HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
khm_Khmr.jsonl.tsv	9/26/2024	Khmer (km)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
700,992	9,864,172	5,800,059 (58.80 %)	1.5B	5.48 GB	2,113,232,593

Top 10 domains

Domain	Docs	% of total
khtoem.com	22K	3.16
voanews.com	21K	3.07
wordpress.com	18K	2.56
monoroom.info	16K	2.30
khmeread.com	14K	1.98
nokorwatnews.com	12K	1.78
freshnewsasia.com	11K	1.57
sabay.com.kh	10K	1.44
rasmeinews.com	9.2K	1.31
postkhmer.com	8.6K	1.23

Top 10 TLDs

Domain	Docs	% of total
com	419K	59.75
com.kh	58K	8.27
icu	38K	5.44
org	37K	5.31
gov.kh	34K	4.86
info	26K	3.67
net	17K	2.39
vn	13K	1.90
news	11K	1.61
org.kh	11K	1.52

Documents size (in segments)

<= 25 segments **88.59%** (621K documents)
> 25 segments **11.41%** (80K documents)
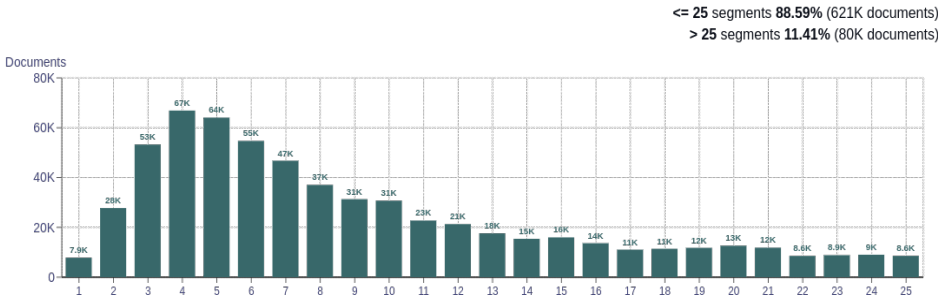
Documents by collection

cc22 (238K)
cc18 (92K)
cc21 (83K)
18 Others (287K)
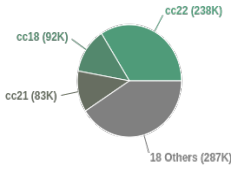
Language Distribution

Number of segments

- Khmer (km) - 9M
- English (en) - 505K
- Italian (it) - 148K
- German (de) - 28K
- French (fr) - 21K
- Dutch (nl) - 16K
- Thai (th) - 11K
- Spanish (es) - 9K
- Russian (ru) - 8.9K
- Chinese (zh) - 6.5K
- 154 Others - 76K
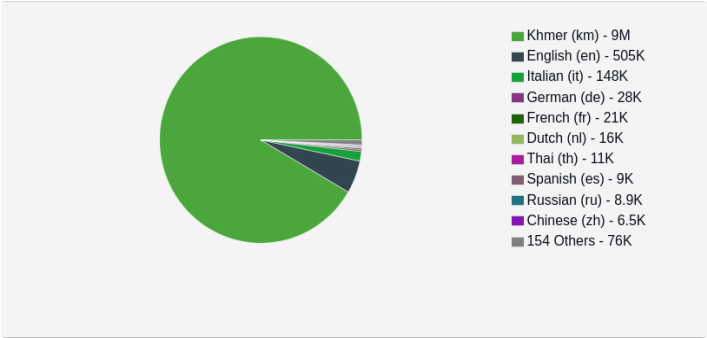
Percentage of segments in Khmer (km) inside documents
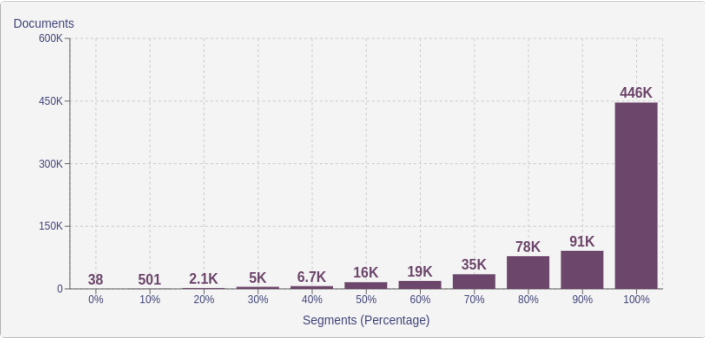
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (701K documents)

Segment length distribution by token

<= 49 tokens = **1.6M** segments | **2M** duplicates
> 50 tokens = **6.3M** segments | **2M** duplicates

- Unique segments
- Duplicated segments

Segment noise distribution

Too long	0.00 %
Too short	15.51 %
URLs	1.14 %
Bad encoding	0.00 %
Contains PII	0.20 %
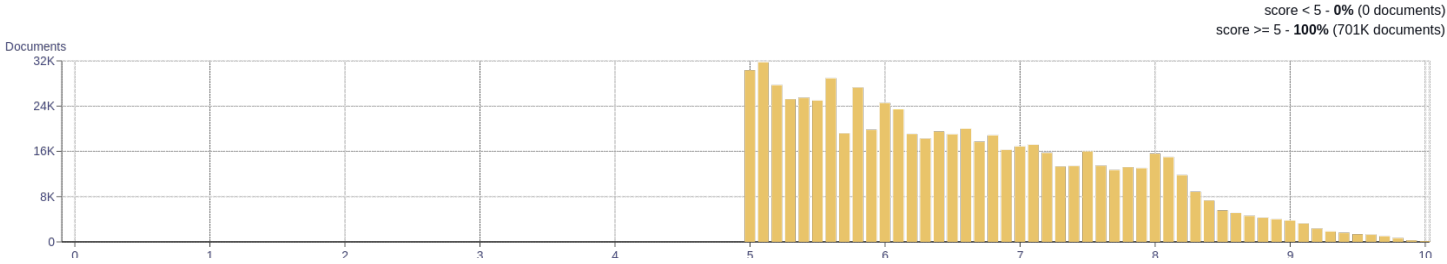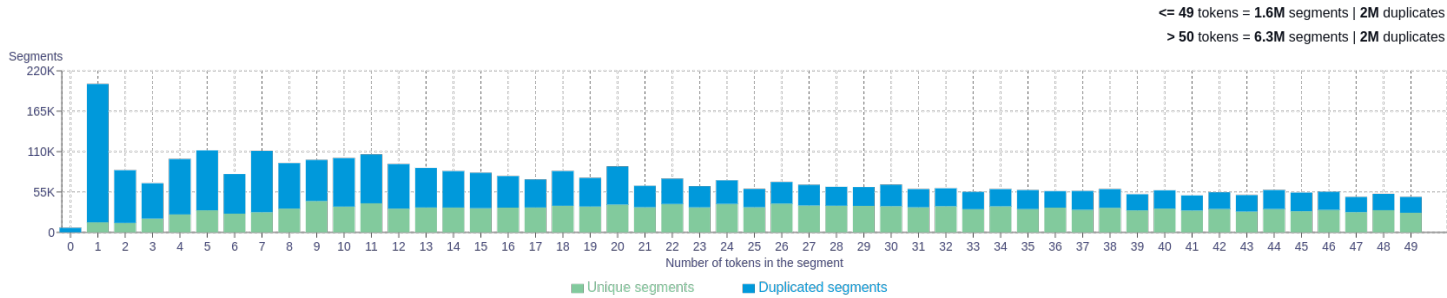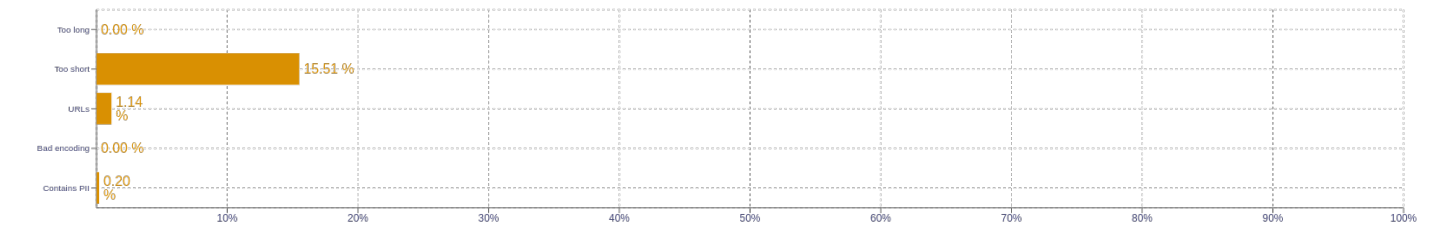
Frequent n-grams

Size	n-grams
1	ឌ | 47850531 ៖ | 40593396 ត | 35971657 ឥ | 31393488 ម | 31022979
2	ឌ ឌ | 503541 world cup | 502430 ៖ ឌ | 414159 ឌ ស | 407861 ម ឌ | 405408
3	fifa world cup | 45367 ម world cup | 45159 qatar world cup | 35895 ក សម មិ | 32137 world cup qatar | 31281
4	fifa world cup qatar | 15154 pro evolution soccer ស៩ | 7634 opens in new window | 5251 click to share on | 5250 ពពល world cup ឌ | 4833
5	to cnrp radio at national | 2547 read more khmer news and | 2547 radio at national rescue party | 2547 please read more khmer news | 2547 news and listen to cnrp | 2547

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