HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
tuk_Latn.jsonl.tsv	11/27/2024	Turkmen (tk)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
171,037	3,355,083	1,961,128 (58.45 %)	89M	594.25 MB	566,811,877

Top 10 domains

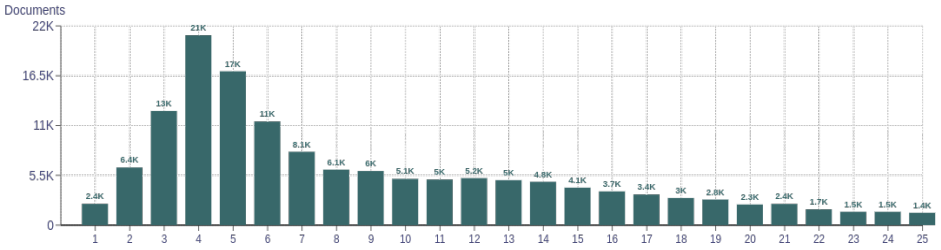
Domain	Docs	% of total
azathabar.com	56K	32.48
ertir.com	6.6K	3.88
atavatan-turkmenistan.com	6.5K	3.80
wikipedia.org	6.5K	3.77
turkmenportal.com	6.1K	3.57
inform.kz	5.5K	3.22
egemen.kz	3.3K	1.91
medeniyet.gov.tm	2.9K	1.71
business.com.tm	2.9K	1.69
turkmenistan.gov.tm	2.9K	1.68

Top 10 TLDs

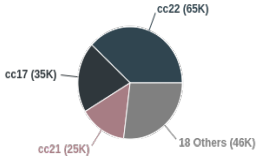
Domain	Docs	% of total
com	99K	58.03
gov.tm	24K	13.88
kz	13K	7.84
org	12K	6.80
com.tm	5.3K	3.10
info	5.1K	2.99
net.tr	2.7K	1.57
tm	1.1K	0.64
edu.tm	1.1K	0.64
ir	1.1K	0.63

Documents size (in segments)

<= 25 segments **84.13%** (144K documents)
> 25 segments **15.87%** (27K documents)

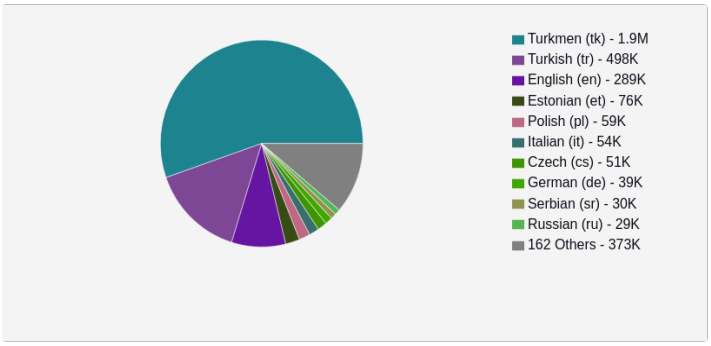


Documents by collection

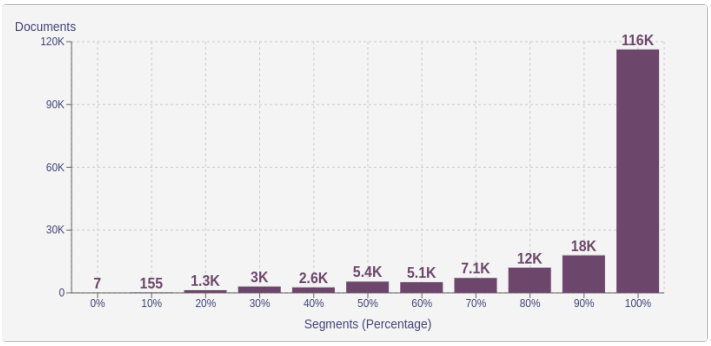


Language Distribution

Number of segments

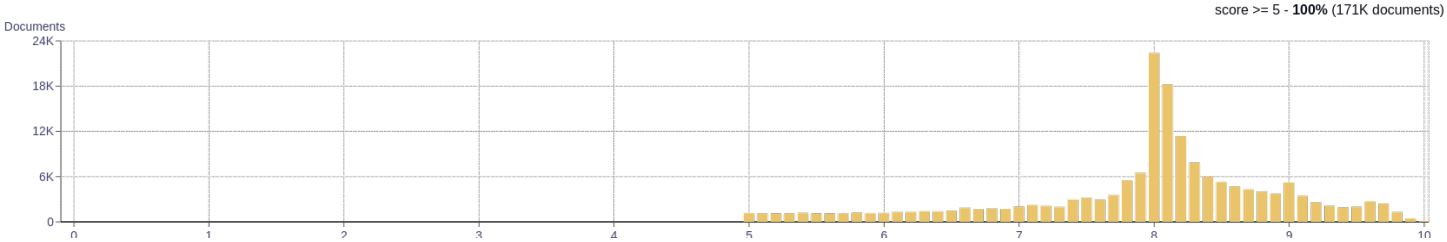


Percentage of segments in Turkmen (tk) inside documents



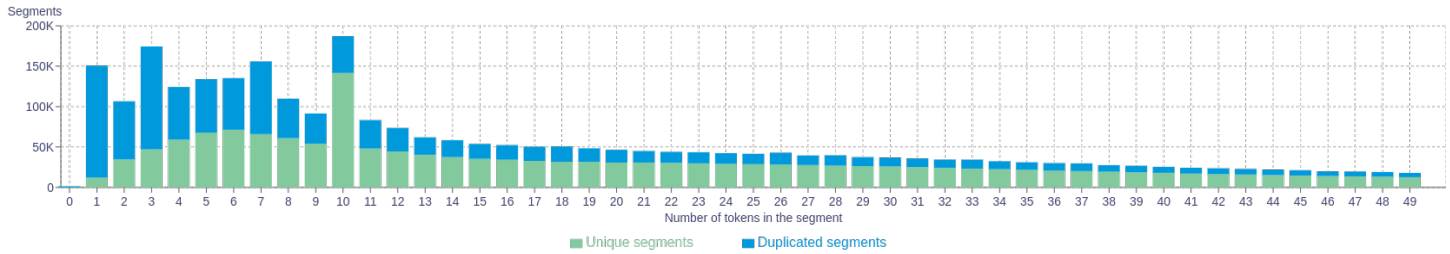
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (171K documents)

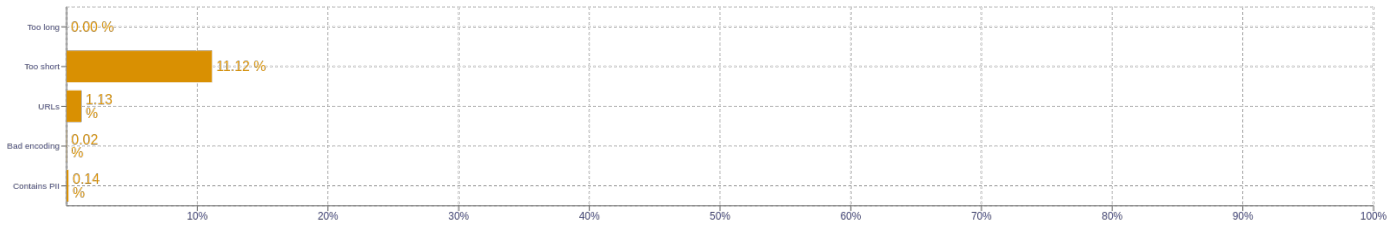


Segment length distribution by token

<= 49 tokens = **1.6M** segments | **1.2M** duplicates
> 50 tokens = **492K** segments | **144K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	we | 1302507 · bilen | 755095 · hem | 554741 · bu | 536344 · üçin | 448763
2	şeýle hem | 67810 · nji ýylyň | 52851 · nji ýylda | 52191 · hormatly prezidentimiz | 47294 · gurbanguly berdimuhamedow | 34061
3	hormatly prezidentimiz gurbanguly | 29168 · prezidentimiz gurbanguly berdimuhamedow | 18804 · şunuň bilen baglylykda | 11135 · türkmenistanyň prezidenti gurbanguly | 9286 · ministrler kabinetiniň başlygynyň | 9215
4	hormatly prezidentimiz gurbanguly berdimuhamedow | 18721 · hormatly prezidentimiz gurbanguly berdimuhamedowyň | 8685 · ministrler kabinetiniň başlygynyň orunbasary | 7009 · şol bir wagtyň özünde | 6190 · türkmenistanyň prezidenti gurbanguly berdimuhamedow | 5802
5	fmly fmly fmly fmly fmly | 4902 · binalarda we söweş sungaty boýunça | 3780 · ýapyk binalarda we söweş sungaty | 3779 · saglygy goraýyş we derman senagaty | 3565 · we söweş sungaty boýunça v | 3319

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