HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
kab_Latn.jsonl.tsv	9/20/2024	Kabyle (kab)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
15,100	345,218	181,035 (52.44 %)	13M	54.55 MB	53,861,233

Top 10 domains

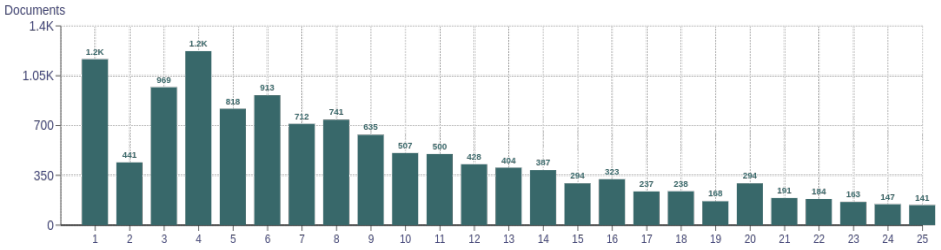
Domain	Docs	% of total
wikipedia.org	4K	26.56
tamurt.info	1.6K	10.77
studybible.info	1.1K	7.48
lerifain.fr	1.1K	7.14
aps.dz	490	3.25
jw.org	323	2.14
depechedekabylie.com	288	1.91
tfmpage.com	278	1.84
amaynu.net	234	1.55
siwel.info	174	1.15

Top 10 TLDs

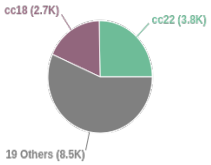
Domain	Docs	% of total
org	5K	33.22
com	3.5K	22.95
info	3.1K	20.73
fr	1.3K	8.43
net	970	6.42
dz	662	4.38
nl	155	1.03
mx	79	0.52
de	79	0.52
com.ar	45	0.30

Documents size (in segments)

<= 25 segments **80.95%** (12K documents)
> 25 segments **19.05%** (2.9K documents)

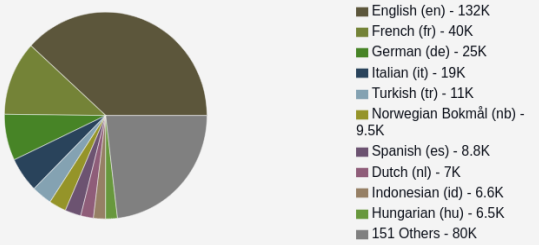


Documents by collection



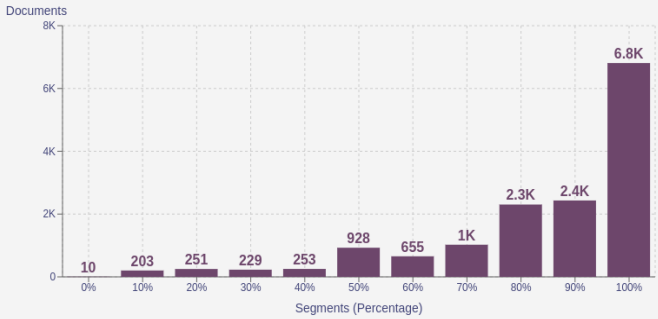
Language Distribution

Number of segments



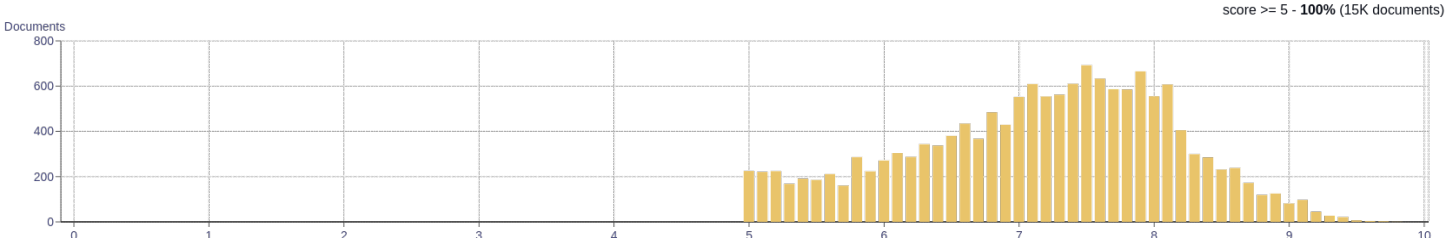
*Kabyle (kab) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Kabyle (kab) inside documents



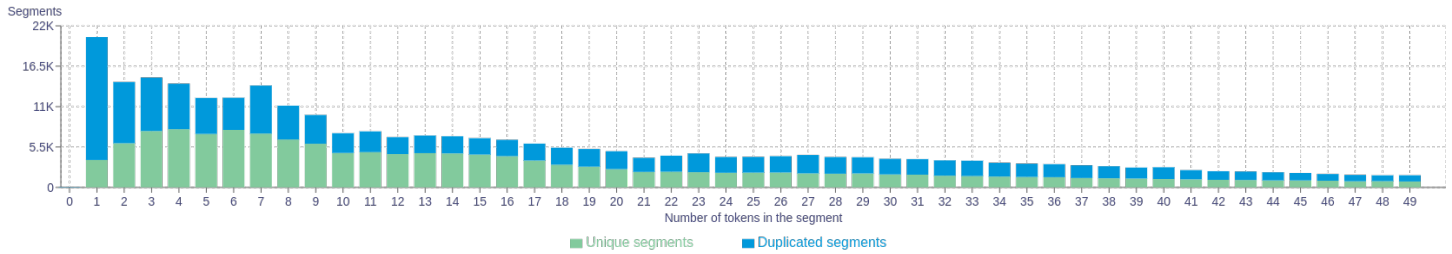
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (15K documents)

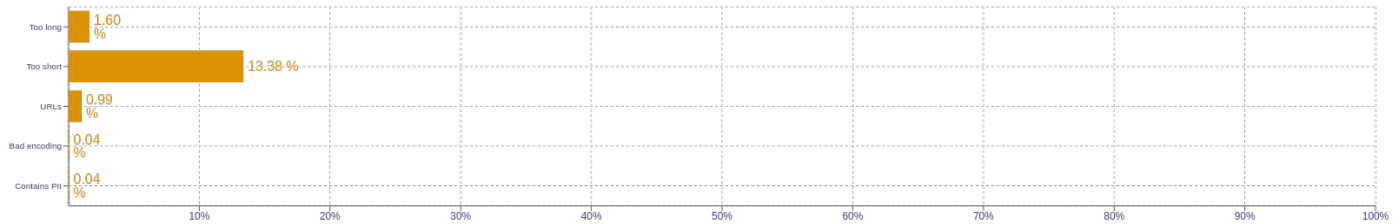


Segment length distribution by token

<= 49 tokens = **147K** segments | **139K** duplicates
> 50 tokens = **59K** segments | **25K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	d | 561188 · n | 545429 · s | 255762 · i | 253451 · a | 129347
2	s s | 61700 · i d | 39676 · d d | 36365 · yes yes | 26700 · sidi ṛebbi | 20990
3	s s s | 57631 · d d d | 32425 · yes yes yes | 26640 · yus yus yus | 10800 · n sidi ṛebbi | 6722
4	s s s s | 55112 · d d d d | 31467 · yes yes yes yes | 26582 · yus yus yus yus | 10452 · n n n n | 4901
5	s s s s s | 53136 · d d d d d | 30821 · yes yes yes yes yes | 26524 · yus yus yus yus yus | 10104 · k k k k k | 4741

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