HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
uig_Arab.jsonl.tsv	12/13/2024	Uyghur (ug)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
442,397	8,982,392	4,386,967 (48.84 %)	274M	1,738,795,684	3.02 GB

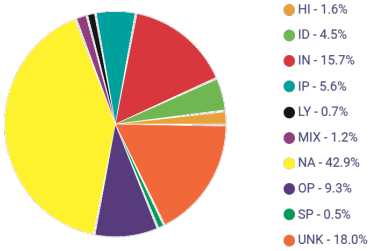
Top 10 domains

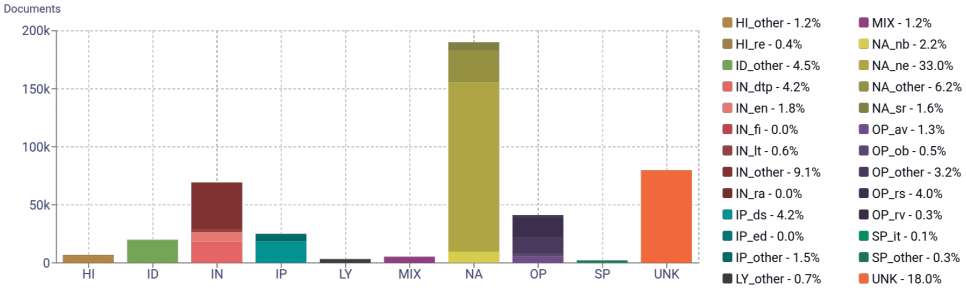
Domain	Docs	% of total
people.com.cn	36K	8.20%
ts.cn	20K	4.44%
okyan.com	18K	4.18%
nur.cn	13K	2.88%
misranim.com	11K	2.49%
izdinix.com	11K	2.42%
chinabroadcast.cn	11K	2.41%
rfa.org	10K	2.27%
karwan.cn	9.1K	2.07%
wikipedia.org	8.3K	1.87%

Top 10 TLDs

Domain	Docs	% of total
com	201K	45.33%
cn	121K	27.26%
com.cn	40K	9.02%
org	27K	6.05%
kz	15K	3.43%
net	10K	2.25%
biz	9.1K	2.05%
net.tr	3.6K	0.81%
cc	3.4K	0.76%
info	3.1K	0.70%

Register labels

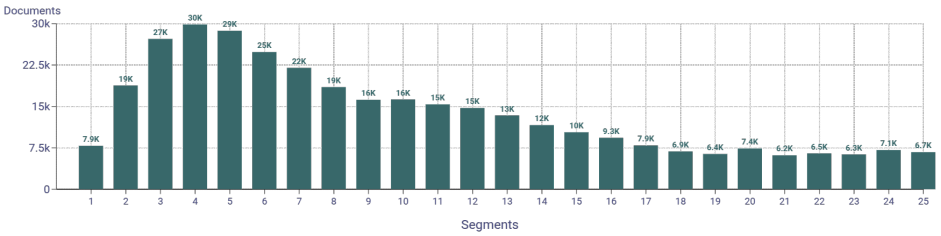




MT:4.7% | 21K Documents

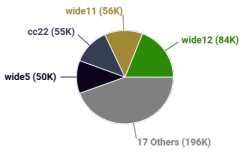
Documents size (in segments)

<= 25 segments **78.29%** (346K documents)
> 25 segments **21.71%** (96K documents)



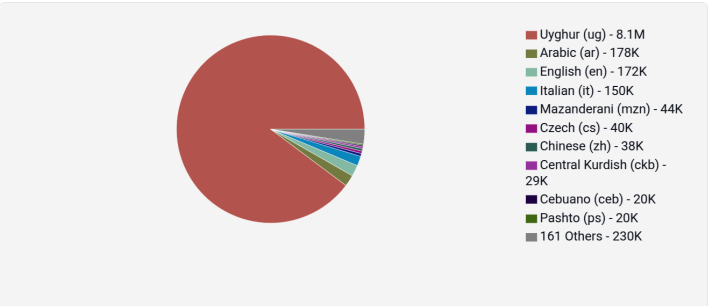
Documents by collection

CC = 24.56%
IA = 75.44%

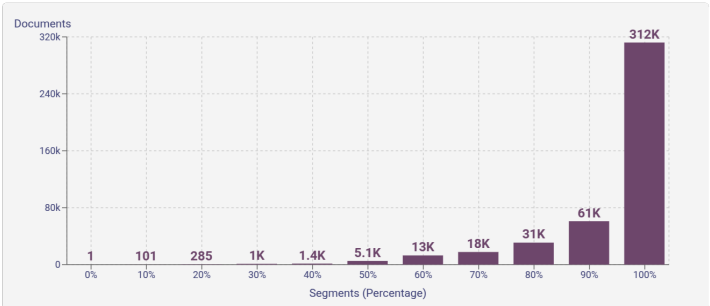


Language Distribution

Number of segments in the Uyghur (ug) corpus

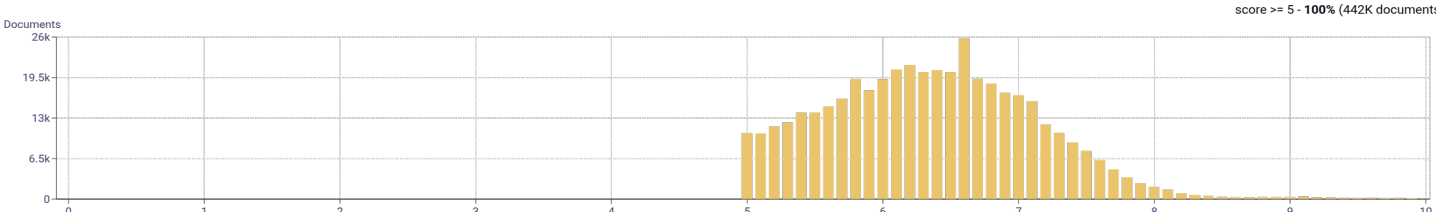


Percentage of segments in Uyghur (ug) inside documents



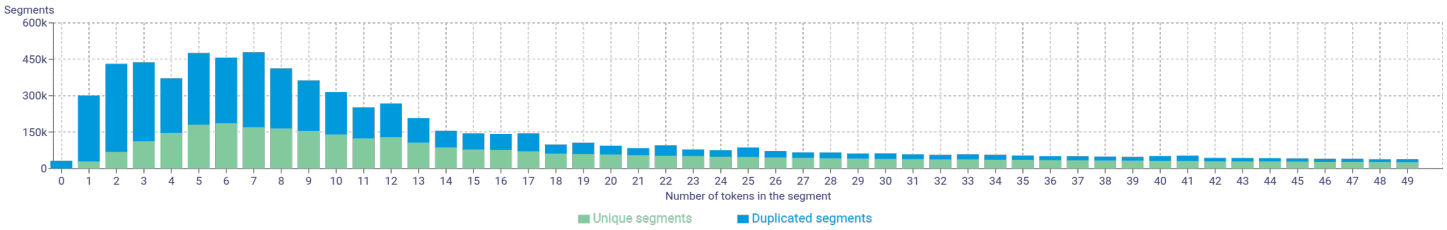
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (442K documents)

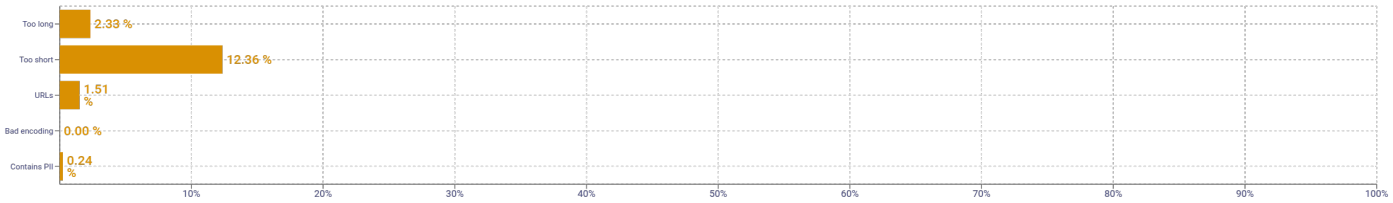


Segment length distribution by token

≤ 49 tokens = **3.3M** segments | **4.1M** duplicates
> 50 tokens = **1.6M** segments | **519K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	1714127 | بىلەن · 833294 | يولۇپ · 750398 | دەپ · 712248 | مەن · 699952 | ئۇ
2	70494 | شۇنىڭ بىلەن · 50144 | مۇنداق دېدى · 47549 | تاپىتونوم رايونلۇق · 44314 | خەلق تورى · 39666 | ھەر خىل
3	25576 | ۋقىمى حالىق تورايىنا · 25576 | حالىق تورايىنا ءتان · 19700 | شىنجاڭ ئۇيغۇر ئاپتونوم · 19558 | مەزمۇنلار پۈتۈنلەي مىسرنىم · 19550 | پۈتۈنلەي مىسرنىم مۇنبىرىدىن
4	25576 | ۋقىمى حالىق تورايىنا ءتان · 19550 | مەزمۇنلار پۈتۈنلەي مىسرنىم مۇنبىرىدىن · 19539 | پۈتۈنلەي مىسرنىم مۇنبىرىدىن كۆچۈرۈلگەن · 12878 | مەنشىك ۋقىمى حالىق تورايىنا · 12878 | بەتتىك مەنشىك ۋقىمى حالىق
5	19539 | مەزمۇنلار پۈتۈنلەي مىسرنىم مۇنبىرىدىن كۆچۈرۈلگەن · 12878 | مەنشىك ۋقىمى حالىق تورايىنا ءتان · 12878 | بەتتىك مەنشىك ۋقىمى حالىق تورايىنا · 12878 | بۆل بەتتىك مەنشىك ۋقىمى حالىق · 12694 | ەگەر ياسقارۇ جاعىنا پىكىرىڭىز بولسا

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