HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
mt_1.jsonl.tsv	3/16/2024	Maltese (mt)

Volumes

Docs	Segments	Unique segments	Tokens	Size
111,123	11,174,217	13,949 (0.12 %)	134M	743.31 MB

Type-Token Ratio

Maltese (mt)
0.02

Top 10 domains

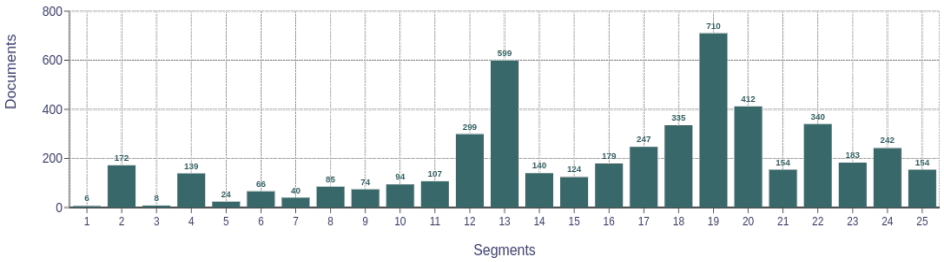
Domain	Docs	% of total
diebuchsuche.com	67K	59.99
europa.eu	7.3K	6.53
airbnb.com	2.7K	2.40
wondershare.com	1.7K	1.52
knisja.mt	1.4K	1.28
netnews.com.mt	1.4K	1.22
sgames.org	1.2K	1.05
uhm.org.mt	1K	0.92
wikipedia.org	949	0.85
gov.mt	935	0.84

Top 10 TLDs

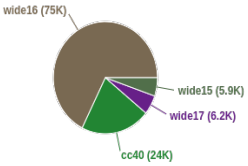
Domain	Docs	% of total
com	81K	73.10
eu	8.3K	7.43
org	5.6K	5.07
com.mt	5.5K	4.95
mt	3.5K	3.19
org.mt	2.3K	2.10
net	1.3K	1.17
gr	505	0.45
fr	354	0.32
pt	269	0.24

Documents size (in segments)

<= 25 segments **4.44%** (4.9K documents)
> 25 segments **95.56%** (106K documents)

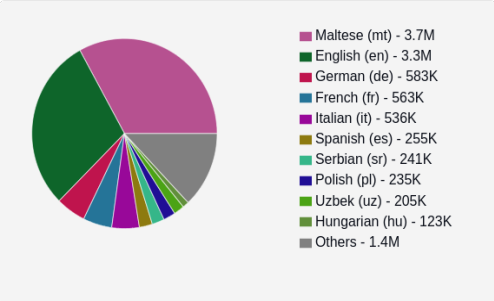


Documents by collection

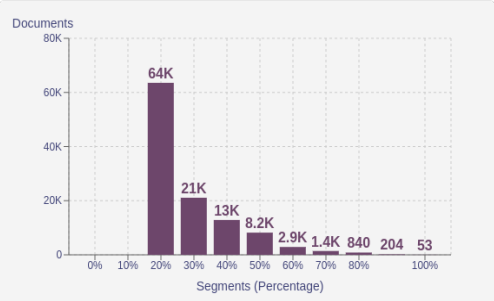


Language Distribution

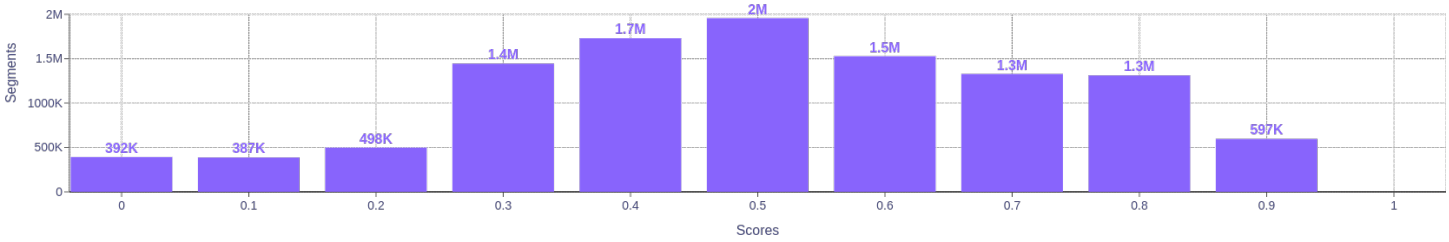
Number of segments



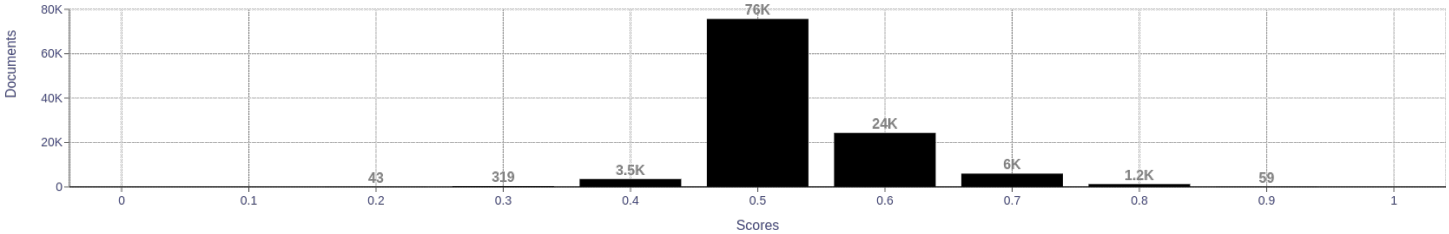
Percentage of segments in Maltese (mt) inside documents



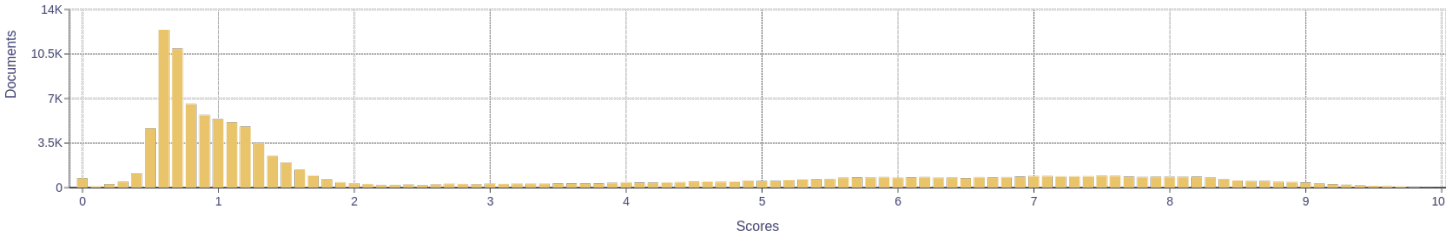
Distribution of segments by fluency score



Distribution of documents by average fluency score



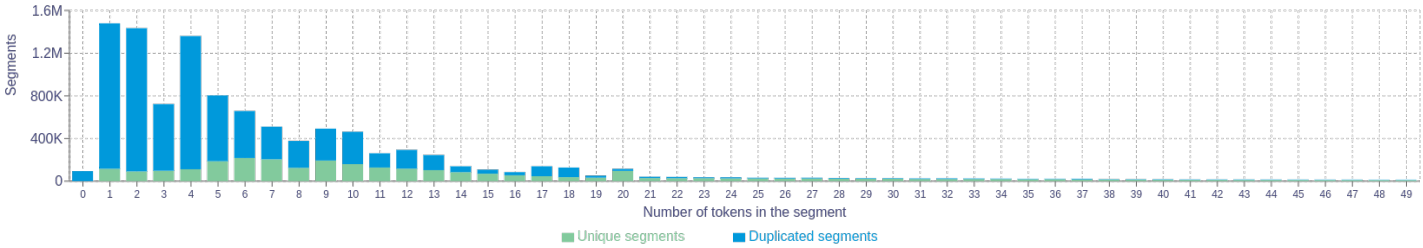
Distribution of documents by document score



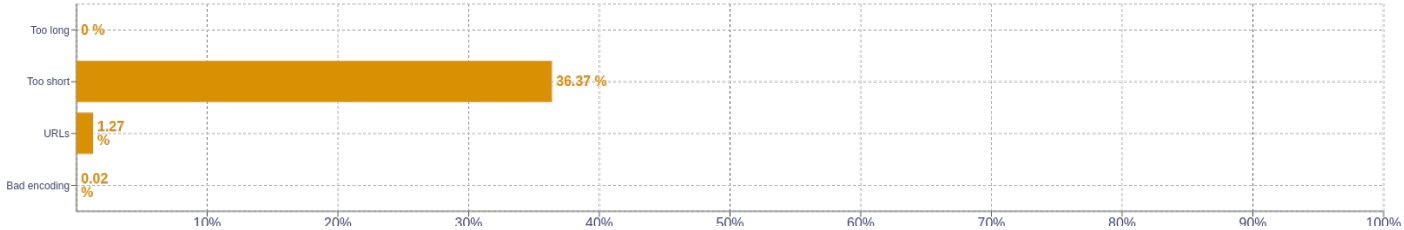
Segment length distribution by token

<= 49 tokens = **2.8M** segments | **7.8M** duplicates

> 50 tokens = **496K** segments | **81K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	ta | 2291265 · u | 1921093 · li | 1644494 · the | 1156434 · of | 727414
2	data minn | 201969 · id-dħul ta | 201052 · dettalji aktar | 201031 · notazzjonijiet alternattivi | 200847 · watch ktieb | 200362
3	made by freepik | 66687 · icons made by | 66687 · is licensed by | 66669 · www.flaticon.com is licensed | 66667 · licensed by cc | 66667
4	icons made by freepik | 66687 · www.flaticon.com is licensed by | 66667 · made by freepik from | 66667 · is licensed by cc | 66667 · from www.flaticon.com is licensed | 66667
5	www.flaticon.com is licensed by cc | 66667 · made by freepik from www.flaticon.com | 66667 · icons made by freepik from | 66667 · from www.flaticon.com is licensed by | 66667 · freepik from www.flaticon.com is licensed | 66667

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