HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
HPLT-v2-ara_Arab.tsv	9/30/2024	Arabic (ar)

Volumes

Docs	Segments	Characters	Size
82,670,005	2,200,259,820	277,388,083,983	460.13 GB

Top 10 domains

Domain	Docs	% of total
blogspot.com	1.1M	1.39%
wikipedia.org	679K	0.82%
alittihad.ae	461K	0.56%
yoo7.com	395K	0.48%
ahlamontada.com	319K	0.39%
aljazeera.net	310K	0.38%
rt.com	300K	0.36%
mazagfm.com	289K	0.35%
paltoday.ps	266K	0.32%
web.app	246K	0.30%

Top 10 TLDs

Domain	Docs	% of total
com	48M	58.45%
net	11M	13.74%
org	4.5M	5.43%
pl	1.6M	1.95%
ae	1.2M	1.45%
nl	951K	1.15%
info	890K	1.08%
ma	668K	0.81%
ps	668K	0.81%
be	657K	0.79%

Register labels

- HI - 3.0%
- ID - 5.2%
- IN - 10.0%
- IP - 10.7%
- LY - 0.5%
- MIX - 1.5%
- NA - 42.2%
- OP - 9.5%
- SP - 0.5%
- UNK - 17.1%

MT:10.6% | 8.8M Documents

- HI_other - 1.8%
- HI_re - 1.2%
- ID_other - 5.2%
- IN_dtp - 3.3%
- IN_en - 1.2%
- IN_fi - 0.0%
- IN_lt - 0.3%
- IN_other - 5.1%
- IN_ra - 0.0%
- IP_ds - 9.1%
- IP_ed - 0.0%
- IP_other - 1.6%
- LY_other - 0.5%
- MIX - 1.5%
- NA_nb - 0.9%
- NA_ne - 34.1%
- NA_other - 2.9%
- NA_sr - 4.3%
- OP_av - 0.9%
- OP_ob - 2.2%
- OP_other - 2.6%
- OP_rs - 3.1%
- OP_rv - 0.7%
- SP_it - 0.3%
- SP_other - 0.1%
- UNK - 17.1%
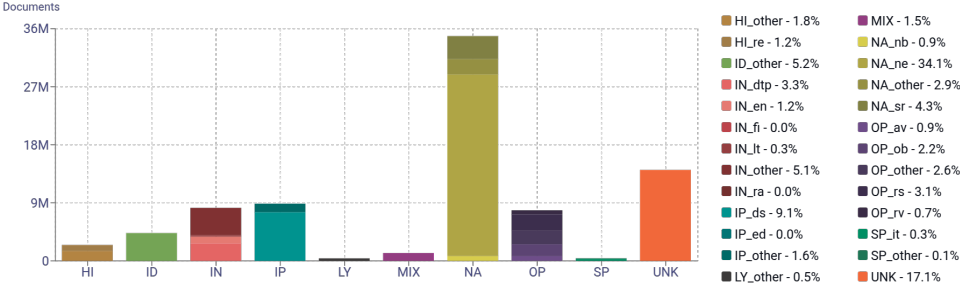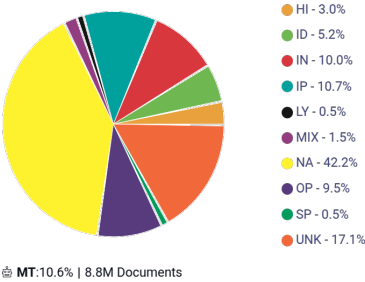
Documents size (in segments)

<= 25 segments **73.01%** (60M documents)
> 25 segments **26.99%** (22M documents)
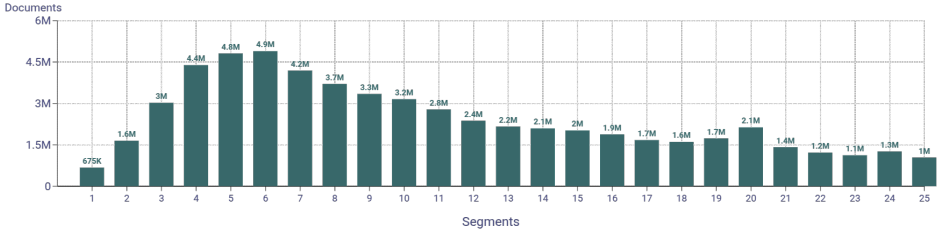
Documents by collection

CC = 60.27%
IA = 39.73%

- cc18 (12M)
- cc22 (21M)
- 19 Others (49M)
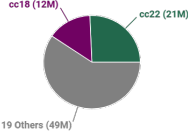
Language Distribution

Number of segments in the Arabic (ar) corpus

- Arabic (ar) - 2B
- English (en) - 60M
- Italian (it) - 49M
- Egyptian Arabic (arz) - 30M
- Persian (fa) - 12M
- French (fr) - 11M
- German (de) - 3.7M
- Dutch (nl) - 3.5M
- Spanish (es) - 3.1M
- Cebuano (ceb) - 2.7M
- 165 Others - 29M
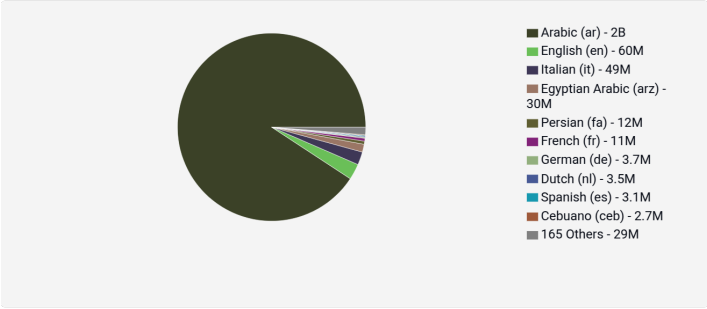
Percentage of segments in Arabic (ar) inside documents
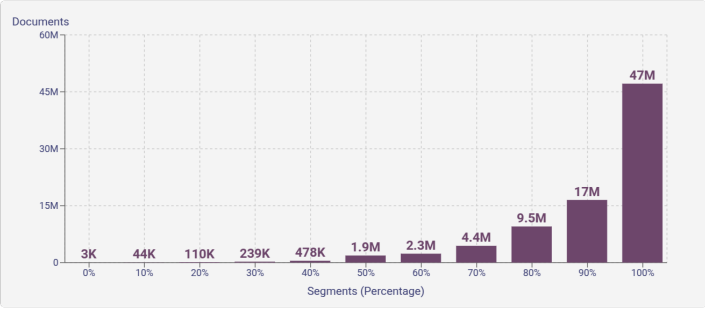
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (83M documents)

Segment noise distribution

- Too long - 0.52 %
- Too short - 11.80 %
- URLs - 0.86 %
- Bad encoding - 0.00 %
- Contains PII - 0.08 %
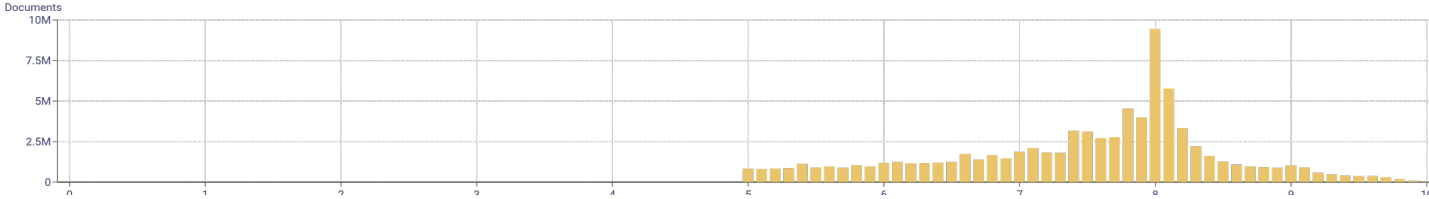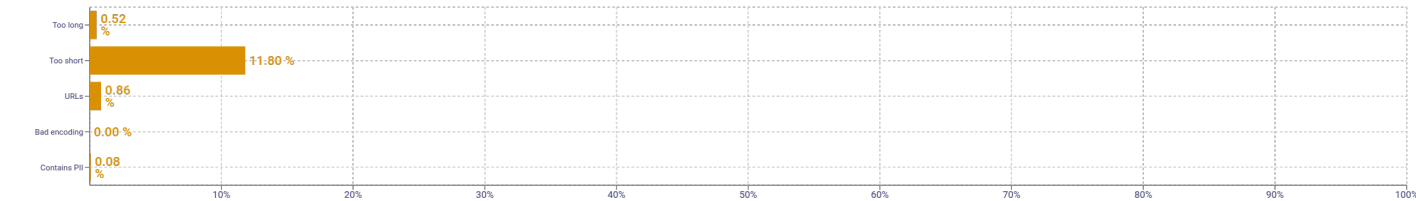
About HPLT Analytics

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

Register labels

Name	Abbr.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	HI
Recipe	re
Informational persuasion	IP
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	IN
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	OP
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av