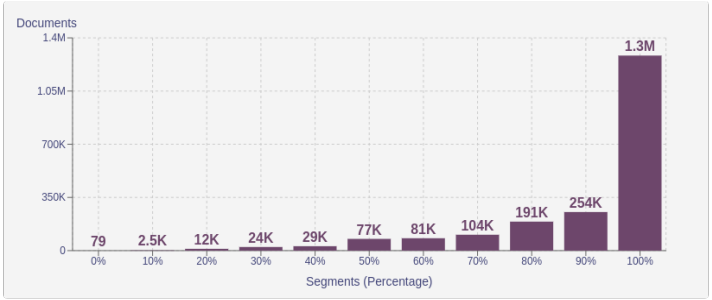


Language Distribution

Number of segments

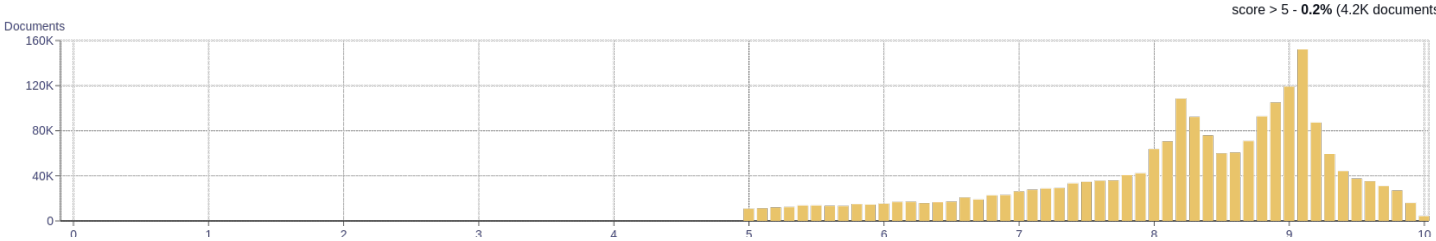


Percentage of segments in Telugu (te) inside documents



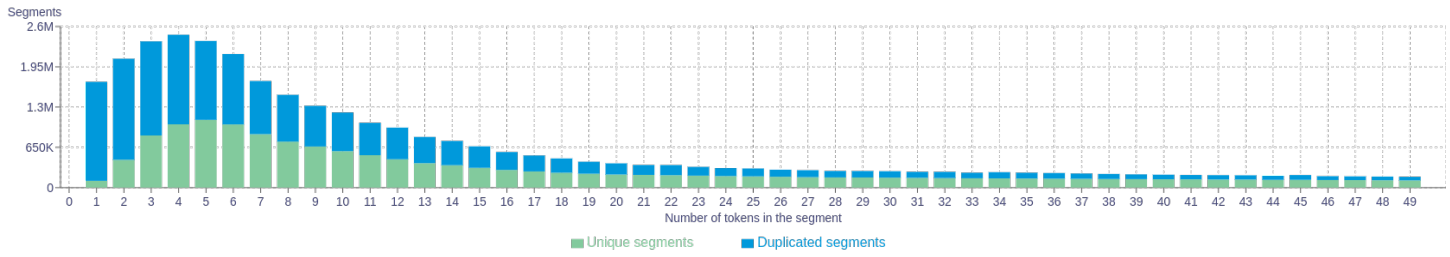
Distribution of documents by document score

score <= 5 - **99.8%** (2.1M documents)
score > 5 - **0.2%** (4.2K documents)

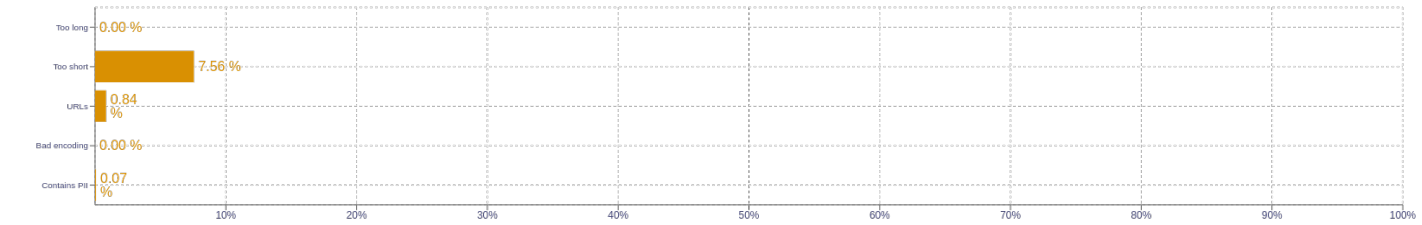


Segment length distribution by token

<= 49 tokens = **15M** segments | **18M** duplicates
> 50 tokens = **6.5M** segments | **2.2M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	ఈ | 8518633 · అ | 3929727 · అని | 2864260 · లో | 2814374 · తన | 2067602
2	ఈ సినిమా | 265875 · read more | 225178 · ఈ సందర్భంగా | 164946 · ఈ రోజు | 159926 · ఈ చిత్రం | 129845
3	from the original | 45157 · archived from the | 44441 · the original on | 40804 · ఎక్కువ మంది చదివినవి | 28153 · ఉత్పత్తులు లేదా సేవల | 24984
4	archived from the original | 44441 · from the original on | 40791 · వ్యాపార ప్రకటనలు వివిధ దేశాల్లోని | 24787 · విచారణ చేసి కొనుగోలు చేయాలి | 24787 · లోపాలకు ఈనాడు యాజమాన్యం బాధ్యత | 24787
5	archived from the original on | 40173 · వ్యాపార ప్రకటనలు వివిధ దేశాల్లోని వ్యాపారస్తులు | 24787 · లోపాలకు ఈనాడు యాజమాన్యం బాధ్యత వహించదు | 24787 · లేదా లోపాలకు ఈనాడు యాజమాన్యం బాధ్యత | 24787 · నెట్లో కనిపించే వ్యాపార ప్రకటనలు వివిధ | 24787

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