HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
kea_Latn.jsonl.tsv	11/4/2024	Kabuverdianu (kea)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
1,962	43,911	26,764 (60.95 %)	1.3M	5.96 MB	6,102,519

Top 10 domains

Domain	Docs	% of total
dypk-portal.com	549	27.98
jw.org	292	14.88
dexamsabi.com	240	12.23
blogspot.com	196	9.99
deltacultura.org	89	4.54
blogspot.pt	37	1.89
santiagomagazine.cv	26	1.33
blogspot.jp	20	1.02
kriolita.com	19	0.97
letras.mus.br	18	0.92

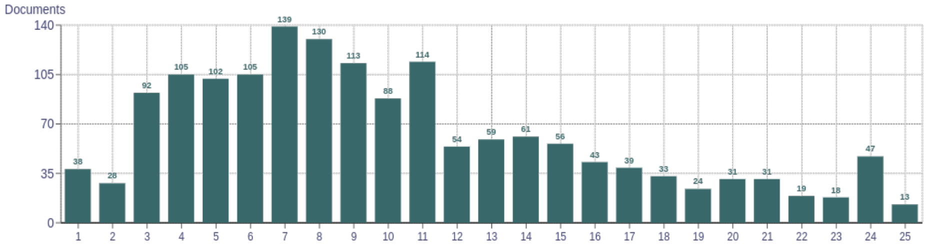
Top 10 TLDs

Domain	Docs	% of total
com	1.2K	61.26
org	440	22.43
cv	74	3.77
com.br	56	2.85
pt	46	2.34
jp	20	1.02
mus.br	18	0.92
gov	16	0.82
info	16	0.82
biz	14	0.71

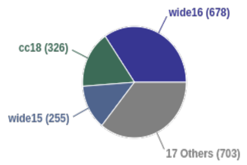
Documents size (in segments)

<= 25 segments **80.63%** (1.6K documents)

> 25 segments **19.37%** (378 documents)

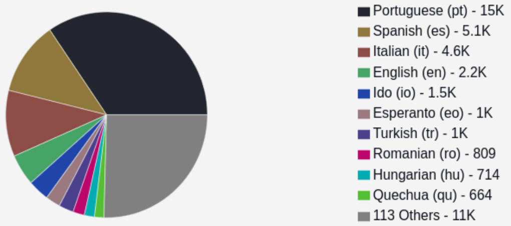


Documents by collection



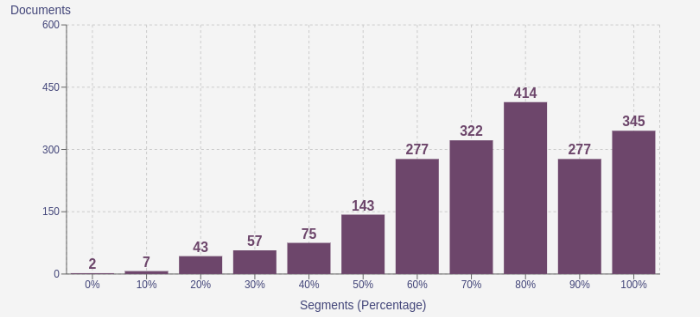
Language Distribution

Number of segments



*Kabuverdianu (kea) identification might be inaccurate because language is not supported by Fasttext

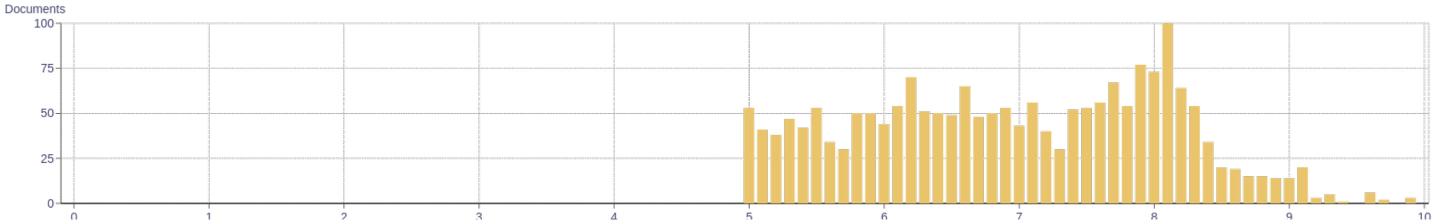
Percentage of segments in Kabuverdianu (kea) inside documents



Distribution of documents by document score

score < 5 - **0%** (0 documents)

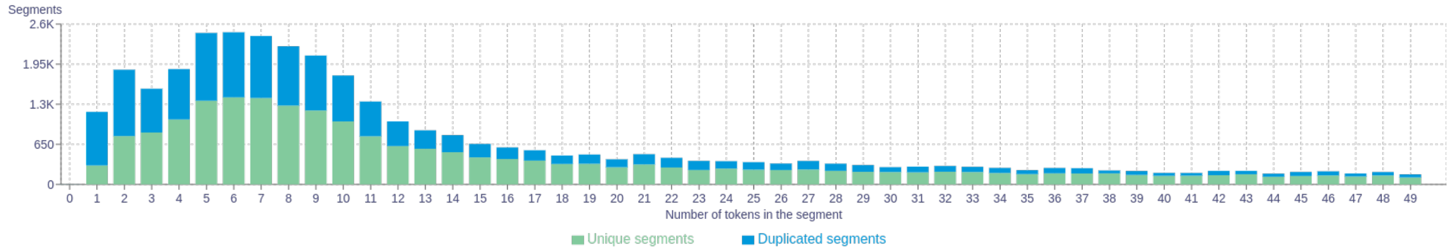
score >= 5 - **100%** (2K documents)



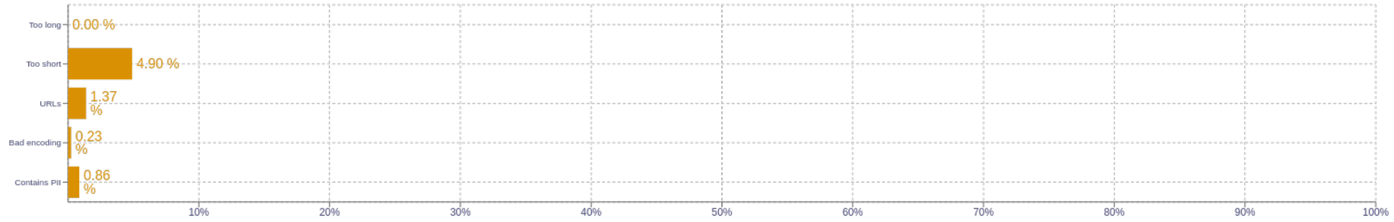
Segment length distribution by token

<= 49 tokens = **21K** segments | **14K** duplicates

> 50 tokens = **8.8K** segments | **2.9K** duplicates



Segment noise distribution



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