

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

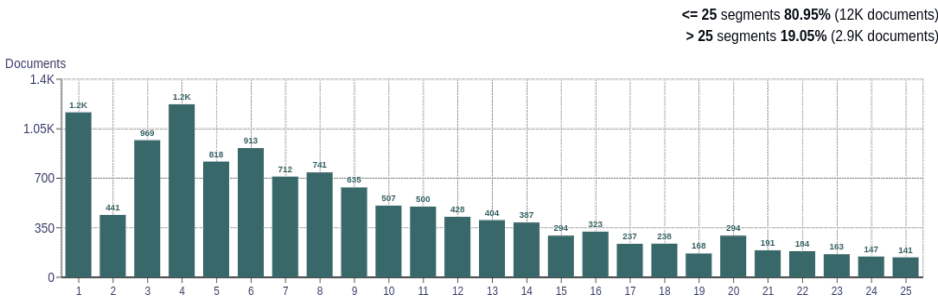
Top 10 domains

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
tftpape.com	278	1.84
amaynu.net	234	1.55
siwel.info	174	1.15

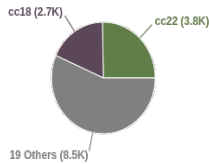
Top 10 TLDs

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)

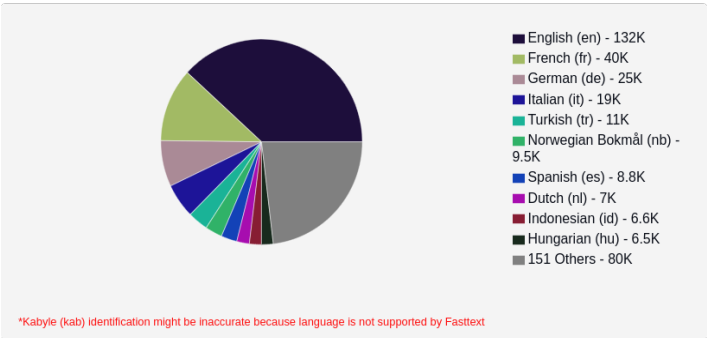


Documents by collection

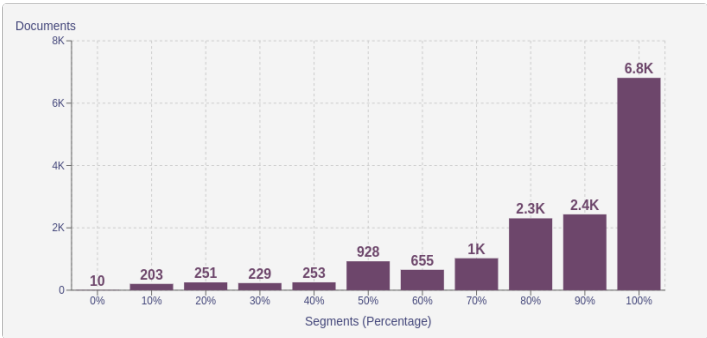


Language Distribution

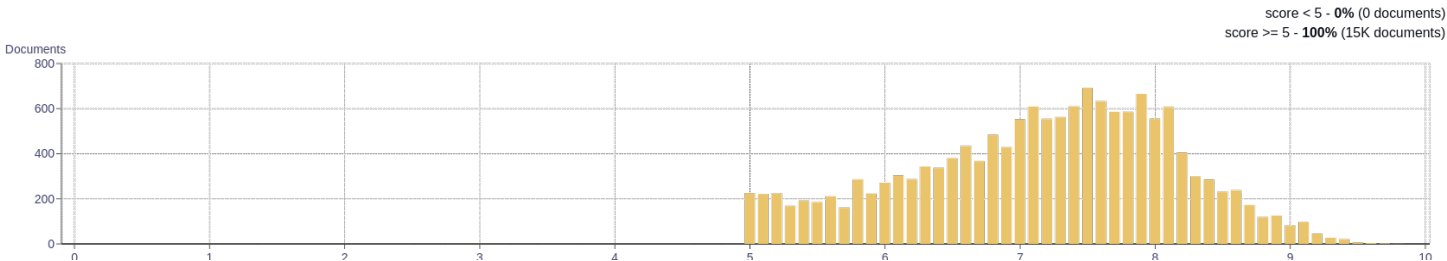
Number of segments



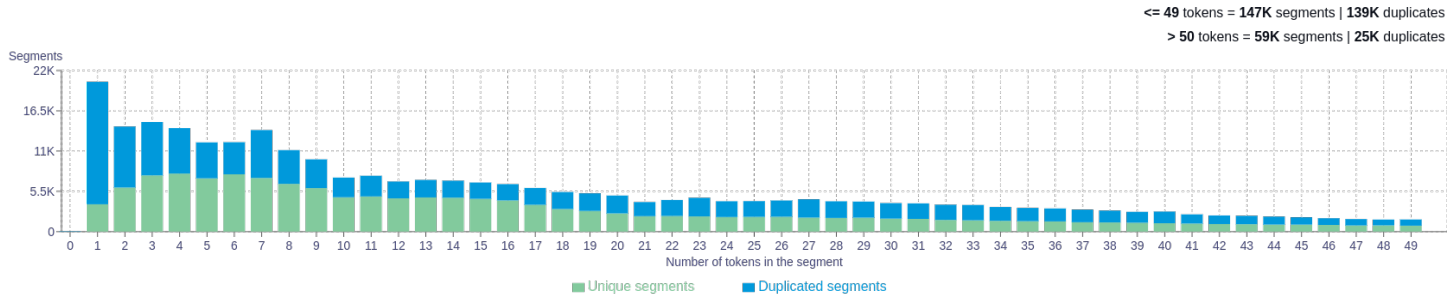
Percentage of segments in Kabyle (kab) inside documents



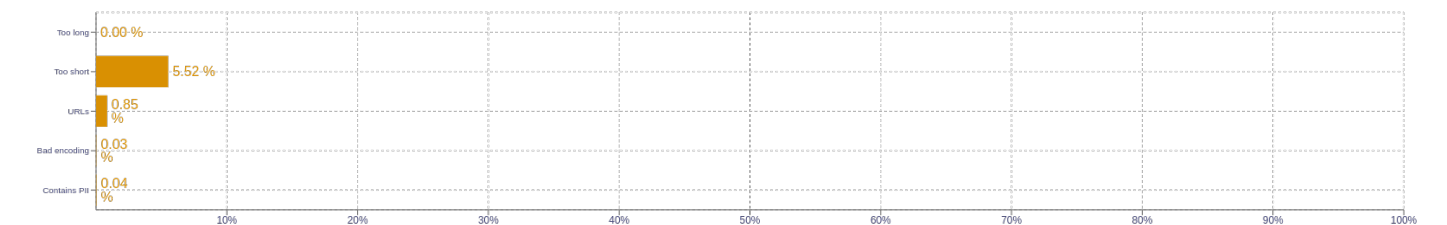
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

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