

General overview

Corpus	Date	Language
snd_Arab.jsonl.tsv	9/6/2024	Sindhi (sd)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
100,298	2,825,658	1,849,842 (65.47 %)	105M	425,901,658	719.43 MB

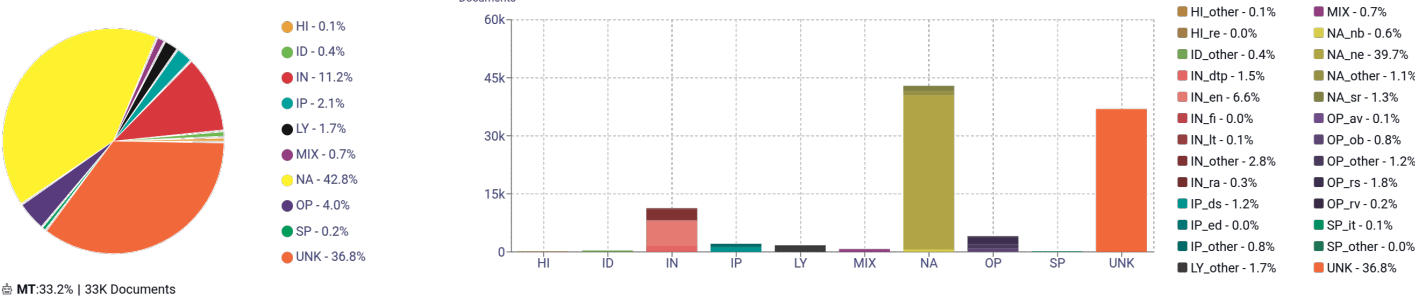
Top 10 domains

Domain	Docs	% of total
awamiawaz.com	9.7K	9.63%
wikipedia.org	7.1K	7.06%
awamiawaz.pk	5.5K	5.44%
voiceofsindh.co...	5K	4.99%
thetimenews.tv	4.2K	4.16%
sarwan.pk	3.9K	3.85%
dailysindhyar.com	2.5K	2.53%
ktnews.tv	2.5K	2.46%
blogfa.com	2.5K	2.45%
sindhsalamat.com	2K	2.00%

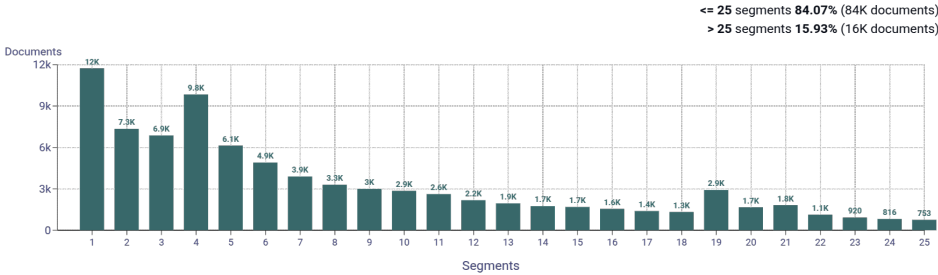
Top 10 TLDs

Domain	Docs	% of total
com	58K	57.99%
org	10K	10.17%
pk	9.5K	9.45%
tv	7.4K	7.37%
com.pk	6.8K	6.73%
net	2K	2.02%
zone	1.2K	1.15%
ir	965	0.96%
co.uk	749	0.75%
ru	281	0.28%

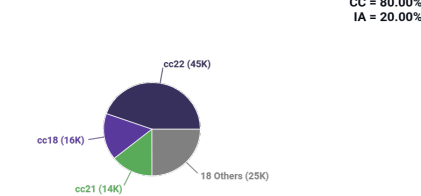
Register labels



Documents size (in segments)

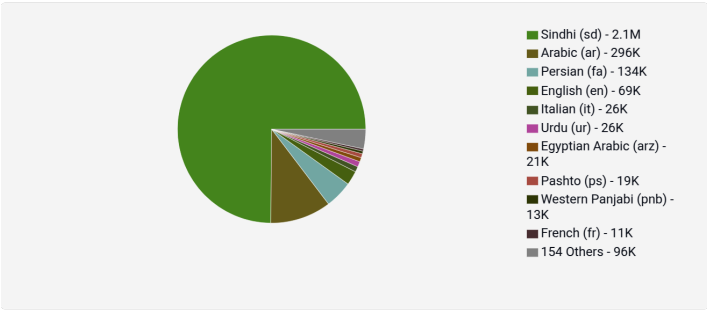


Documents by collection

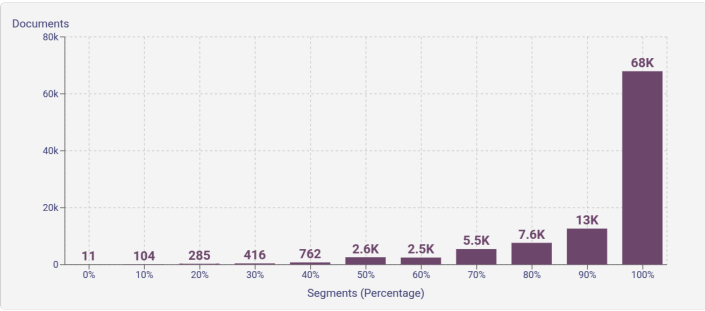


Language Distribution

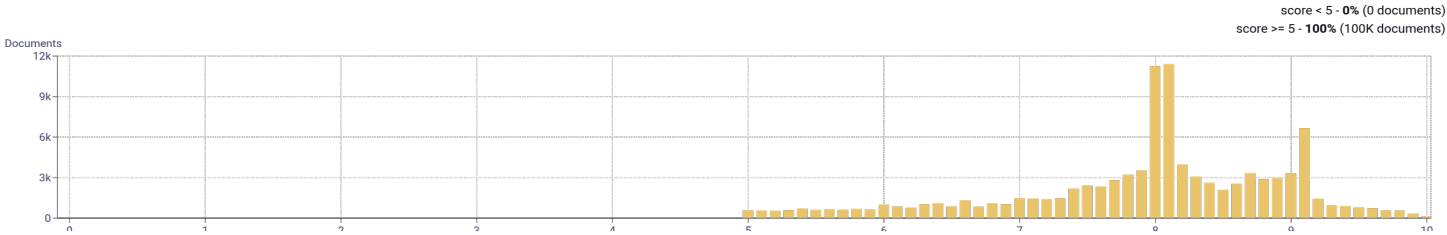
Number of segments in the Sindhi (sd) corpus



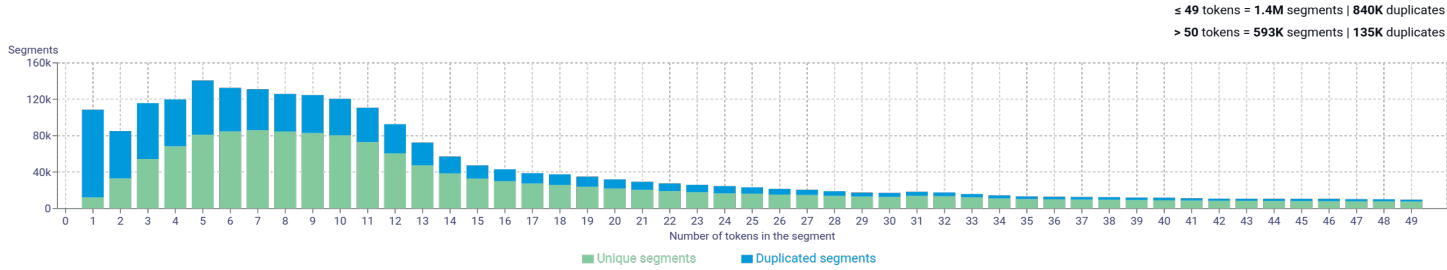
Percentage of segments in Sindhi (sd) inside documents



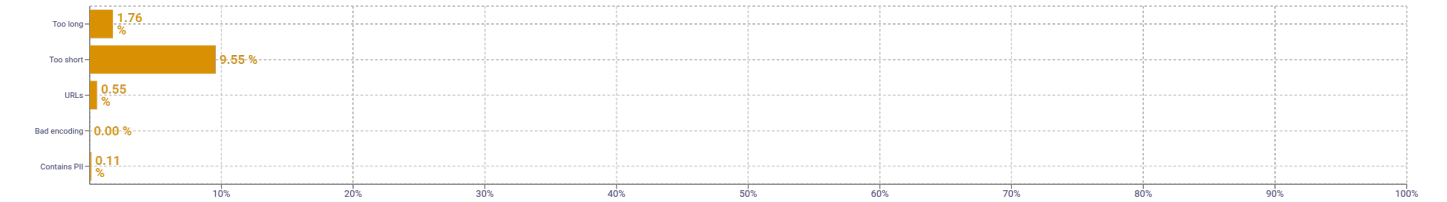
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	سند   221402   ا   253265   ڪيو   333327   جا   378486   نه   594656
2	ويب ڏيک   19851   ڪيو وڃي   21735   ڪيو ويندو   22272   اسلام آباد   25821   ڪٿي وٺي   27683
3	مديا سان ڳالها نيندي   3735   پيداوار جي تفصيل   4351   استعمال ڪيو ويندو   4435   ايس اي او   4532   پي ٽي آءِ   6571
4	سيد مراد علي شاهه   1874   سائين جي ايم سيد   2847   اسمبلي جو اجلاس طلب   2103   قومي اسمبلي جو اجلاس   2121   سند جي وڏي وزير   2334
5	وڏي وزير سيد مراد علي   1489   ڊيليو ڊيليو ڊيليو ڊيليو   1637   جواب شامل ڪريجنواب منسوخ ڪريڻ   1758   وزيراعظم جي اعتماد جو ووٽ   2102   قومي اسمبلي جو اجلاس طلب   2103

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	dt
FAQ	fi
Legal terms & conditions	lt
Opinion	OP
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av