HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
ara_Arab.jsonl.tsv	8/4/2025	Arabic (ar)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
82,669,942	2,196,313,084	1,046,496,186 (47.65 %)	57B	277,388,083,983	460.13 GB

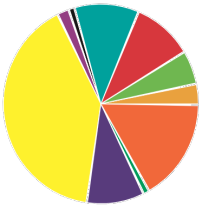
Top 10 domains

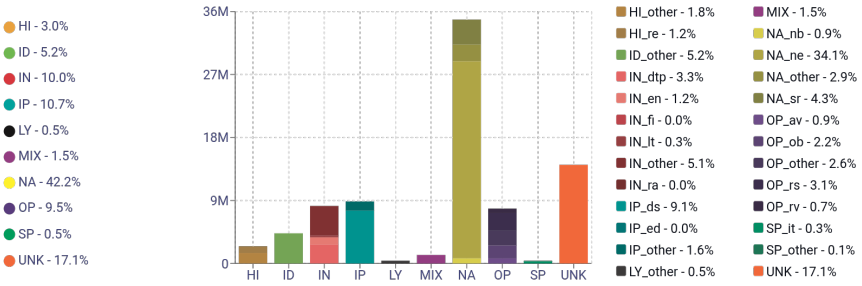
Domain	Docs	% of total
blogspot.com	1.1M	1.39%
wikipedia.org	679K	0.82%
alittihad.ae	461K	0.56%
yoo7.com	395K	0.48%
ahlamontada.com	319K	0.39%
aljazeera.net	310K	0.38%
rt.com	300K	0.36%
mazagfm.com	289K	0.35%
paltoday.ps	266K	0.32%
web.app	246K	0.30%

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

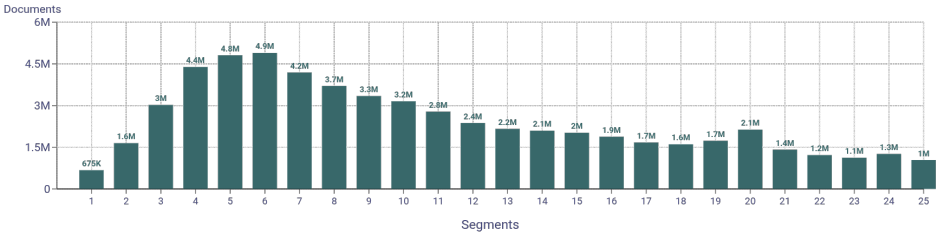




MT:10.6% | 8.8M Documents

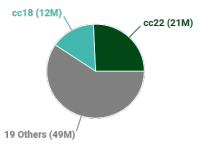
Documents size (in segments)

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



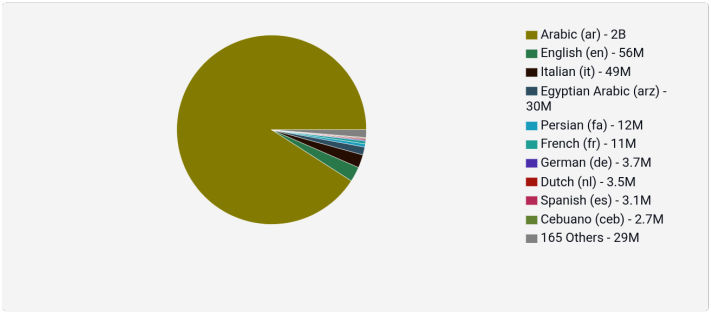
Documents by collection

CC = 60.27%
IA = 39.73%

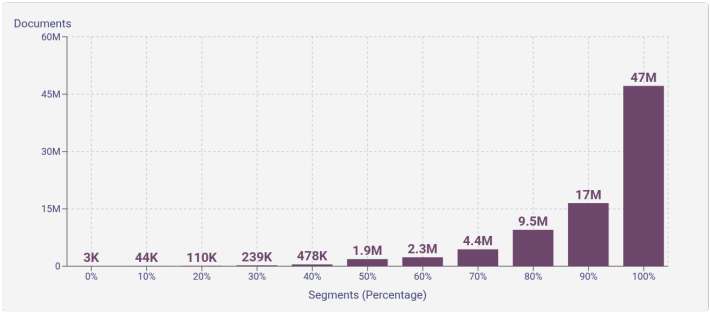


Language Distribution

Number of segments in the Arabic (ar) corpus

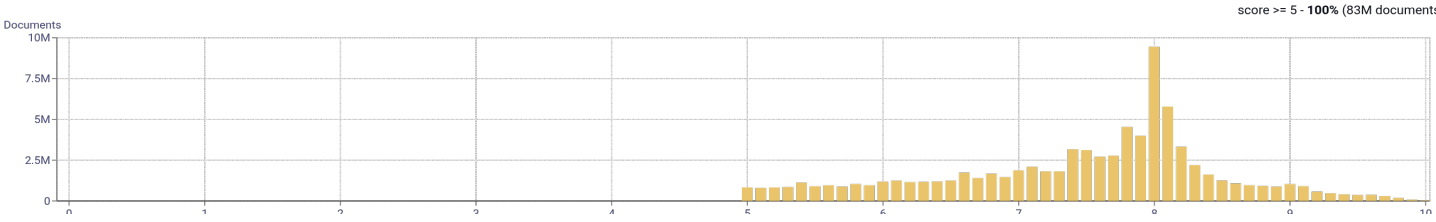


Percentage of segments in Arabic (ar) inside documents



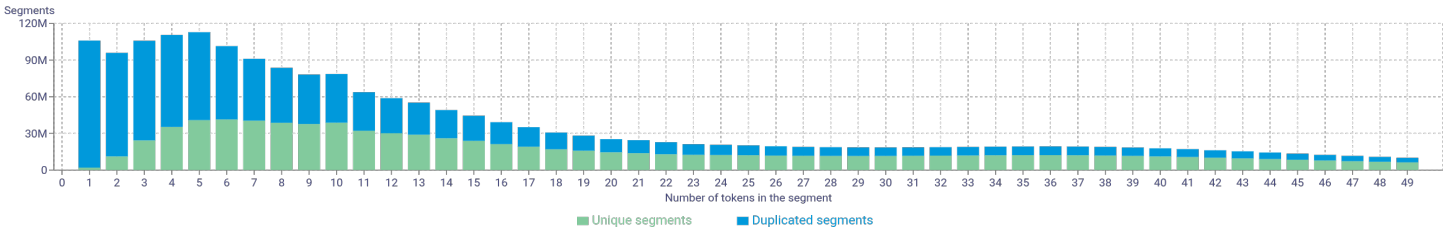
Distribution of documents by document score

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

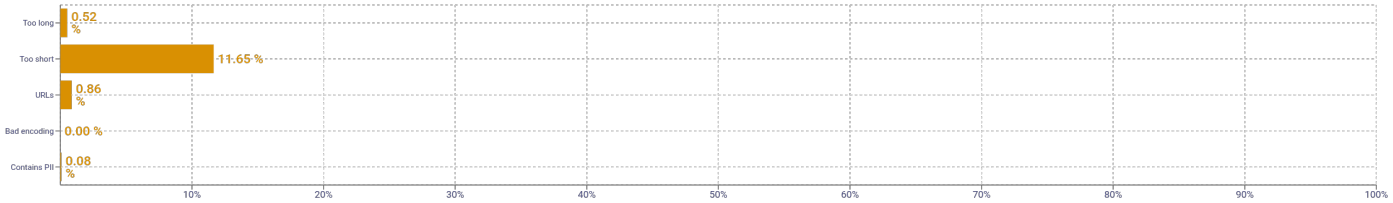


Segment length distribution by token

≤ **49** tokens = **1.9B** segments | **1.1B** duplicates
> **50** tokens = **279M** segments | **99M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	153895509 | الله · 141854449 | كسارة · 87385913 | في · 85850941 | ان · 65893497 | الحجر
2	10586582 | الحجر الجيري · 10172095 | آلة طحن · 9834098 | أكمل القراءة · get price | 8956907 · 8648184 | الكرة مطحنة
3	8414274 | احصل على السعر · 3076150 | آلة كسارة الحجر · اتصل بالموردال whatsapp | 1120935 · 987728 | آلة صنع الرمل · 887962 | إله إلا الله
4	press copyright contact us | 403330 · copyright contact us creators | 392448 · shanghai gm machinery co | 374367 · ltd 363887 | هي إحدى شركات · find complete details about | 361460
5	press copyright contact us creators | 392435 · ltd 357052 | هي إحدى شركات التكنولوجيا · 333261 | أفضل مجموعة من شركات التصنيع · 301392 | أعلى بلدان العرض أو المناطق · 271181 | أبي هريرة رضي الله عنه

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