

General overview

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

Volumes

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

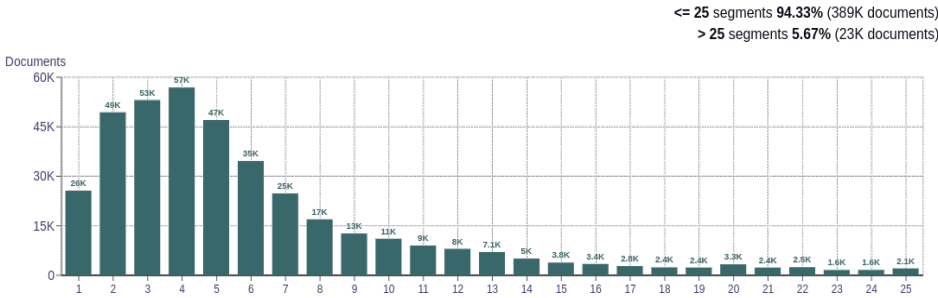
Top 10 domains

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

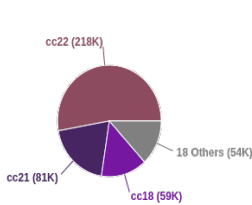
Top 10 TLDs

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

Documents size (in segments)

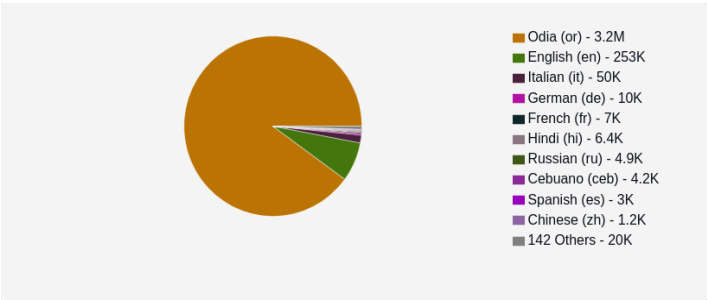


Documents by collection

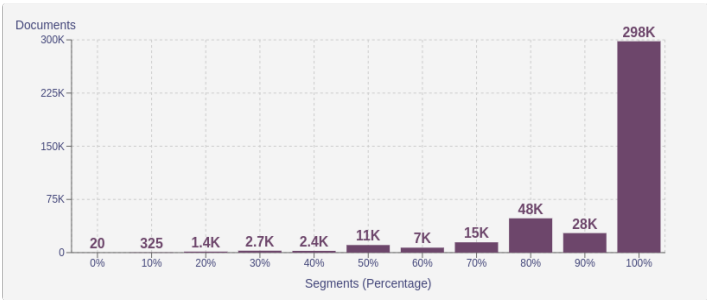


Language Distribution

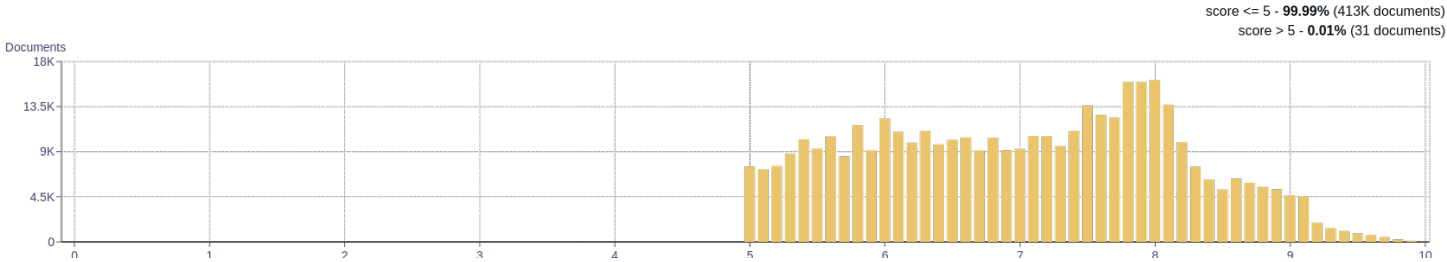
Number of segments



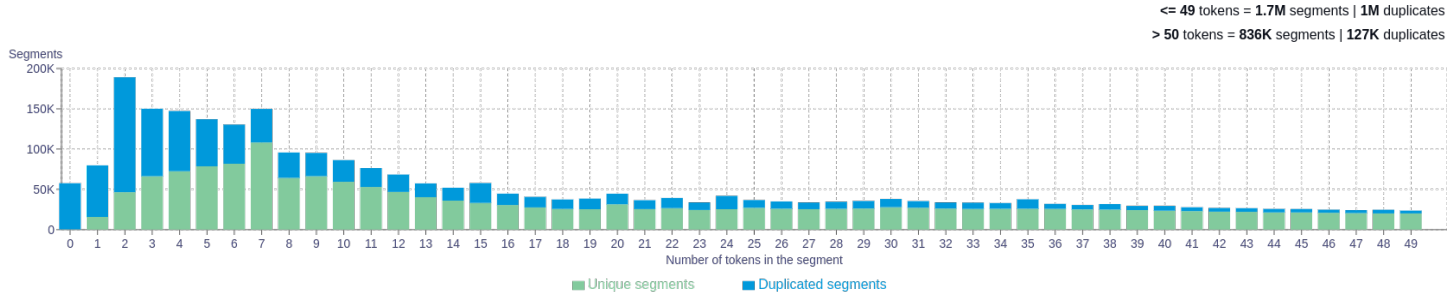
Percentage of segments in Odia (ory) inside documents



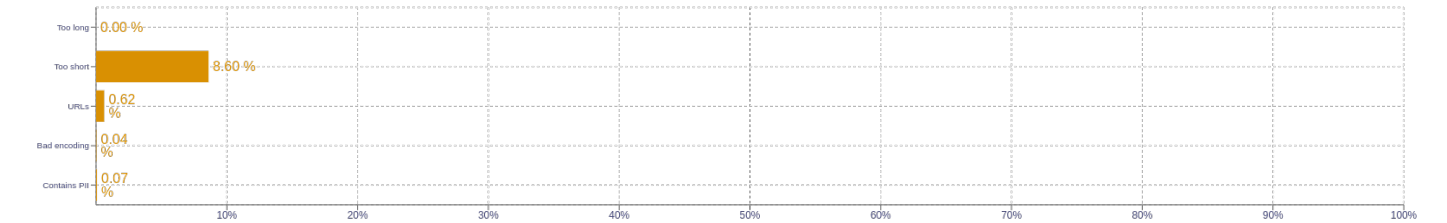
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>ସେ   659801</div> <div>ଏକ   600284</div> <div>କରିବା   421415</div> <div>ଗୋଲି   371775</div> <div>ଝା   312114</div>
2	<div>କହିଛନ୍ତି ଝା   53427</div> <div>ଭାର୍ଯ୍ୟ ସରମାର   31309</div> <div>ରୋଇସ୍ତବା ବେବେ   26936</div> <div>ସେ କହିଛନ୍ତି   26410</div> <div>ଲୋଡି ବଦ୍ଦା   25708</div>
3	<div>ପୁରୁଷମନ୍ତ୍ରୀ ନବୀନ ପଟ୍ଟନାୟକ   12409</div> <div>ଏଠାରେ ବୁଲିବୁ କରୁଛି ।   10140</div> <div>ସାଇବା ପାଇଁ ଏଠାରେ   10137</div> <div>ସମ୍ଭବ୍ତ ବଡ଼ ଖବର   10125</div> <div>ରେ। ସମ୍ଭବ୍ତ ବଡ଼   10125</div>
4	<div>ସମ୍ଭବ୍ତ ବଡ଼ ଖବର ସାଇବା   10125</div> <div>ରେ। ସମ୍ଭବ୍ତ ବଡ଼ ଖବର   10125</div> <div>ଭିସୋର୍ରେ ଖବର ଏବେ ଚେଲିଭୁରାମ୍   10125</div> <div>ସାଇବା ପାଇଁ ଏଠାରେ ବୁଲିବୁ   10125</div> <div>ପଢ଼ୁଛୁ ଓଡ଼ିଶା ଭିସୋର୍ରେ ଖବର   10125</div>
5	<div>ରେ। ସମ୍ଭବ୍ତ ବଡ଼ ଖବର ସାଇବା   10125</div> <div>ଭିସୋର୍ରେ ଖବର ଏବେ ଚେଲିଭୁରାମ୍ ରେ।   10125</div> <div>ବଡ଼ ଖବର ସାଇବା ପାଇଁ ଏଠାରେ   10125</div> <div>ସାଇବା ପାଇଁ ଏଠାରେ ବୁଲିବୁ କରୁଛି ।   10125</div> <div>ଚେଲିଭୁରାମ୍ ରେ। ସମ୍ଭବ୍ତ ବଡ଼ ଖବର   10125</div>

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 of 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>