HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
npi_Deva.jsonl.tsv	9/25/2024	Nepali (npi)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
2,777,574	37,138,196	22,365,347 (60.22 %)	1.2B	17.88 GB	7,221,503,597

Top 10 domains

Domain	Docs	% of total
onlinekhabar.com	36K	1.31
abhiyan.com.np	29K	1.05
ekantipur.com	28K	1.00
ujyaaloonline.com	25K	0.91
ktmkhabar.com	25K	0.89
wikipedia.org	22K	0.81
setopati.com	22K	0.80
blogspot.com	21K	0.76
ratopati.com	19K	0.67
sajhasabal.com	18K	0.64

Top 10 TLDs

Domain	Docs	% of total
com	2.4M	87.77
com.np	107K	3.83
org	70K	2.52
net	31K	1.12
gov.np	31K	1.10
tv	26K	0.93
org.np	14K	0.49
com.au	6.7K	0.24
co.il	6K	0.22
news	5K	0.18

Documents size (in segments)

<= 25 segments **88.7%** (2.5M documents)
> 25 segments **11.3%** (314K documents)
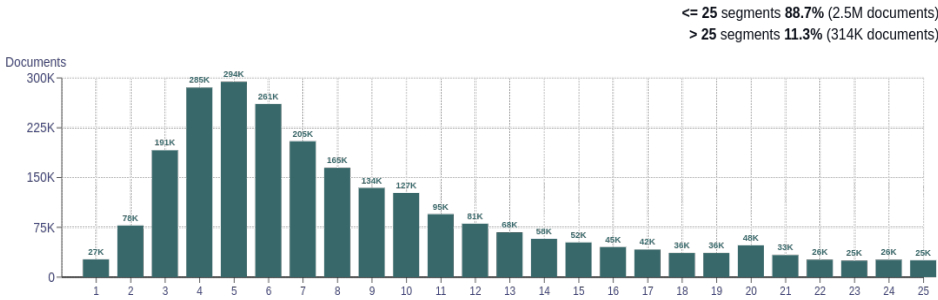
Documents by collection
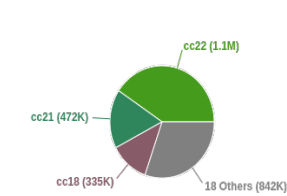
Language Distribution

Number of segments

- Nepali (ne) - 31M
- Hindi (hi) - 2.8M
- English (en) - 1.3M
- Italian (it) - 752K
- Marathi (mr) - 666K
- Sanskrit (sa) - 238K
- Newari (new) - 111K
- Dotyali (dty) - 61K
- French (fr) - 38K
- Maithili (mai) - 26K
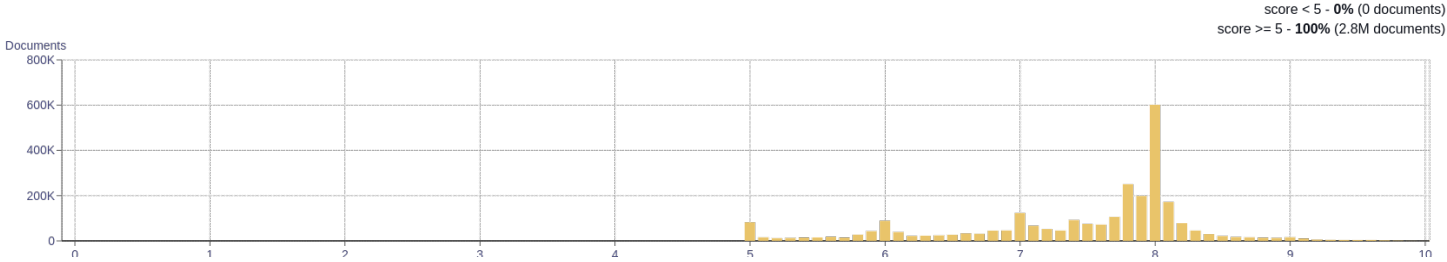- 158 Others - 262K
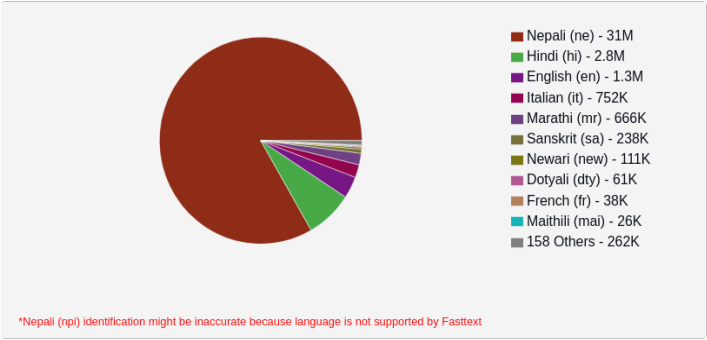
*Nepali (npi) identification might be inaccurate because language is not supported by Fasttext

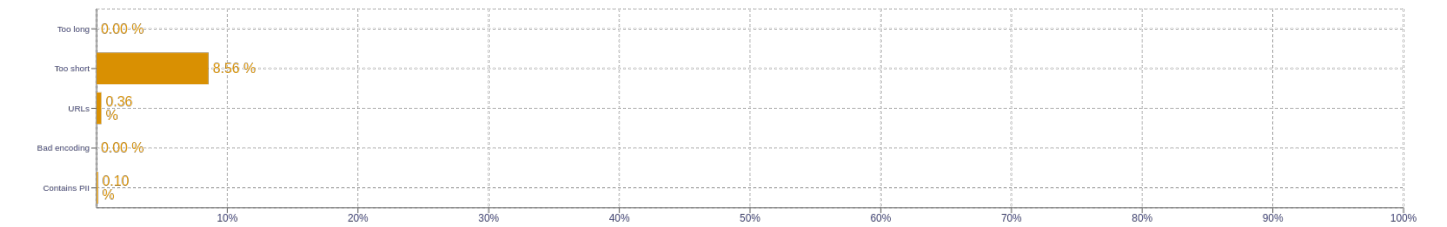### Percentage of segments in Nepali (npi) inside documents
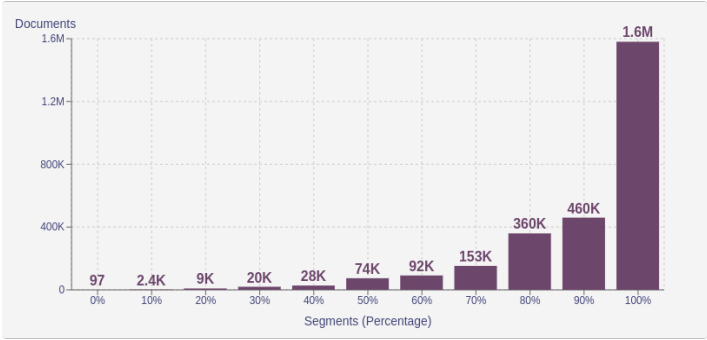
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (2.8M documents)

Segment length distribution by token

<= 49 tokens = **16M** segments | **13M** duplicates
> 50 tokens = **7.9M** segments | **1.9M** duplicates

Number of tokens in the segment

■ Unique segments ■ Duplicated segments

Segment noise distribution

Too long	0.00 %
Too short	8.56 %
URLs	0.36 %
Bad encoding	0.00 %
Contains PII	0.10 %
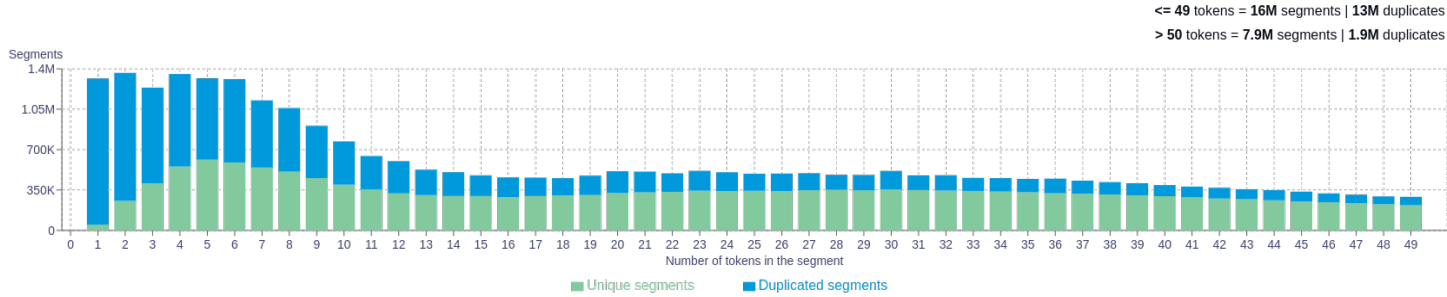
Frequent n-grams

Size	n-grams
1	नेपाल | 2495926 नेपाली | 2231383 गरिएको | 2186314 काम | 2120619 भएका | 1710736
2	जानकारी दिनुभयो | 188639 प्रहरी कार्यालय | 181360 वडा नं | 176503 read more | 176500 केपी शर्मा | 155024
3	जिल्ला प्रहरी कार्यालय | 100852 प्रधानमन्त्री केपी शर्मा | 85087 प्रमुख जिल्ला अधिकारी | 76609 केपी शर्मा ओलीले | 68217 अध्यक्ष पुष्पकमल दाहाल | 51563
4	प्रधानमन्त्री केपी शर्मा ओलीले | 41964 आवको सोही अवधिको तुलनामा | 25032 गत आवको सोही अवधिको | 24899 चालू आवको तेस्रो त्रैमासमा | 24806 लाख खुद मुनाफा आर्जन | 24724
5	गत आवको सोही अवधिको तुलनामा | 24853 चालू आवको तेस्रो त्रैमासमा बैङ्कको | 24579 बैङ्कको खराब कर्जा शून्य दशमलव | 24499 त्रैमासमा बैङ्कको खराब कर्जा शून्य | 24481 आवको तेस्रो त्रैमाससम्ममा बैङ्कले रु | 24478

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