

General overview

Corpus	Analytics date	Language
lao_Lao.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

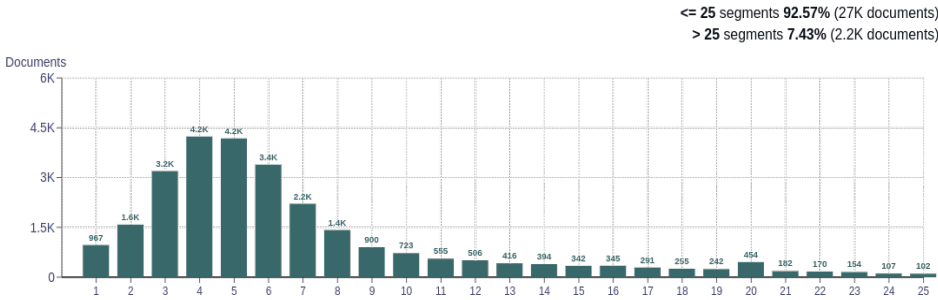
Top 10 domains

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

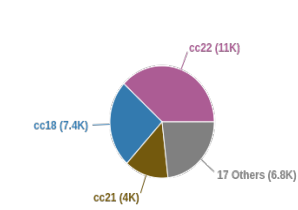
Top 10 TLDs

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)

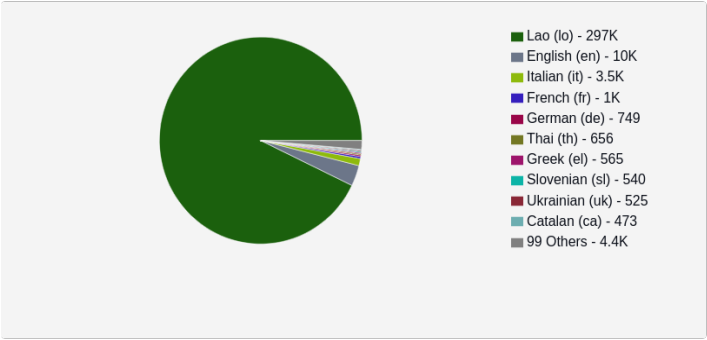


Documents by collection

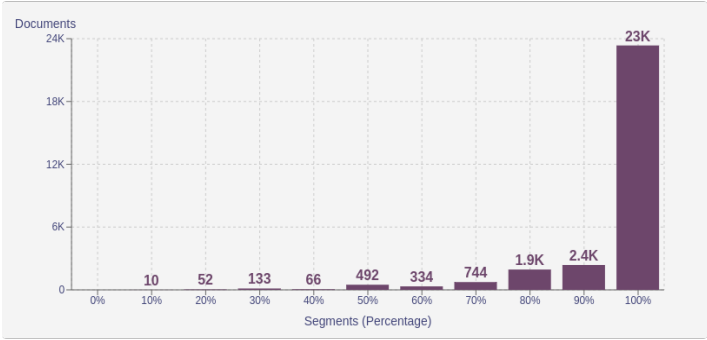


Language Distribution

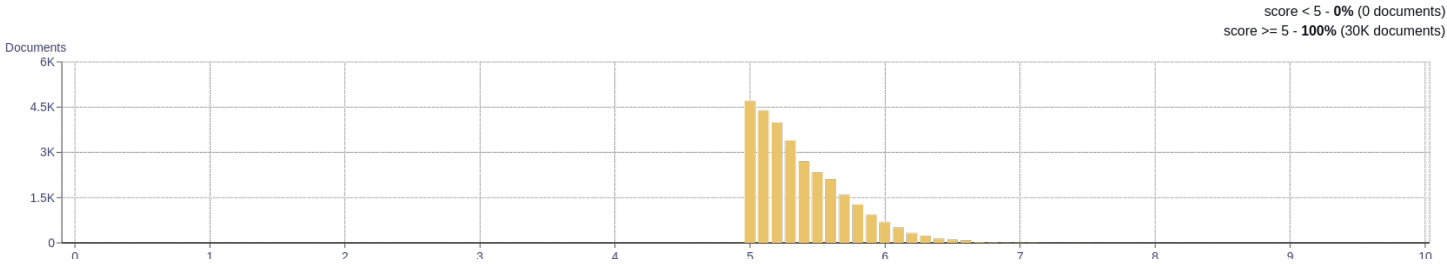
Number of segments



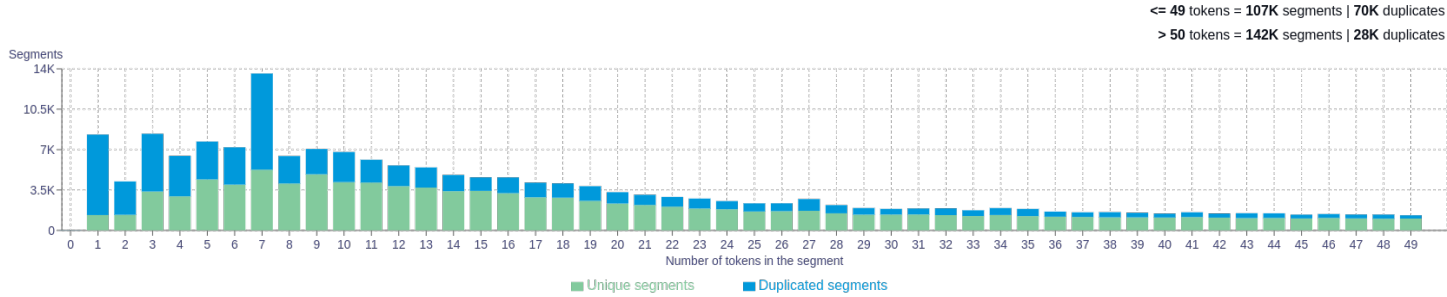
Percentage of segments in Lao (lo) inside documents



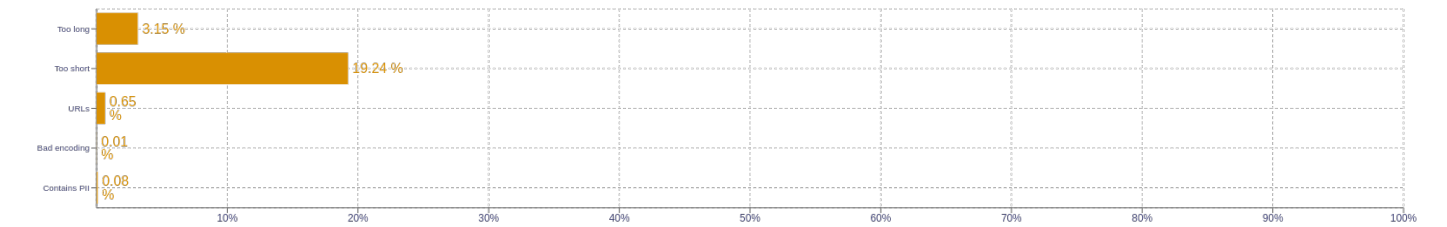
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div><div>າ</div><div> </div><div>156290</div><div><div>ສ້າງ</div><div> </div><div>129958</div></div><div><div>ລາວ</div><div> </div><div>92003</div></div><div><div>ແມ່ນ</div><div> </div><div>84606</div></div><div><div>ພັກ</div><div> </div><div>74319</div></div></div>
2	<div><div>ກອງ ປະຊຸມ</div><div> </div><div>53088</div></div> <div><div>ສ່ຳ າ</div><div> </div><div>42316</div></div> <div><div>ສະພາ ແຫ່ງຊາດ</div><div> </div><div>25340</div></div> <div><div>ນໍ າ</div><div> </div><div>24536</div></div> <div><div>າ ພັຍ</div><div> </div><div>20624</div></div>
3	<div><div>ສ່ າ ພັຍ</div><div> </div><div>20609</div></div> <div><div>ສ່ າ ພັຍ</div><div> </div><div>8441</div></div> <div><div>ນໍ າ ໃຊ້</div><div> </div><div>8161</div></div> <div><div>ຄະນະ ບໍລິຫານ ງານ</div><div> </div><div>7607</div></div> <div><div>ຈໍ າ ນວນ</div><div> </div><div>7230</div></div>
4	<div><div>ຫງ ວຽນ ຊວນ ຟຸດ</div><div> </div><div>3904</div></div> <div><div>ຄະນະ ບໍລິຫານ ງານ ສູນກາງ</div><div> </div><div>3721</div></div> <div><div>ແນວ ລາວ ສ້າງ ຊາດ</div><div> </div><div>3676</div></div> <div><div>ບໍລິຫານ ງານ ສູນກາງ ພັກ</div><div> </div><div>3156</div></div> <div><div>ກົມ ການເມືອງ ສູນກາງ ພັກ</div><div> </div><div>2971</div></div>
5	<div><div>ຄະນະ ບໍລິຫານ ງານ ສູນກາງ ພັກ</div><div> </div><div>3083</div></div> <div><div>ກອງ ປະຊຸມ ໂຕຍໍ ສັງ ທິ</div><div> </div><div>1614</div></div> <div><div>ສູນກາງ ແນວ ລາວ ສ້າງ ຊາດ</div><div> </div><div>1590</div></div> <div><div>ຄະນະ ໂຄສະນາ ອົບຮົມ ສູນກາງ ພັກ</div><div> </div><div>1588</div></div> <div><div>ກຳມະການ ກົມ ການເມືອງ ສູນກາງ ພັກ</div><div> </div><div>1328</div></div>

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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>, , , 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>