HPLT Analytics report

HPLTAnalytics

General overview

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

Volumes

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

Top 10 domains

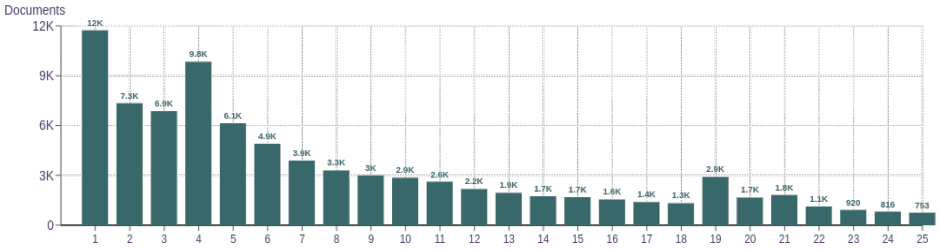
Domain	Docs	% of total
awamiawaz.com	9.7K	9.63
wikipedia.org	7.1K	7.06
awamiawaz.pk	5.5K	5.44
voiceofsindh.com.pk	5K	4.99
thetimenews.tv	4.2K	4.16
sarwan.pk	3.9K	3.85
dailysindhyar.com	2.5K	2.53
ktnnews.tv	2.5K	2.46
blogfa.com	2.5K	2.45
sindhsalamat.com	2K	2.00

Top 10 TLDs

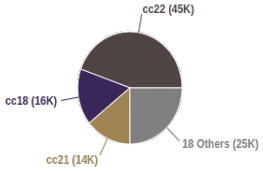
Domain	Docs	% of total
com	58K	57.99
org	10K	10.17
pk	9.5K	9.45
tv	7.4K	7.37
com.pk	6.8K	6.73
net	2K	2.02
zone	1.2K	1.15
ir	965	0.96
co.uk	749	0.75
ru	281	0.28

Documents size (in segments)

<= 25 segments **84.07%** (84K documents)
> 25 segments **15.93%** (16K documents)

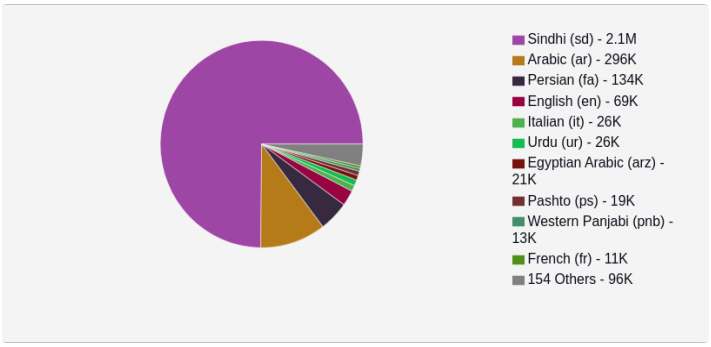


Documents by collection

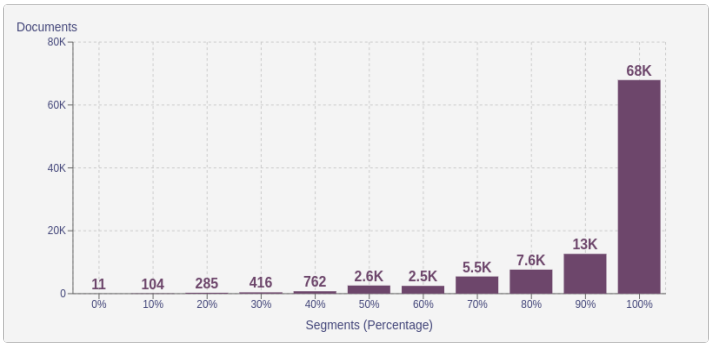


Language Distribution

Number of segments

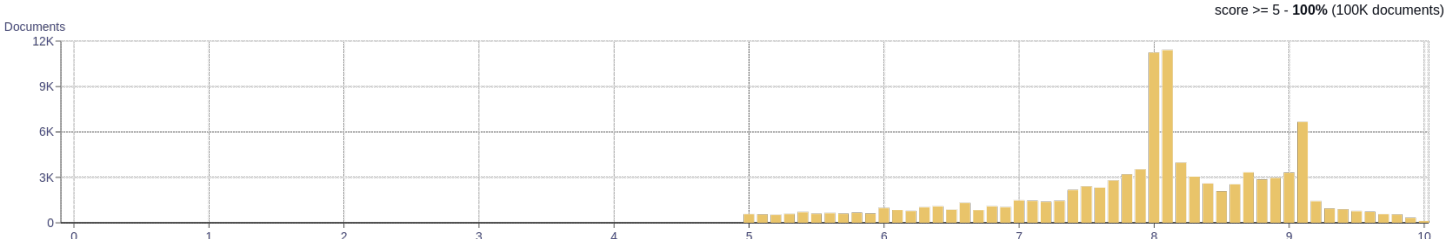


Percentage of segments in Sindhi (sd) inside documents



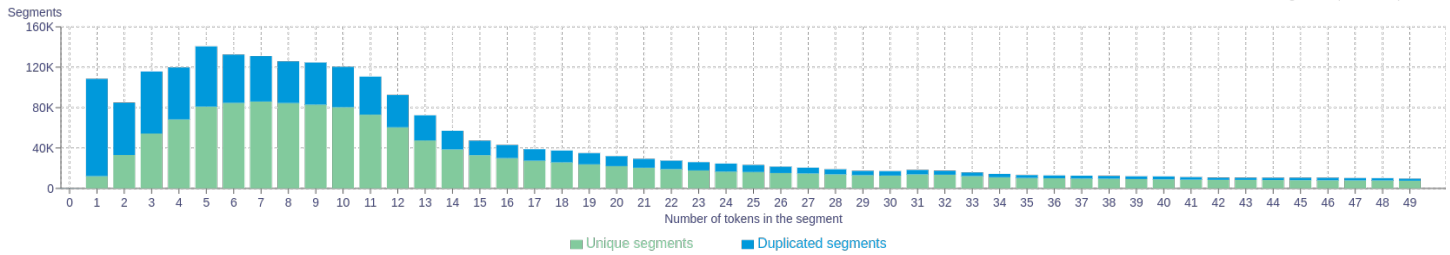
Distribution of documents by document score

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

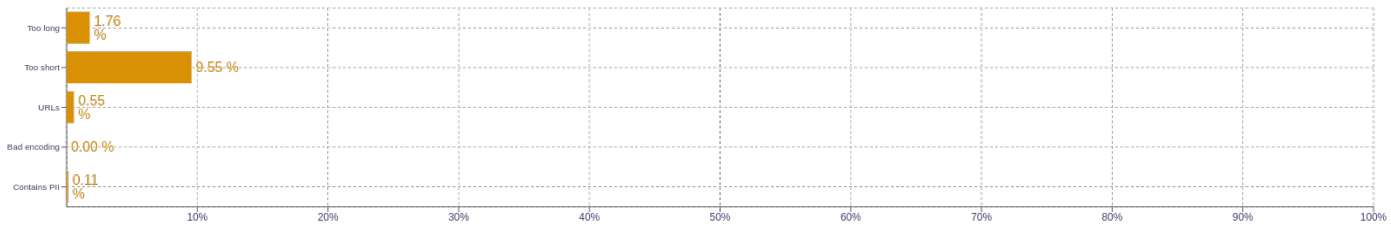


Segment length distribution by token

<= 49 tokens = **1.4M** segments | **840K** duplicates
> 50 tokens = **593K** segments | **135K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	594656 | نه, 378486 | جا, 333327 | ڪيو, 253265 | ا, 221402 | سنڌ
2	27683 | ڪٿي وڃي, 25821 | اسلام آباد, 22272 | ڪيو ويندو, 21735 | ڪيو وڃي, 19851 | ويب ڊيسڪ
3	6571 | پي ٽي آءِ, 4532 | ايس اي او, 4435 | استعمال ڪيو ويندو, 4351 | پيداوار جي تفصيل, 3735 | ميڊيا سان ڳالهائيندي
4	2334 | سنڌ جي وڏي وزير, 2121 | قومي اسمبلي جو اجلاس, 2103 | اسمبلي جو اجلاس طلب, 2047 | سائين جي ايم سيد, 1874 | سيد مراد علي شاهه
5	2103 | قومي اسمبلي جو اجلاس طلب, 2102 | وزيراعظم جي اعتماد جو ووٽ, 1758 | جواب شامل ڪريجواب منسوخ ڪريو, 1637 | ڊيليو ڊيليو ڊيليو ڊيليو ڊيليو, 1489 | وڏي وزير سيد مراد علي

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