HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
ory_Orya.jsonl.tsv	9/22/2024	Odia (ory)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
412,895	3,595,619	2,426,565 (67.49 %)	121M	778,355,779	1.91 GB

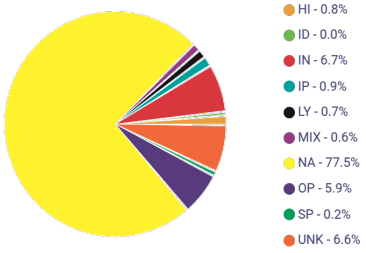
Top 10 domains

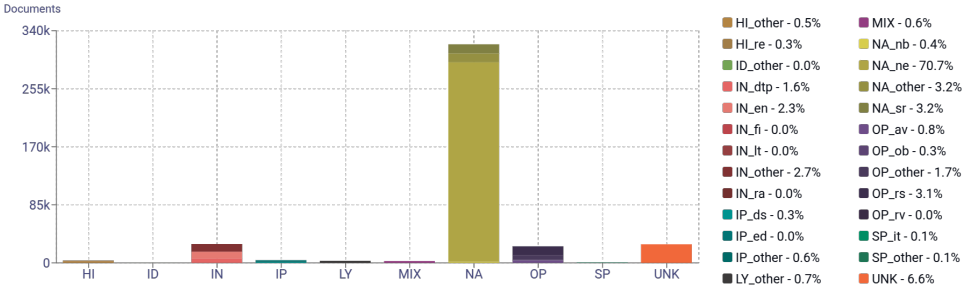
Domain	Docs	% of total
prameyanews7.com	39K	9.48%
sambad.in	39K	9.39%
kanaknews.com	30K	7.18%
dharitri.com	20K	4.95%
samajalive.in	19K	4.55%
odishareporter.in	17K	4.02%
news18.com	13K	3.14%
eodishasamachar...	10K	2.54%
wikipedia.org	10K	2.49%
prameya.com	5.7K	1.38%

Top 10 TLDs

Domain	Docs	% of total
com	254K	61.61%
in	123K	29.77%
org	24K	5.74%
me	4.3K	1.04%
live	1.8K	0.44%
net	1.2K	0.28%
nic.in	948	0.23%
gov.in	910	0.22%
is	562	0.14%
co.in	479	0.12%

Register labels

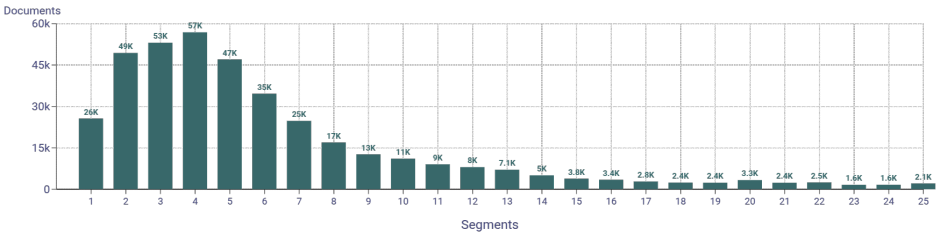




MT:1.7% | 7.1K Documents

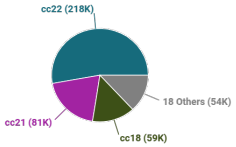
Documents size (in segments)

<= 25 segments **94.33%** (389K documents)
> 25 segments **5.67%** (23K documents)



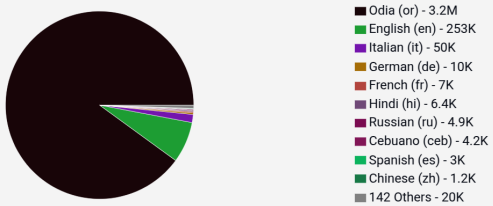
Documents by collection

CC = 91.93%
IA = 8.07%



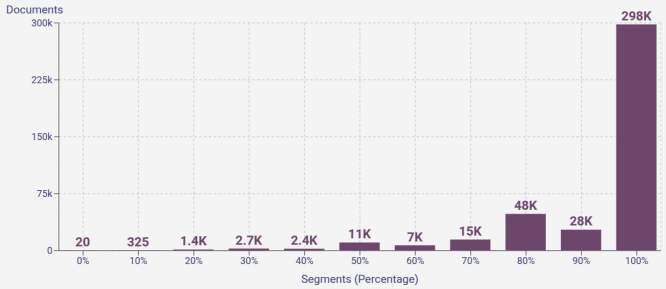
Language Distribution

Number of segments in the Odia (ory) corpus



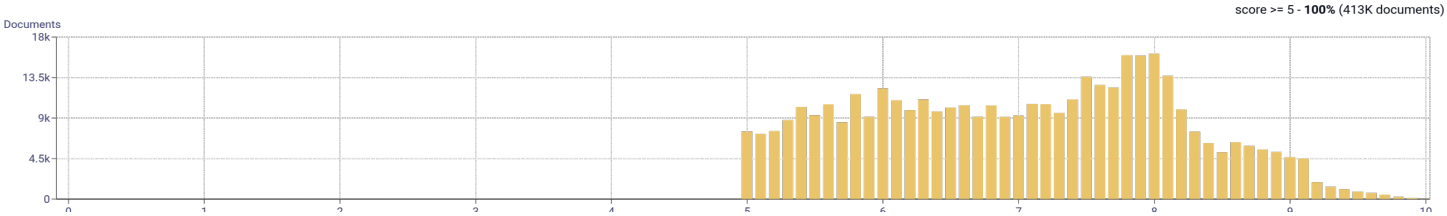
*Odia (ory) identification might be inaccurate because it is not supported by FastSpell

Percentage of segments in Odia (ory) inside documents



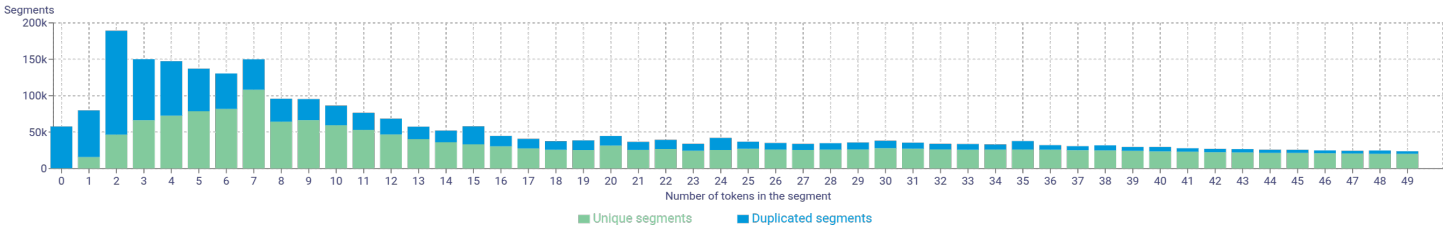
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (413K documents)

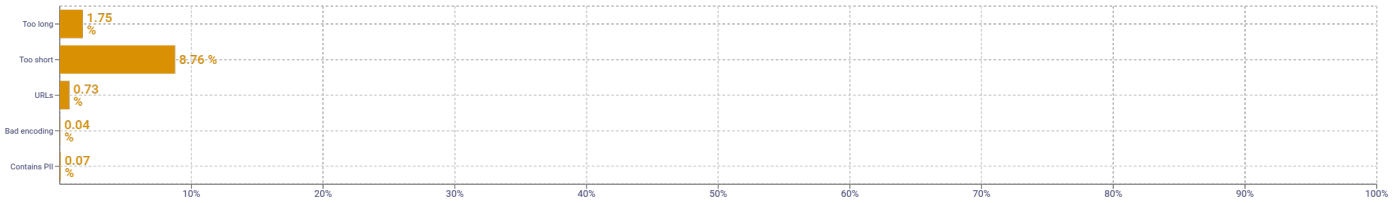


Segment length distribution by token

≤ **49** tokens = **1.7M** segments | **1M** duplicates
> **50** tokens = **836K** segments | **127K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	ସେ | 659801 · ଏକ | 600284 · ଜଣିବା | 421415 · ବୋଲି | 371775 · ଏହା | 312114
2	ଜଣିଛନ୍ତି ଯେ | 53427 · ରାଜ୍ୟ ସରକାର | 31309 · ହୋଇଥିବା ବେଳେ | 26936 · ସେ ଜଣିଛନ୍ତି | 26410 · ଜୋଡି ଟଡ୍‌କା | 25708
3	ମୁଖ୍ୟମନ୍ତ୍ରୀ ନବୀନ ପଟ୍ଟନାୟକ | 12409 · ଏଠାରେ କ୍ଲିକ୍ କରନ୍ତୁ । | 10140 · ପାଇବା ପାଇଁ ଏଠାରେ | 10137 · ସମସ୍ତ ତଡ ଖବର | 10125 · ରେ। ସମସ୍ତ ତଡ | 10125
4	ସମସ୍ତ ତଡ ଖବର ପାଇବା | 10125 · ରେ। ସମସ୍ତ ତଡ ଖବର | 10125 · ରିପୋର୍ଟର ଖବର ଏବେ ଟେଲିଗ୍ରାମ୍ | 10125 · ପାଇବା ପାଇଁ ଏଠାରେ କ୍ଲିକ୍ | 10125 · ପଢନ୍ତୁ ଓଡ଼ିଶା ରିପୋର୍ଟର ଖବର | 10125
5	ରେ। ସମସ୍ତ ତଡ ଖବର ପାଇବା | 10125 · ରିପୋର୍ଟର ଖବର ଏବେ ଟେଲିଗ୍ରାମ୍ ରେ। | 10125 · ତଡ ଖବର ପାଇବା ପାଇଁ ଏଠାରେ | 10125 · ପାଇବା ପାଇଁ ଏଠାରେ କ୍ଲିକ୍ କରନ୍ତୁ । | 10125 · ଟେଲିଗ୍ରାମ୍ ରେ। ସମସ୍ତ ତଡ ଖବର | 10125

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