HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
wol_Latn.jsonl.tsv	12/10/2024	Wolof (wo)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
5,679	161,474	57,857 (35.83 %)	6.7M	27.2 MB	27,386,737

Top 10 domains

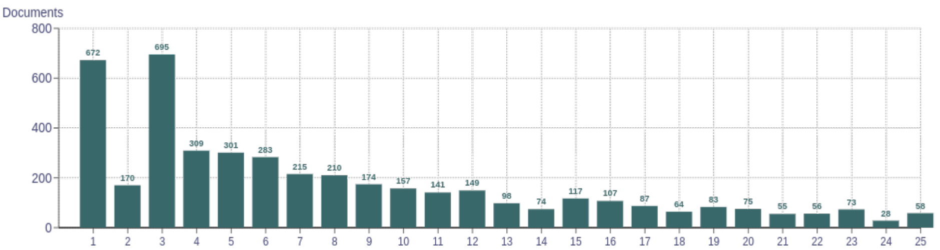
Domain	Docs	% of total
wikipedia.org	2.5K	43.69
bible.is	1K	17.64
jw.org	298	5.25
iqna.ir	181	3.19
wolof-online.com	170	2.99
jangwolof.com	134	2.36
sacred-texts.com	111	1.95
sciencegraph.net	102	1.80
senegalhit.com	62	1.09
osad-sn.com	52	0.92

Top 10 TLDs

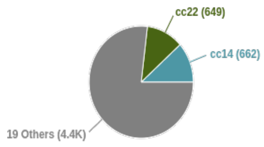
Domain	Docs	% of total
org	3.1K	54.08
com	1.1K	18.86
is	1K	17.64
ir	189	3.33
net	177	3.12
in	43	0.76
info	26	0.46
fr	15	0.26
sn	15	0.26
va	12	0.21

Documents size (in segments)

<= 25 segments **78.38%** (4.5K documents)
> 25 segments **21.62%** (1.2K documents)

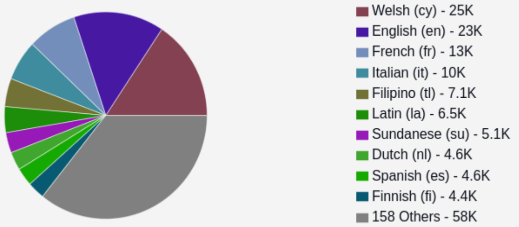


Documents by collection



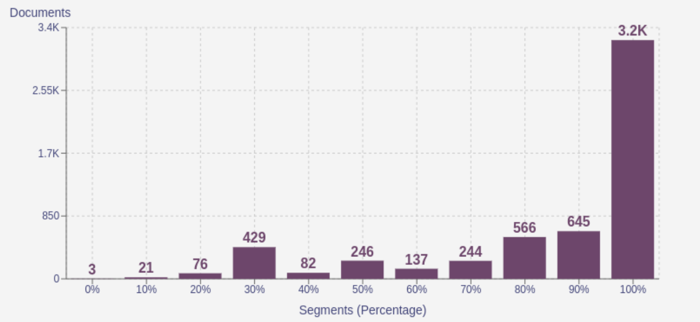
Language Distribution

Number of segments



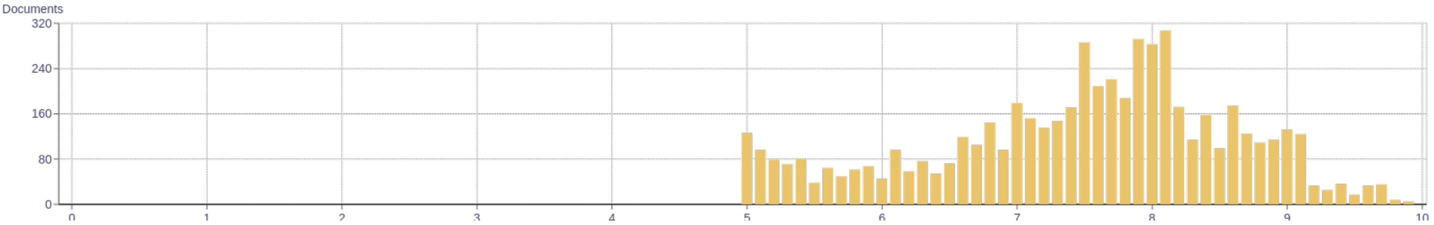
*Wolof (wo) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Wolof (wo) inside documents



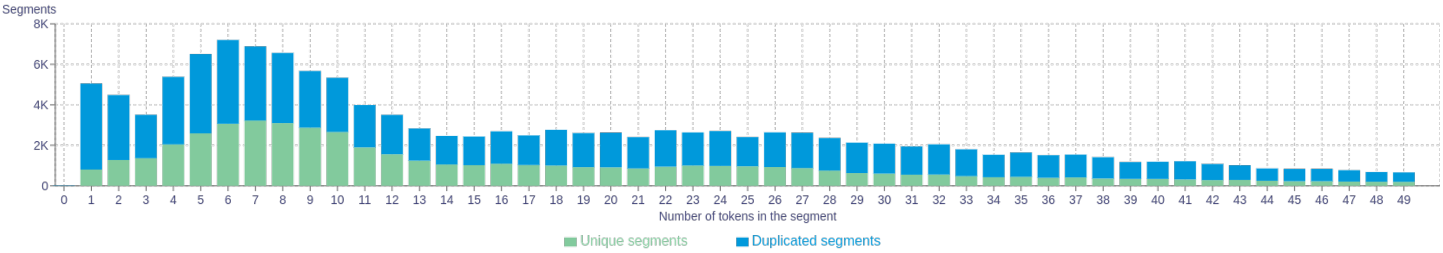
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (5.7K documents)

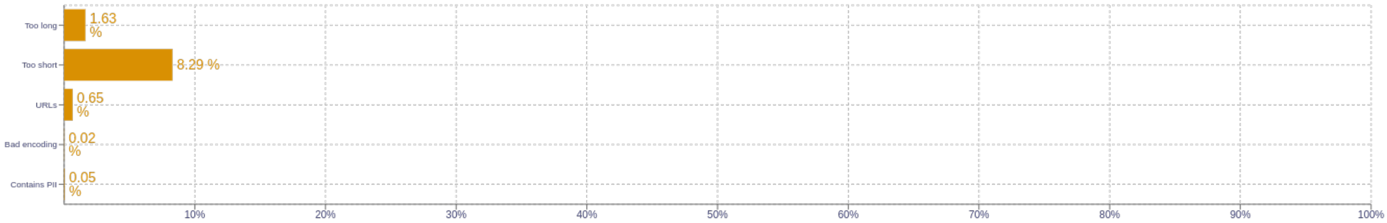


Segment length distribution by token

<= 49 tokens = **50K** segments | **84K** duplicates
> 50 tokens = **28K** segments | **20K** duplicates



Segment noise distribution



Frequent n-grams

<table>
<tr><th>Size</th><th>n-grams</th></tr>
<tr><td>1</td><td>lo | 13720
soppi | 13333
yeesu | 12878
nii | 12146
xaa | 11426</td></tr>
<tr><td>2</td><td>soppi gongikuwaay | 3713
lëë xaa | 2429
xaa nii | 1712
doom doom | 1674
votez mankoo | 1664</td></tr>
<tr><td>3</td><td>votez mankoo taxawu | 1664
taxawu senegaal votez | 1651
doom doom doom | 1355
abc def abc | 726
benno bok yaakaar | 721</td></tr>
<tr><td>4</td><td>taxawu senegaal votez mankoo | 1651
mankoo taxawu senegaal votez | 1651
doom doom doom doom | 1158
dund gu dul jeex | 341
sunu boroom subhaanahu wa | 318</td></tr>
<tr><td>5</td><td>votez mankoo taxawu senegaal votez | 1651
taxawu senegaal votez mankoo taxawu | 1651
mankoo taxawu senegaal votez mankoo | 1651
doom doom doom doom doom | 1131
abc def abc def abc | 716</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