HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
kon_Latn.jsonl.tsv	9/21/2024	Kongo (kg)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
2,542	47,477	30,148 (63.50 %)	2.4M	10.82 MB	11,229,852

Top 10 domains

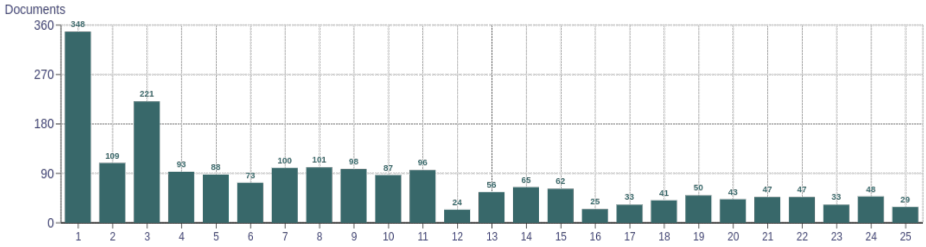
Domain	Docs	% of total
jw.org	1.9K	75.37
wikipedia.org	302	11.88
radiookapi.net	94	3.70
grindr.com	28	1.10
afrikblog.com	22	0.87
contafrica.org	15	0.59
wizi-kongo.com	12	0.47
gotquestions.org	11	0.43
watchtower.org	10	0.39
radiokongodieto.com	6	0.24

Top 10 TLDs

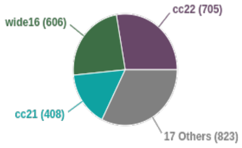
Domain	Docs	% of total
org	2.3K	90.09
com	125	4.92
net	111	4.37
co	3	0.12
info	3	0.12
eu	2	0.08
ch	2	0.08
click	2	0.08
cz	1	0.04
co.uk	1	0.04

Documents size (in segments)

<= 25 segments **79.35%** (2K documents)
> 25 segments **20.65%** (524 documents)

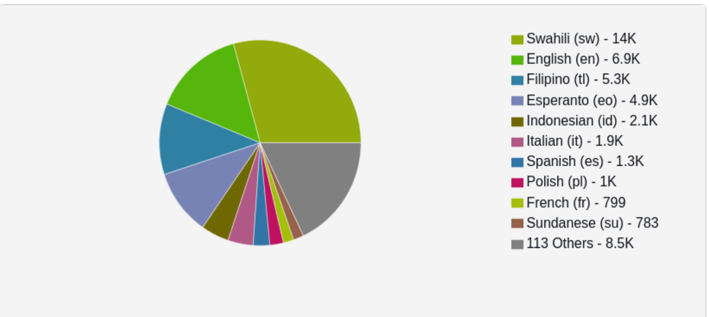


Documents by collection



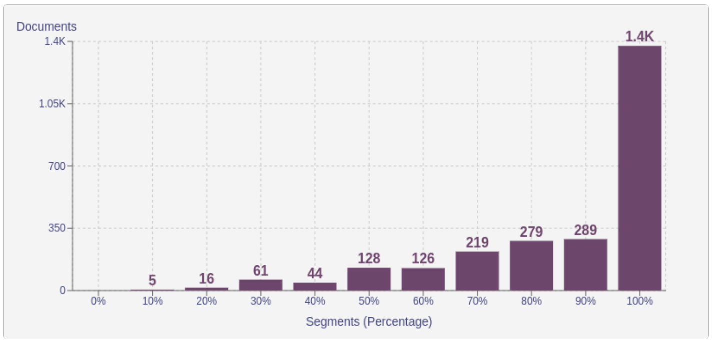
Language Distribution

Number of segments



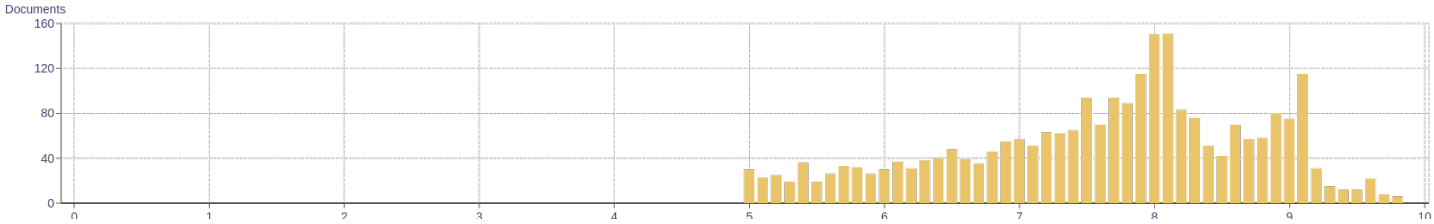
*Kongo (kg) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Kongo (kg) inside documents



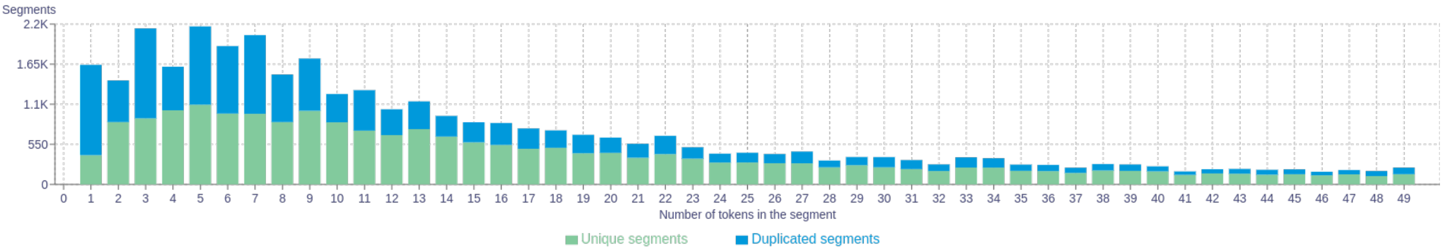
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (2.5K documents)

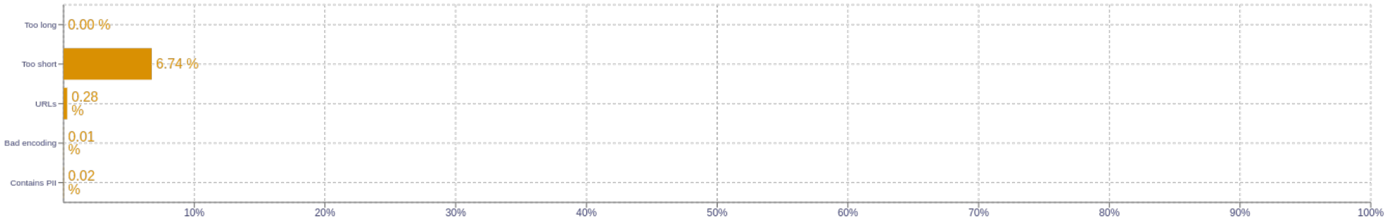


Segment length distribution by token

<= 49 tokens = **21K** segments | **15K** duplicates
> 50 tokens = **12K** segments | **2.7K** duplicates



Segment noise distribution



Frequent n-grams

<table>
<tr><th>Size</th><th>n-grams</th></tr>
<tr><td>1</td><td>nzambi | 20030
bantu | 15995
ve | 15470
mambu | 10925
yehowa | 8945</td></tr>
<tr><td>2</td><td>kuma kia | 1121
wavova kwa | 923
sébastien sasa | 700
père sébastien | 700
kia nzambi | 676</td></tr>
<tr><td>3</td><td>bantu ya nkaka | 777
père sébastien sasa | 696
kimfumu ya nzambi | 666
ndinga ya nzambi | 608
bambangi ya yehowa | 584</td></tr>
<tr><td>4</td><td>mfumu ya kuluta nene | 178
manisa manisa manisa manisa | 174
yehowa mfumu ya kuluta | 169
zinga mutindu bakristu fwete | 152
mutindu bakristu fwete zinga | 151</td></tr>
<tr><td>5</td><td>bimvwama ya ndinga ya nzambi | 195
konso muntu ke na luve | 175
manisa manisa manisa manisa manisa | 172
yehowa mfumu ya kuluta nene | 168
luzingu ya mvula na mvula | 157</td></tr>
</table>

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