HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
HPLT-docslite.tr.tsv	6/23/2024	Turkish (tr)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
27,051,591	3,866,382,084	1,174,862,485 (30.39 %)	48B	294.92 GB	

Top 10 domains

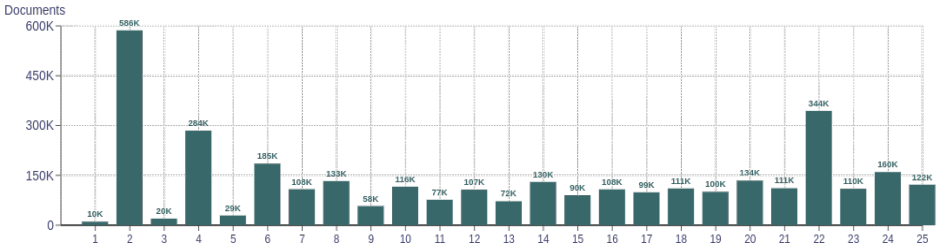
Domain	Docs	% of total
blogspot.com.tr	1.3M	4.88
alibaba.com	516K	1.91
dhgate.com	257K	0.95
haberx.com	208K	0.77
aliexpress.com	172K	0.64
docplayer.biz.tr	163K	0.60
blogspot.com	143K	0.53
nosorgulama.com	112K	0.41
blogspot.de	93K	0.34
diebuchsuche.com	93K	0.34

Top 10 TLDs

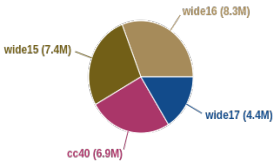
Domain	Docs	% of total
com	17M	64.31
com.tr	2.5M	9.18
net	2.1M	7.79
org	1.3M	4.93
gen.tr	256K	0.95
info	238K	0.88
org.tr	215K	0.79
biz.tr	210K	0.78
xyz	190K	0.70
biz	177K	0.66

Documents size (in segments)

<= 25 segments **12.6%** (3.4M documents)
> 25 segments **87.4%** (24M documents)

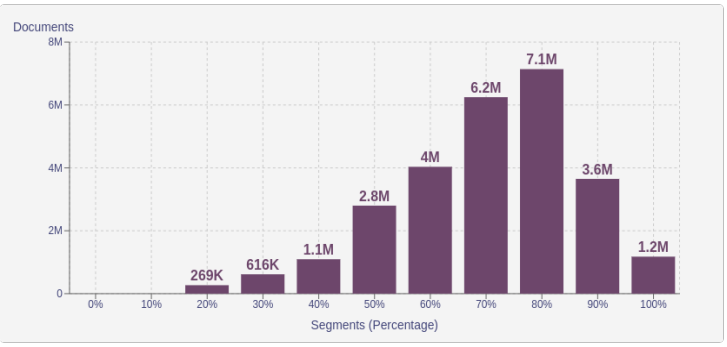


Documents by collection



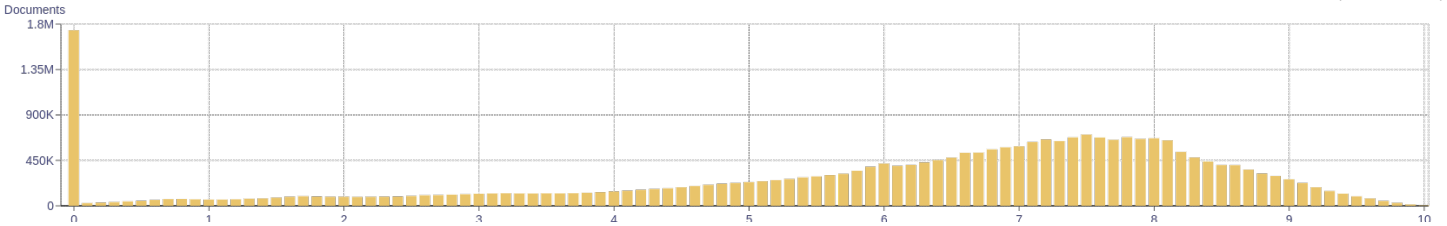
Language Distribution

Percentage of segments in Turkish (tr) inside documents



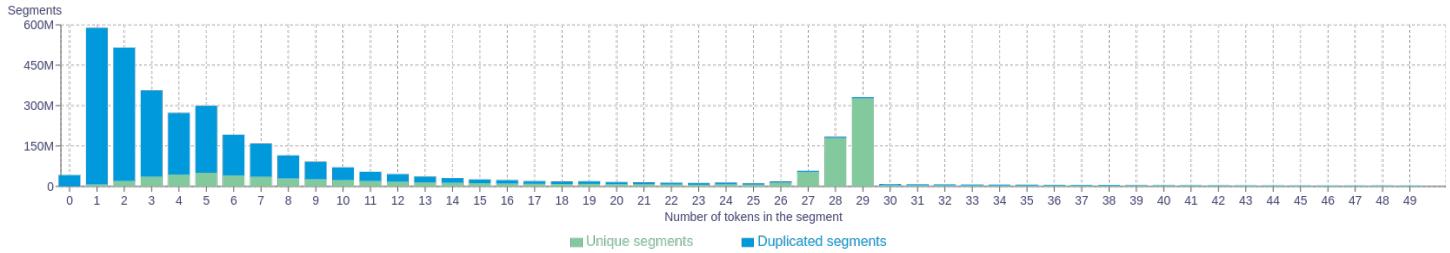
Distribution of documents by document score

score < 5 - **25.93%** (7M documents)
score >= 5 - **74.07%** (20M documents)

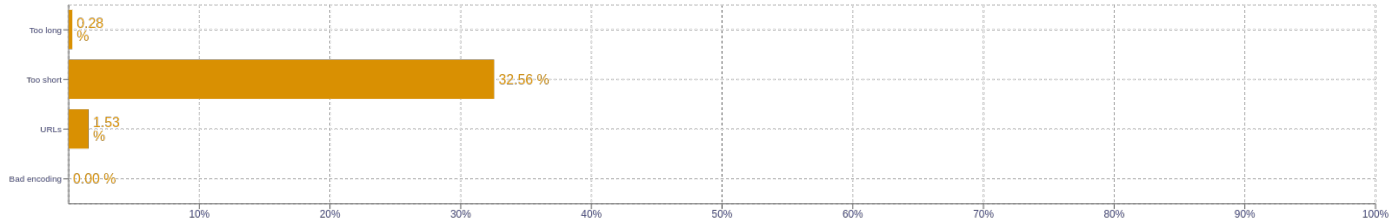


Segment length distribution by token

<= 49 tokens = **1.1B** segments | **2.7B** duplicates
> 50 tokens = **108M** segments | **37M** duplicates



Segment noise distribution



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