HPLT Analytics report

General overview

Corpus	Analytics date	Language
HPLT-docslite.ar.tsv	6/28/2024	Arabic (ar)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
26,800,103	3,655,762,877			297.02 GB	

Top 10 domains

Domain	Docs	% of total
diebuchsuche.com	1.2M	4.45
alibaba.com	536K	2.00
movizhouse.com	252K	0.94
blogspot.com.eg	188K	0.70
blogspot.com	165K	0.62
aliexpress.com	164K	0.61
made-in-china.com	128K	0.48
blogspot.ae	124K	0.46
blogspot.co.il	123K	0.46
mawdoo3.com	115K	0.43

Top 10 TLDs

Domain	Docs	% of total
com	16M	60.32
net	3.3M	12.32
org	1.2M	4.55
pl	507K	1.89
nl	341K	1.27
info	338K	1.26
fr	305K	1.14
be	296K	1.10
ae	270K	1.01
de	248K	0.93

Documents size (in segments)

<= 25 segments **8.29%** (2.2M documents)
> 25 segments **91.71%** (25M documents)
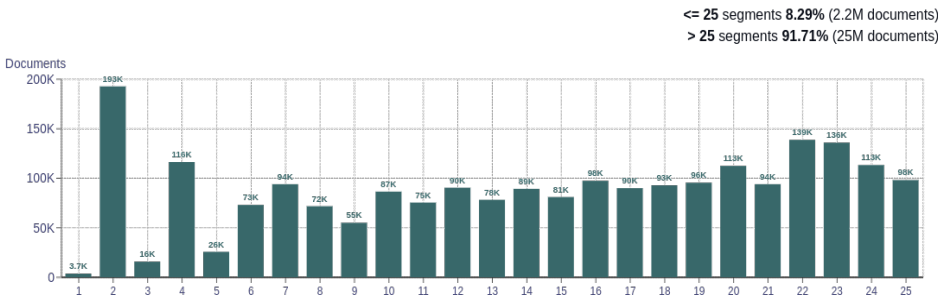
Documents by collection

wide16 (12M)
1 Others (2.3M)
cc40 (8.2M)
wide15 (3.8M)
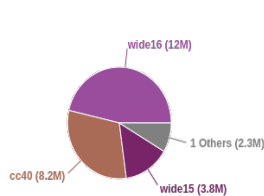
Language Distribution

Number of segments

- Arabic (ar) - 2.8B
- English (en) - 418M
- Egyptian Arabic (arz) - 73M
- Persian (fa) - 68M
- French (fr) - 58M
- German (de) - 30M
- Spanish (es) - 26M
- Italian (it) - 15M
- Dutch (nl) - 10M
- Russian (ru) - 9.6M
- 164 Others - 129M
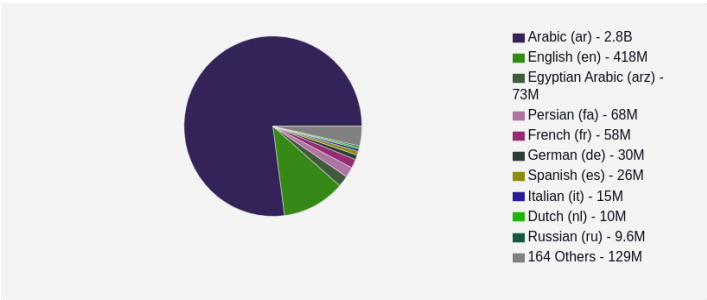
Percentage of segments in Arabic (ar) inside documents

Distribution of documents by document score

score <= 5 - **23.93%** (6.4M documents)
score > 5 - **76.07%** (20M documents)
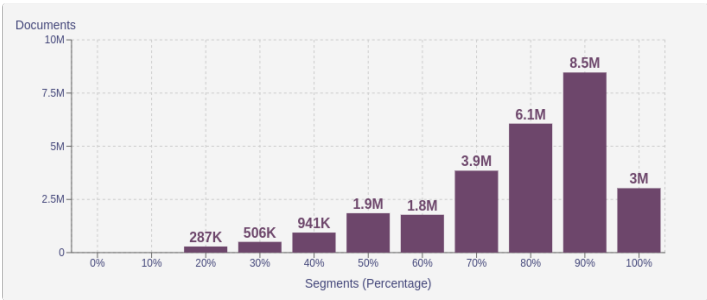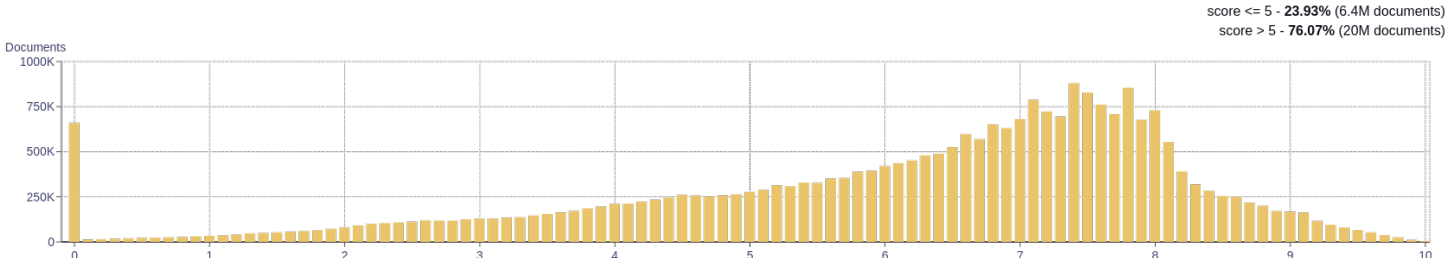
About HPLT Analytics

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