HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
lao_Laoo.jsonl.tsv	9/23/2024	Lao (lo)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
29,504	319,953	221,388 (69.19 %)	22M	221.5 MB	84,390,143

Top 10 domains

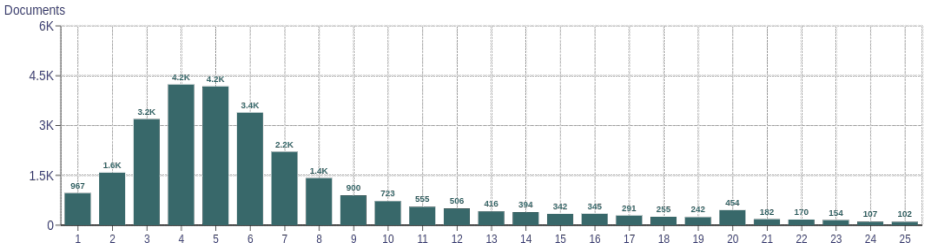
Domain	Docs	% of total
vnanet.vn	5.2K	17.55
vovworld.vn	2K	6.66
na.gov.la	1.3K	4.29
lnr.org.la	1.1K	3.77
lsr.com.la	981	3.32
thoidai.com.vn	941	3.19
wikipedia.org	709	2.40
tapchicongsan.org.vn	534	1.81
cri.cn	473	1.60
nuol.edu.la	358	1.21

Top 10 TLDs

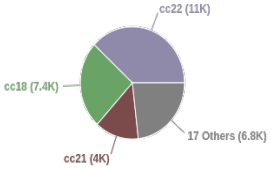
Domain	Docs	% of total
vn	7.5K	25.38
com	7.4K	25.08
gov.la	5.6K	19.11
org	2K	6.67
org.la	1.5K	5.04
com.la	1.4K	4.66
com.vn	941	3.19
edu.la	620	2.10
org.vn	534	1.81
cn	487	1.65

Documents size (in segments)

<= 25 segments **92.57%** (27K documents)
> 25 segments **7.43%** (2.2K documents)

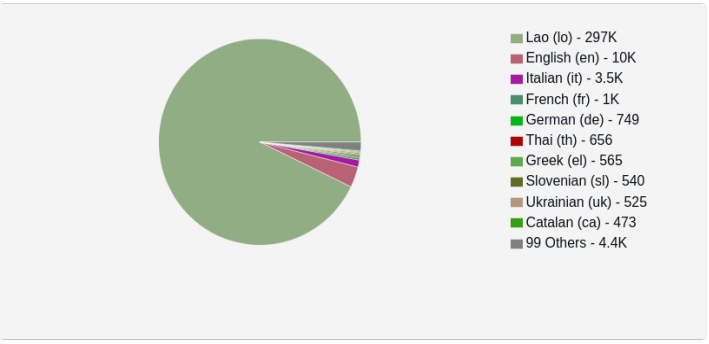


Documents by collection

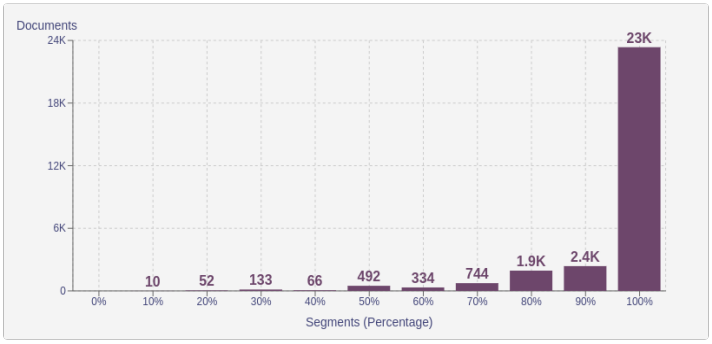


Language Distribution

Number of segments

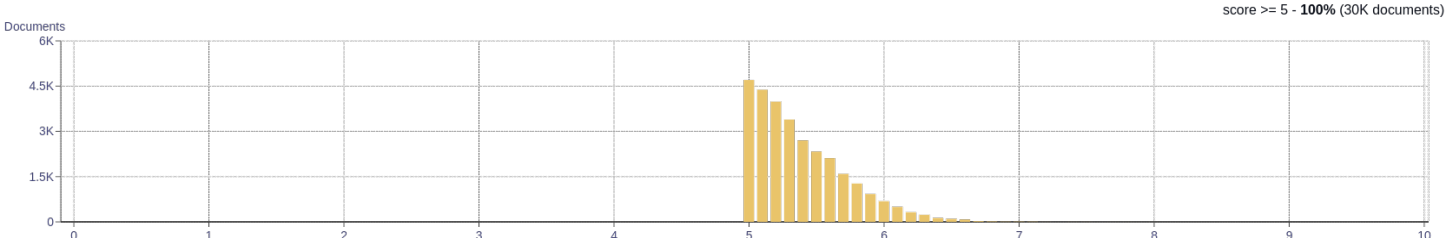


Percentage of segments in Lao (lo) inside documents



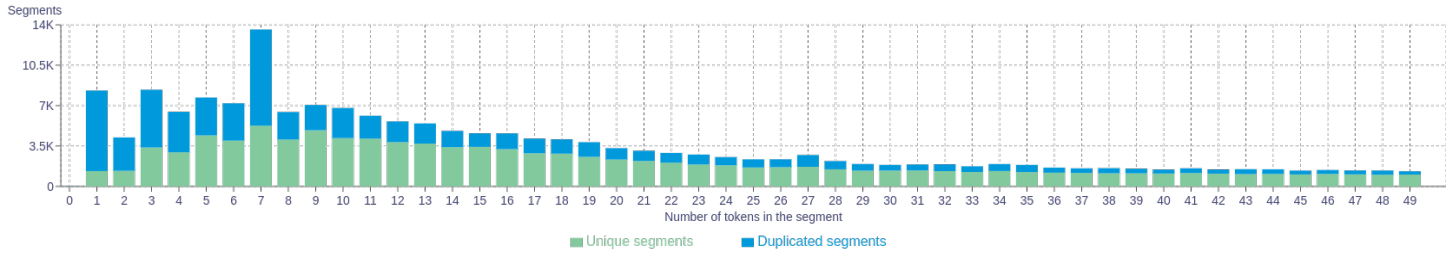
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (30K documents)

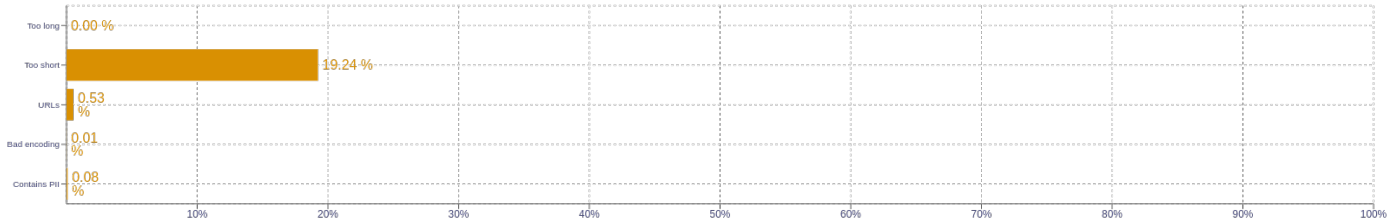


Segment length distribution by token

<= 49 tokens = **107K** segments | **70K** duplicates
> 50 tokens = **142K** segments | **28K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	ๆ | 156290 · ຫ່ານ | 129958 · ລາວ | 92003 · ແມ່ນ | 84606 · ນັກ | 74319
2	ກອງ ປະຊຸມ | 53088 · ສ້ ๆ | 42316 · ສະພາ ແຫ່ງຊາດ | 25340 · ນ໋ ๆ | 24536 · ๆ ລ້ຍ | 20624
3	ສ້ ๆ ລ້ຍ | 20609 · ສ້ ๆ ລ້ົມ | 8441 · ນ໋ ๆ ໃຊ້ | 8161 · ຄະນະ ບໍ່ລິຫານ ງານ | 7607 · ຈ໋ ๆ ນ່ວນ | 7230
4	ຫງ ວຽ້ນ ຊວນ ຝຸກ | 3904 · ຄະນະ ບໍ່ລິຫານ ງານ ສູນກາງ | 3721 · ແນວ ລາວ ສ້ ງ ຊາດ | 3676 · ບໍ່ລິຫານ ງານ ສູນກາງ ນັກ | 3156 · ກົມ ການເມືອງ ສູນກາງ ນັກ | 2971
5	ຄະນະ ບໍ່ລິຫານ ງານ ສູນກາງ ນັກ | 3083 · ກອງ ປະຊຸມ ໃຫ່ ຍ ັ ງ ຮີ | 1614 · ສູນກາງ ແນວ ລາວ ສ້ ງ ຊາດ | 1590 · ຄະນະ ໂຄສະນາ ອົບຮົມ ສູນກາງ ນັກ | 1588 · ກໍ ານະການ ກົມ ການເມືອງ ສູນກາງ ນັກ | 1328

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