

General overview

Corpus	Date	Language
hplt-v3-ars_Arab	10/3/2025	Najdi Arabic

Volumes

Docs	Segments	Unique segments	Duplication ratio	Tokens	Characters	Size
1,810	38,346	29,627 (77.26 %)	22.74%	525K	2,726,160	4.69 MB

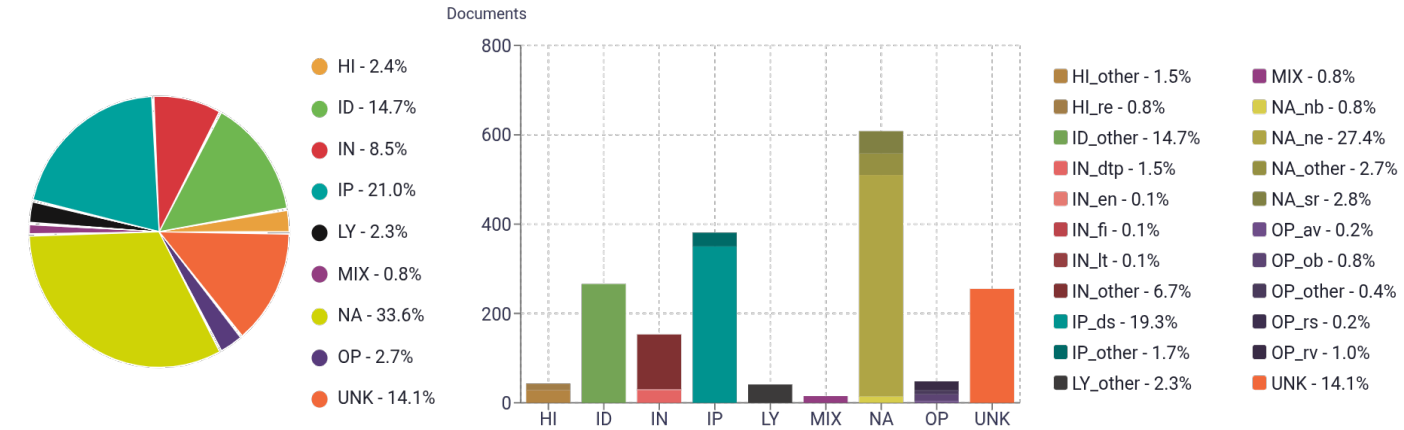
Top 10 domains

Domain	Docs	% of total
unlimit-tech.com	51	2.82%
otlobmehany.com	47	2.60%
alriyadh.com	31	1.71%
asir.me	29	1.60%
jobs-arab.com	28	1.55%
blogspot.com	27	1.49%
alweeam.com.sa	23	1.27%
a1ash.com	23	1.27%
mokhtsar.net	22	1.22%
dubaisession.com	22	1.22%

Top 10 TLDs

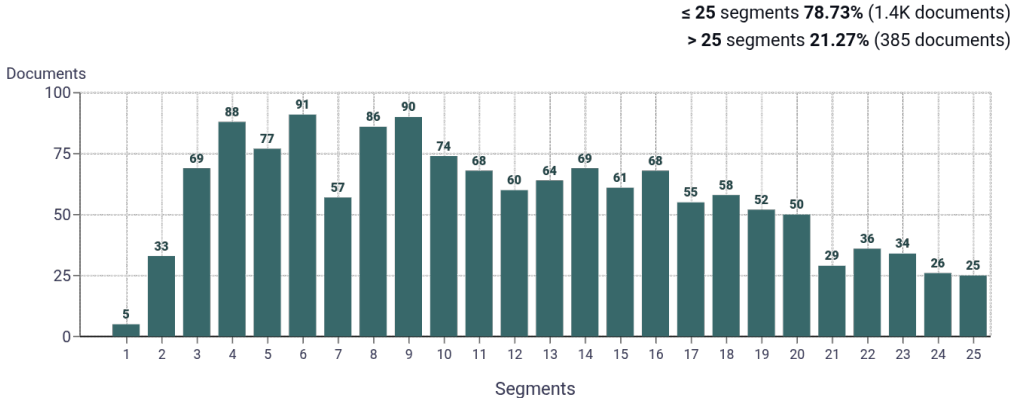
Domain	Docs	% of total
com	1.2K	68.84%
net	285	15.75%
org	63	3.48%
com.sa	41	2.27%
me	33	1.82%
info	19	1.05%
net.sa	12	0.66%
news	9	0.50%
co	9	0.50%
us	7	0.39%

Register labels

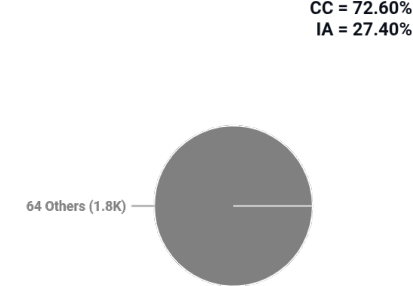


MT:3.9% | 70 Documents

Documents size (in segments) ⓘ

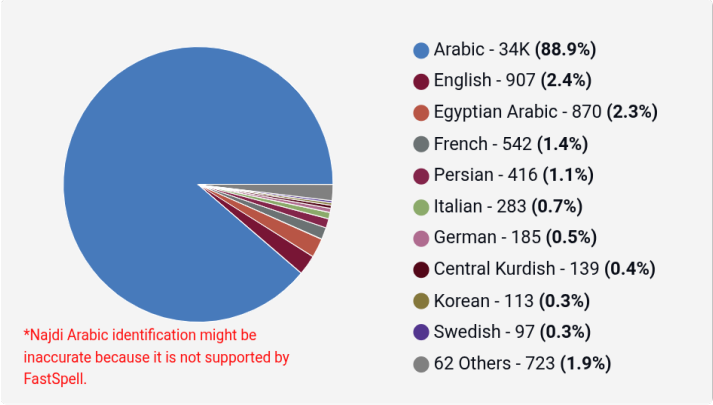


Document collections

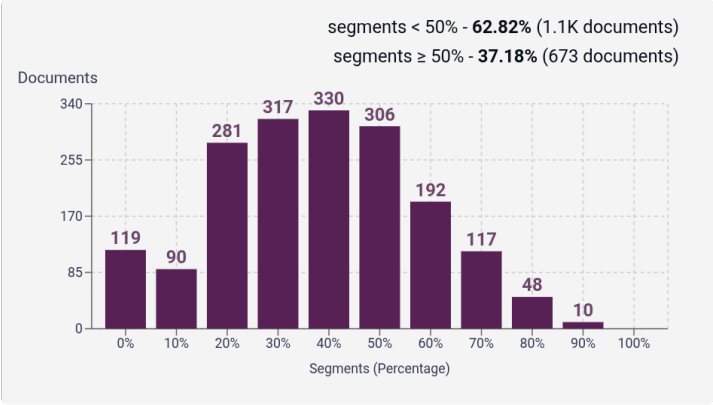


Language Distribution

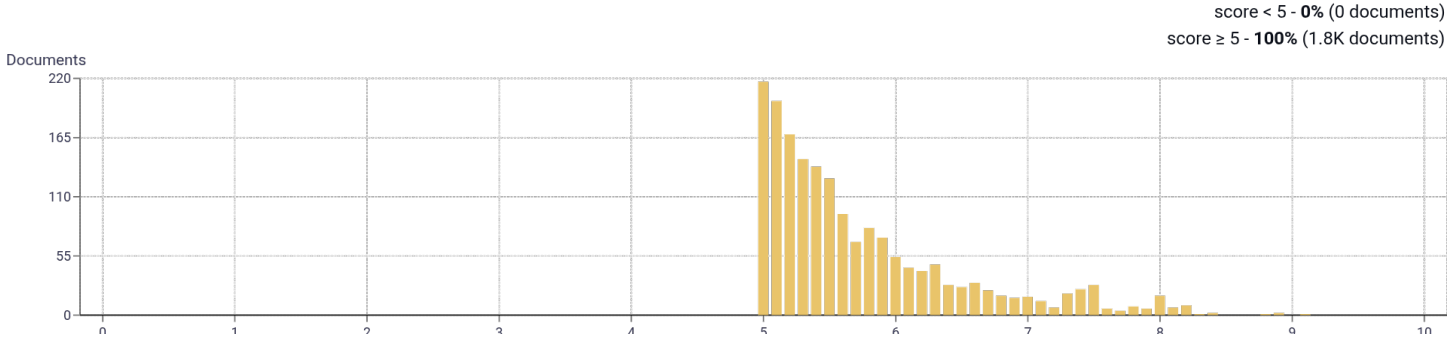
Number of segments in the Najdi Arabic corpus



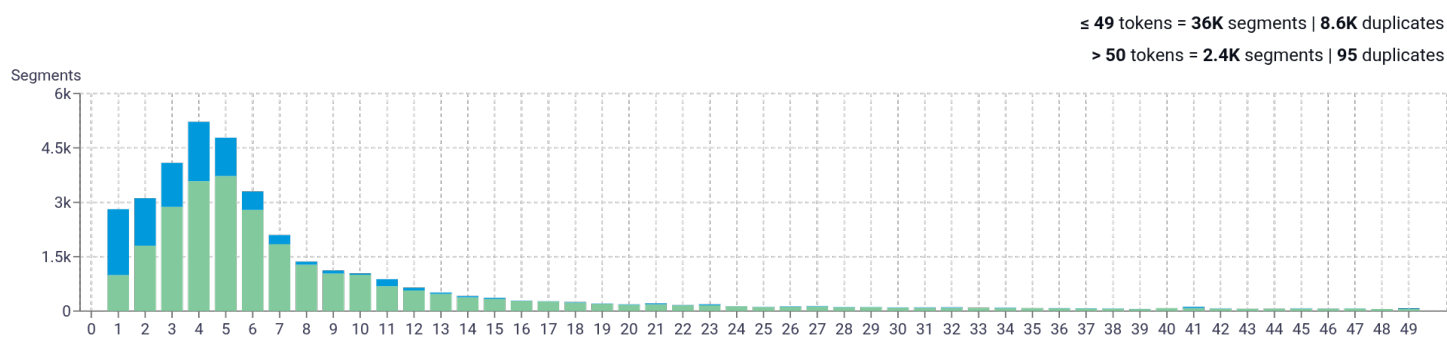
Percentage of segments in Najdi Arabic inside documents



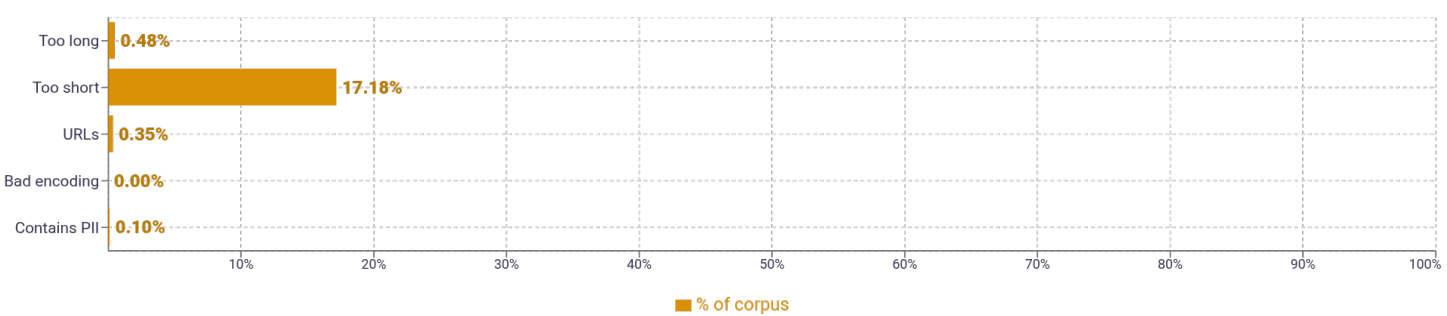
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	3,622 شركة 3,044 بالرياض 2,990 الله 2,153 تنظيف 1,781 الرياض	
2	1,649 شركة تنظيف 477 نقل عفش 425 بنك الراجحي 374 شركة نقل 366 يمن موبايل	
3	247 المملكة العربية السعودية 244 فتح في نافذة 236 نص نص نص 234 شركة نقل عفش 222 نقل عفش بالرياض	
4	244 فتح في نافذة جديدة 225 نص نص نص 171 شكرًا شكرًا شكرًا شكرًا شكرًا 110 شركة تنظيف فلل بالرياض 96 شركة تنظيف خزانات بالرياض	
5	215 نص نص نص نص نص 170 شكرًا شكرًا شكرًا شكرًا شكرًا شكرًا شكرًا 64 فقط الأعضاء المسجلين والمفعلين يمكنهم 64 الأعضاء المسجلين والمفعلين يمكنهم رؤية	
	52 السلام عليكم ورحمة الله وبركاته	

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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 of 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>, , , 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.	Name	Abbr.	Name	Abbr.
Machine-translated	MT	How-to or instructions	HI	Description of a thing or person	dtp
Lyrical	LY	Recipe	re	FAQ	fi
Spoken	SP	Informational persuasion	IP	Legal terms & conditions	lt
Interview	it	Description with intent to sell	ds	Opinion	OP
Interactive discussion	ID	News & opinion blog or editorial	ed	Review	rv
Narrative	NA	Informational description	IN	Opinion blog	ob
News report	ne	Enciclopedia article	en	Denominational religious blog or sermon	rs
Sports report	sr	Research article	ra	Advice	av
Narrative blog	nb				