HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
mlt_Latn.jsonl.tsv	9/27/2024	Maltese (mt)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
367,265	8,675,475	4,669,868 (53.83 %)	239M	1.39 GB	1,433,340,040

Top 10 domains

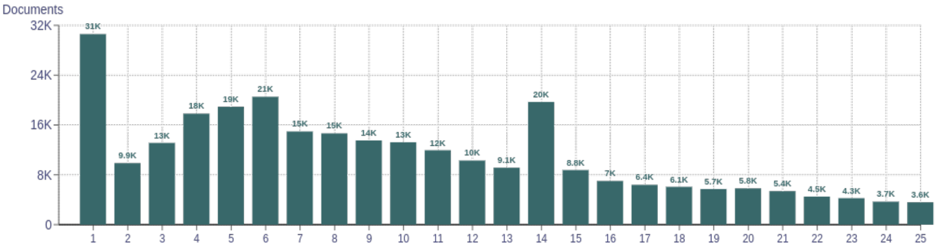
Domain	Docs	% of total
europa.eu	33K	9.03
wikipedia.org	16K	4.37
jiffyrando.com	14K	3.83
itsmygame.org	13K	3.50
airbnb.com	11K	3.10
soft-free-download.com	7.8K	2.13
laikosblog.org	7.2K	1.97
playgame24.com	6.8K	1.84
maltarightnow.com	6.2K	1.69
inewsmalta.com	5.9K	1.60

Top 10 TLDs

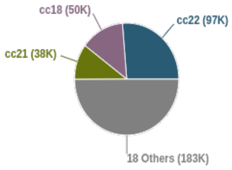
Domain	Docs	% of total
com	188K	51.21
org	60K	16.37
eu	41K	11.07
com.mt	24K	6.53
mt	16K	4.29
net	7.7K	2.10
org.mt	7.5K	2.05
de	2.3K	0.62
ru	1.6K	0.43
nu	1.5K	0.41

Documents size (in segments)

<= 25 segments **76.09%** (279K documents)
> 25 segments **23.91%** (88K documents)

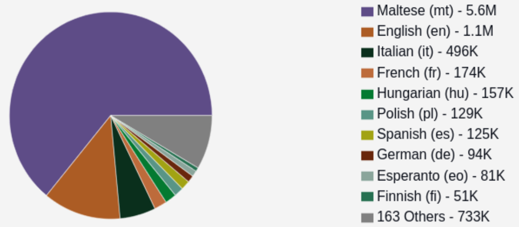


Documents by collection

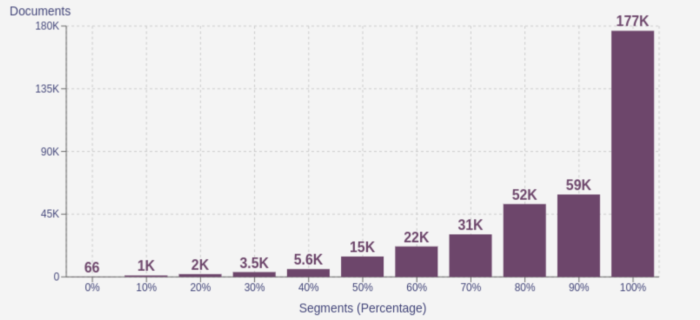


Language Distribution

Number of segments

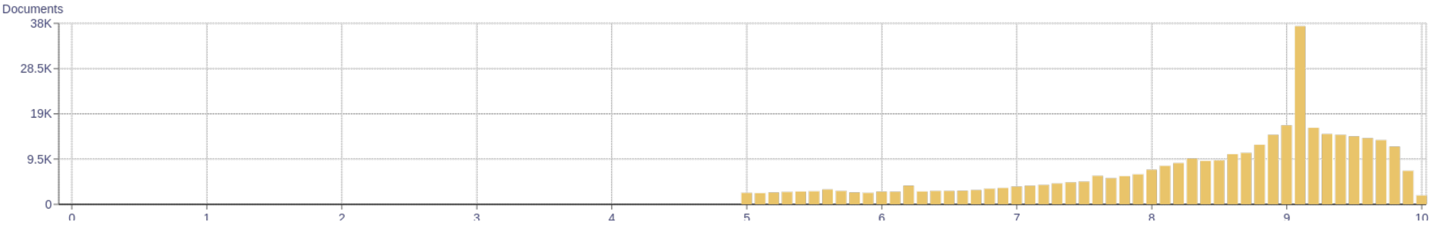


Percentage of segments in Maltese (mt) inside documents



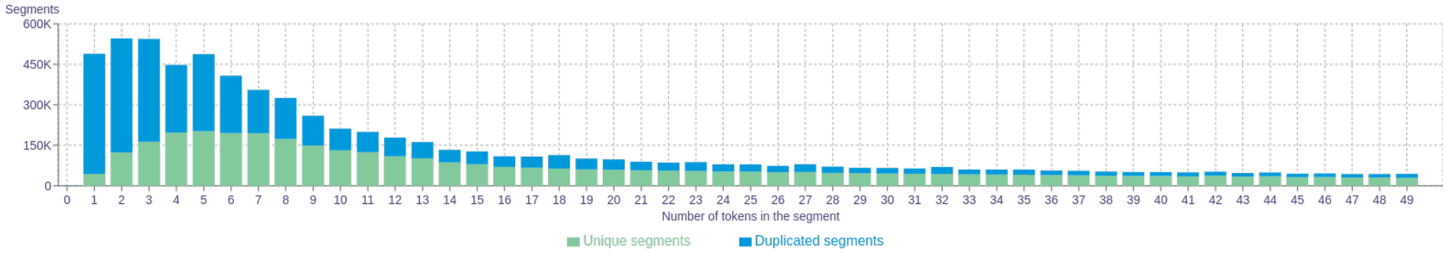
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (367K documents)

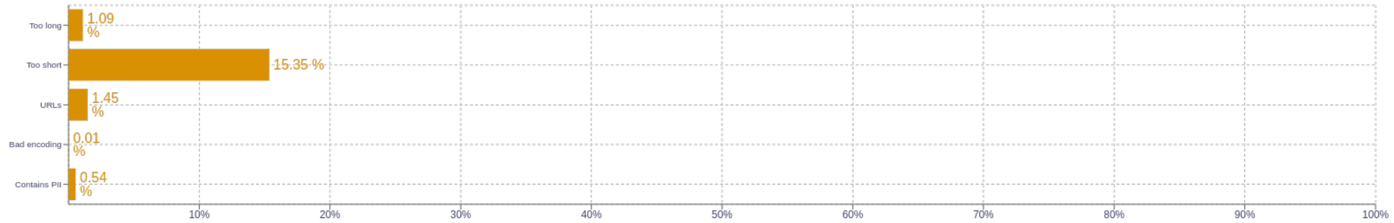


Segment length distribution by token

<= 49 tokens = **3.6M** segments | **3.6M** duplicates
> 50 tokens = **1.5M** segments | **438K** duplicates



Segment noise distribution



Frequent n-grams

<table>
<tr><th>Size</th><th>n-grams</th></tr>
<tr><td>1</td><td>f | 1341364
b | 1080275
tista | 411781
the | 367038
a | 310130</td></tr>
<tr><td>2</td><td>b 'mod | 239815
f 'dan | 136225
f 'din | 54496
of the | 53216
b 'xejn | 52193</td></tr>
<tr><td>3</td><td>b 'mod partikolari | 46873
f 'dan ir-rigward | 18984
f 'dan il-każ | 18966
ewropew u tal-kunsill | 17700
il-logħba hija kellha | 11803</td></tr>
<tr><td>4</td><td>tal-parlament ewropew u tal-kunsill | 17585
tal-kodiċi u paste fil-kodiċi | 11709
rabta biex tniżżel il-logħba | 11709
kopja u jibgħat l-link | 11709
kopja tal-kodiċi u paste | 11709</td></tr>
<tr><td>5</td><td>tal-kodiċi u paste fil-kodiċi html | 11709
l-link lil habib jew ħbieb | 11709
kopja tal-kodiċi u paste fil-kodiċi | 11709
impenjati li jipprovdu l-aqwa żjarat | 10213
affarijiet li għandek tkun taf | 9441</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