HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
HPLT-docslite.he.tsv	6/9/2024	Hebrew (he)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
4,979,580	838,116,375	169,137,393 (20.18 %)	9B	65.29 GB	

Top 10 domains

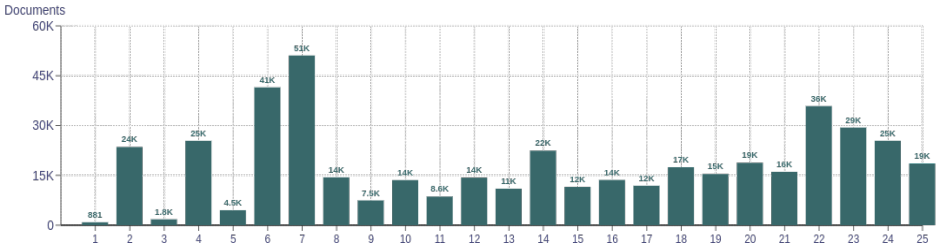
Domain	Docs	% of total
alibaba.com	543K	10.91
blogspot.co.il	237K	4.75
eip.co.il	170K	3.42
aliexpress.com	150K	3.01
diebuchsuche.com	66K	1.32
wikipedia.org	65K	1.30
infomed.co.il	50K	1.01
blogspot.com	45K	0.90
lightinthebox.com	44K	0.89
saloona.co.il	31K	0.63

Top 10 TLDs

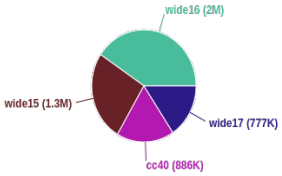
Domain	Docs	% of total
co.il	2M	40.58
com	2M	39.44
org.il	289K	5.81
org	251K	5.04
net	143K	2.87
ac.il	42K	0.84
info	31K	0.63
xyz	14K	0.27
biz	11K	0.23
co	11K	0.22

Documents size (in segments)

<= 25 segments **9.14%** (455K documents)
> 25 segments **90.86%** (4.5M documents)

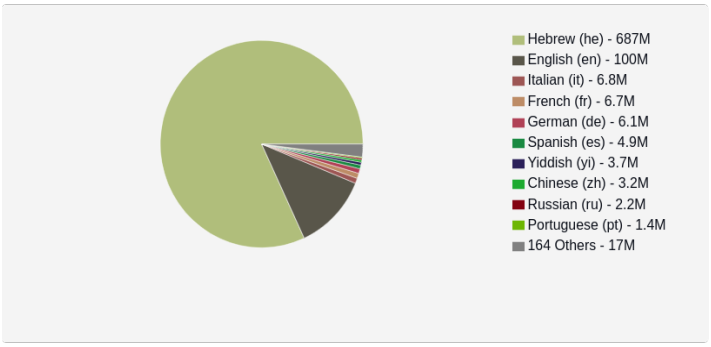


Documents by collection

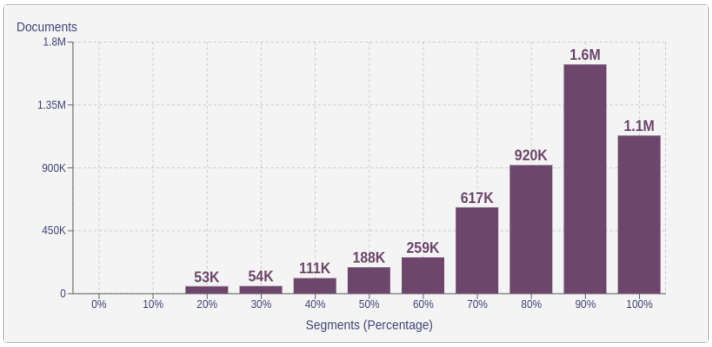


Language Distribution

Number of segments

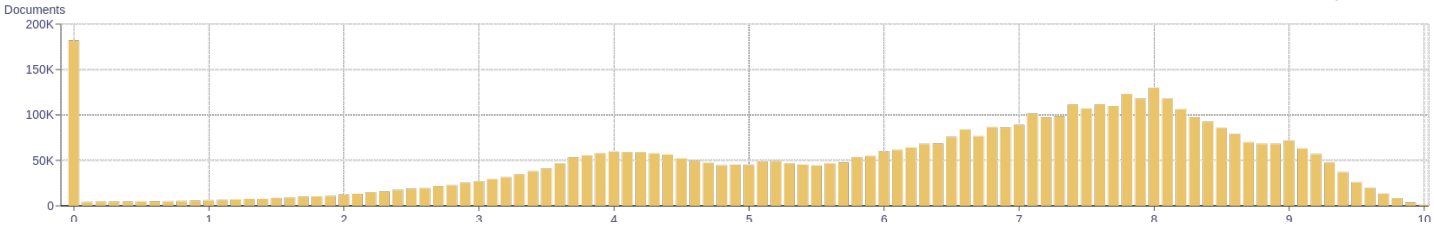


Percentage of segments in Hebrew (he) inside documents



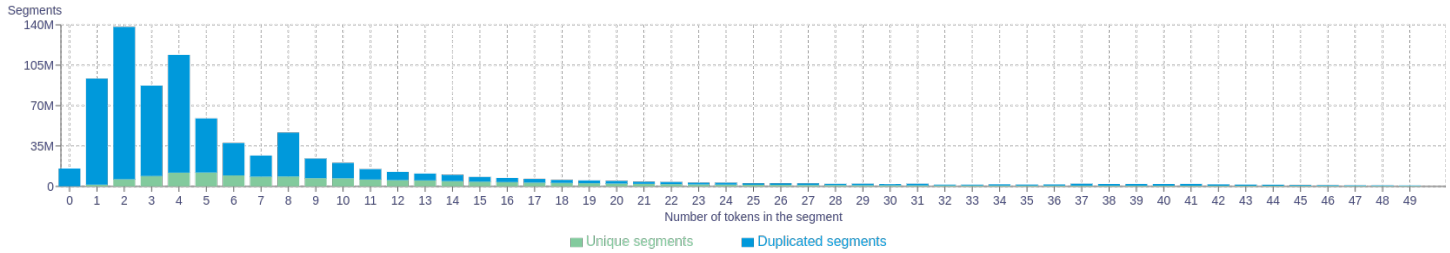
Distribution of documents by document score

score < 5 - **28.79%** (1.4M documents)
score >= 5 - **71.21%** (3.5M documents)

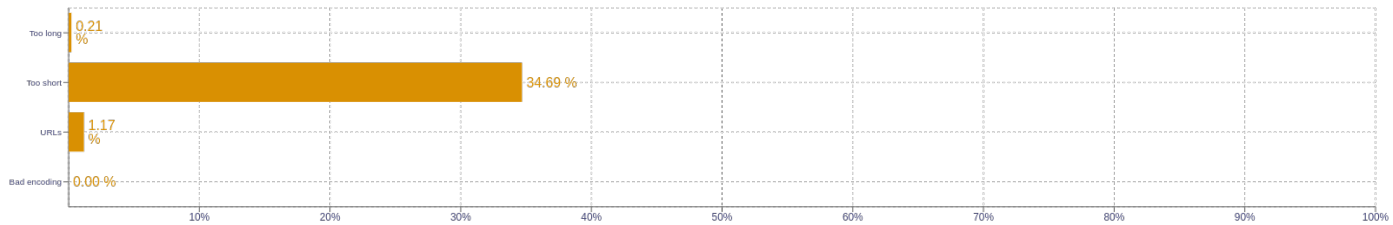


Segment length distribution by token

<= 49 tokens = **150M** segments | **661M** duplicates
> 50 tokens = **28M** segments | **8.5M** 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