HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
ne_1.jsonl.tsv	3/23/2024	Nepali (ne)

Volumes

Docs	Segments	Unique segments	Tokens	Size
863,349	78,392,421	43,583 (0.06 %)	795M	10.41 GB

Type-Token Ratio

Nepali (ne)
0.01

Top 10 domains

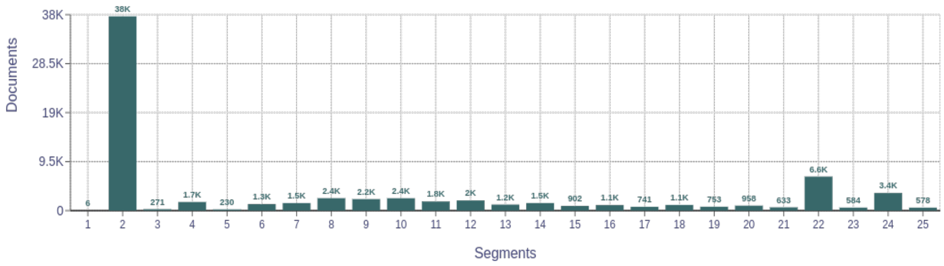
Domain	Docs	% of total
ekantipur.com	13K	1.49
nepal11news.com	5.3K	0.61
blogspot.co.il	4.9K	0.57
karobardaily.com	4.7K	0.55
setopati.com	4.5K	0.52
lokaantar.com	4.1K	0.47
everestdainik.com	4K	0.47
headlinenepal.com	3.8K	0.44
news24nepal.tv	3.8K	0.44
onlinekhabar.com	3.7K	0.43

Top 10 TLDs

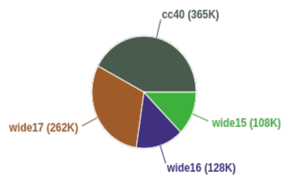
Domain	Docs	% of total
com	777K	90.06
com.np	15K	1.79
org	14K	1.68
net	11K	1.25
tv	7.4K	0.85
gov.np	6.3K	0.72
co.il	4.9K	0.57
kr	2.6K	0.30
com.au	2.4K	0.27
org.np	2.1K	0.25

Documents size (in segments)

<= 25 segments **8.5%** (73K documents)
> 25 segments **91.5%** (790K documents)

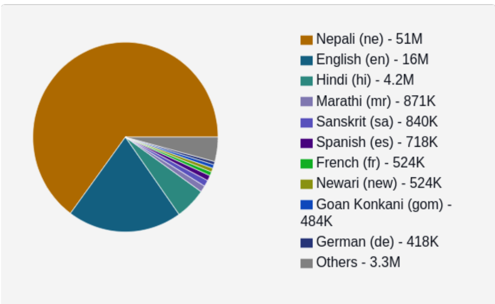


Documents by collection

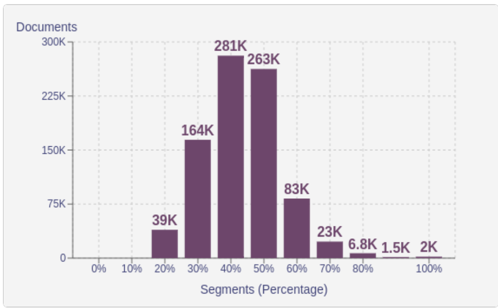


Language Distribution

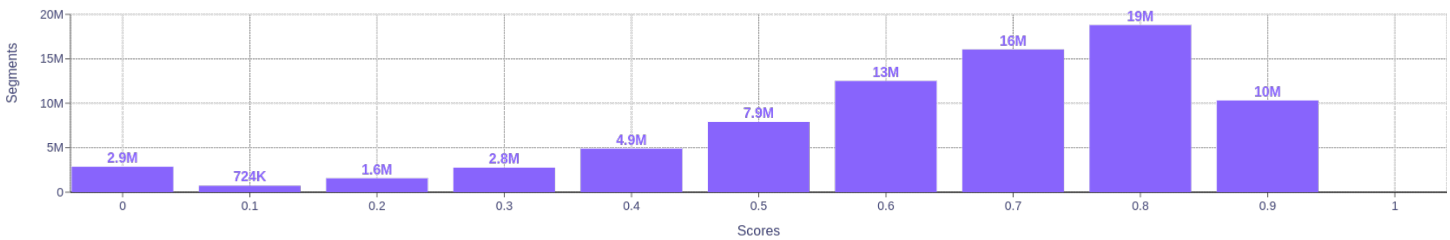
Number of segments



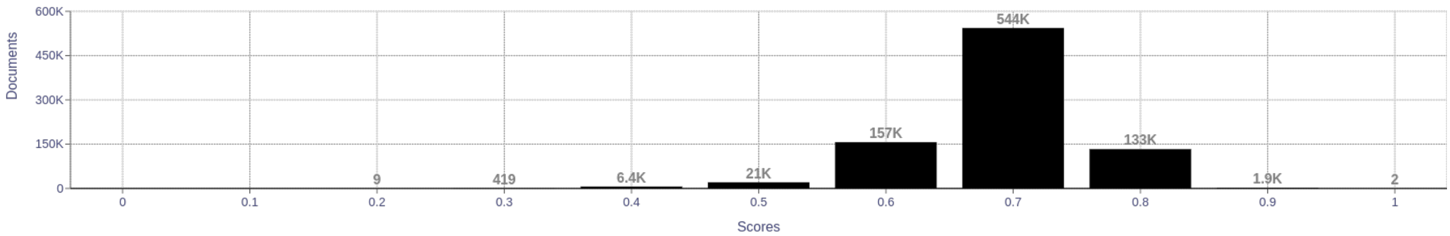
Percentage of segments in Nepali (ne) inside documents



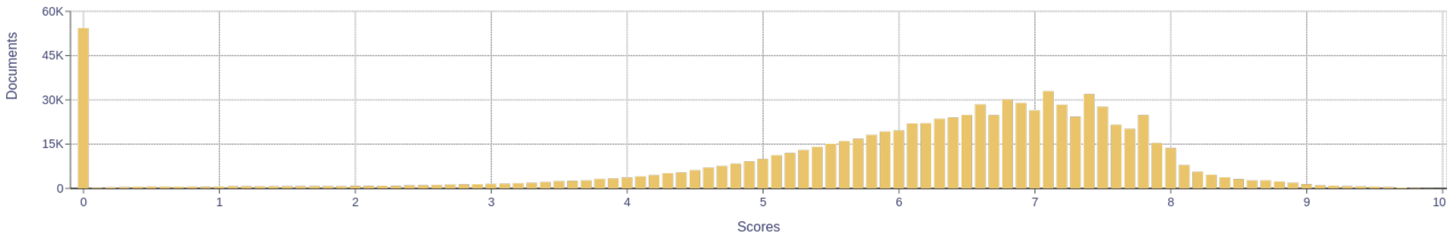
Distribution of segments by fluency score



Distribution of documents by average fluency score



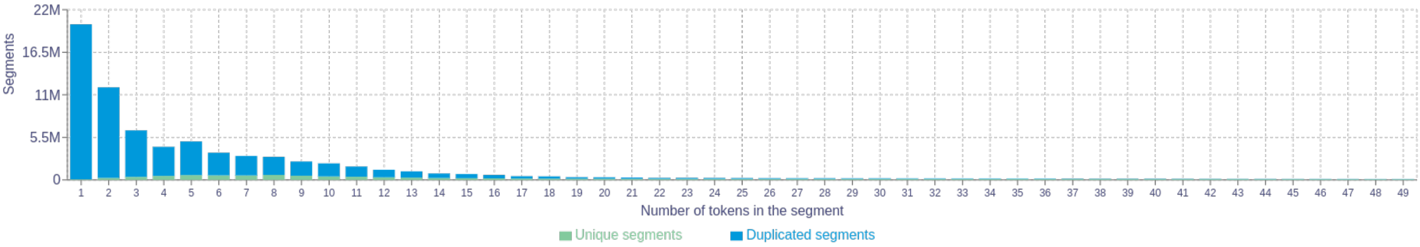
Distribution of documents by document score



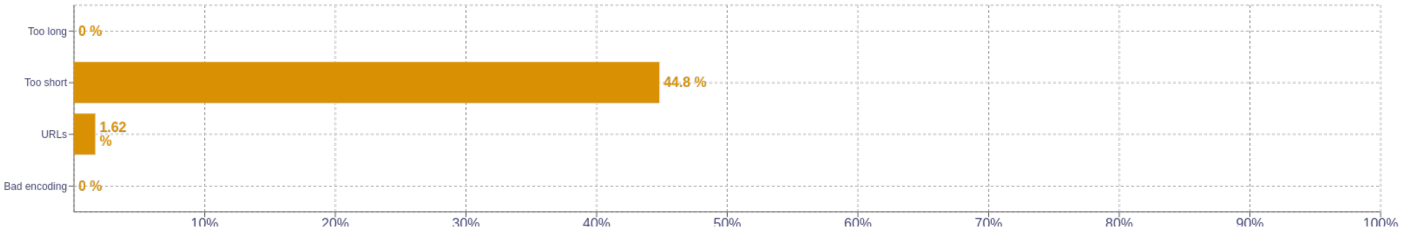
Segment length distribution by token

<= 49 tokens = **11M** segments | **65M** duplicates

> 50 tokens = **3.2M** segments | **696K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	समाचार | 2112597 · नेपाल | 1771404 · com | 1651501 · नेपाली | 1468742 · प्रदेश | 1274807
2	email protected | 325587 · rights reserved | 303236 · all rights | 303005 · read more | 292234 · हाम्रो बारेमा | 248969
3	all rights reserved | 300070 · सूचना विभाग दर्ता | 115639 · विभाग दर्ता नं | 104792 · leave a reply | 97600 · your email address | 94972
4	leave a reply cancel | 86228 · a reply cancel reply | 85758 · will not be published | 84327 · address will not be | 81861 · your email address will | 81827
5	leave a reply cancel reply | 85758 · address will not be published | 81854 · email address will not be | 81826 · your email address will not | 81824 · twittershare to facebookshare to pinterest | 59874

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