HPLT Analytics report

General overview

Corpus	Analytics date	Language
HPLT-docslite.uk.tsv	6/11/2024	Ukrainian (uk)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
9,306,382	1,077,427,089	271,582,584 (25.21 %)	14B	125.9 GB	

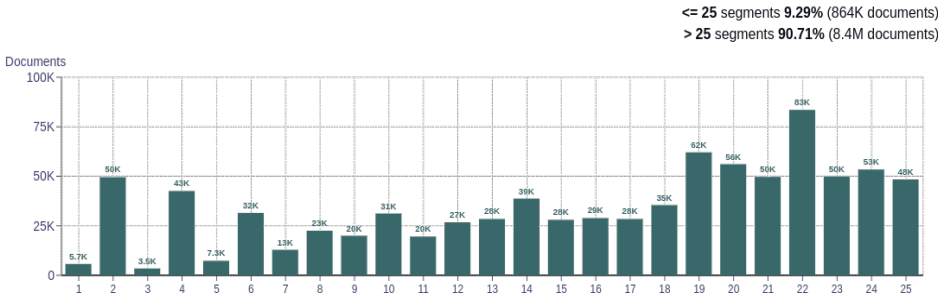### Top 10 domains

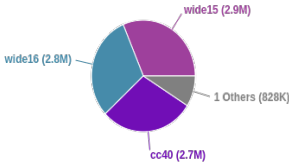| Domain	Docs	% of total
co.ua	312K	3.35
pp.ua	264K	2.84
diebuchsuche.com	103K	1.10
uapatents.com	77K	0.82
mistaua.com	68K	0.73
e-catalog.name	58K	0.63
wikipedia.org	58K	0.62
referatu.net.ua	56K	0.61
maidan.org.ua	55K	0.60
bibliograph.com.ua	54K	0.58

Top 10 TLDs

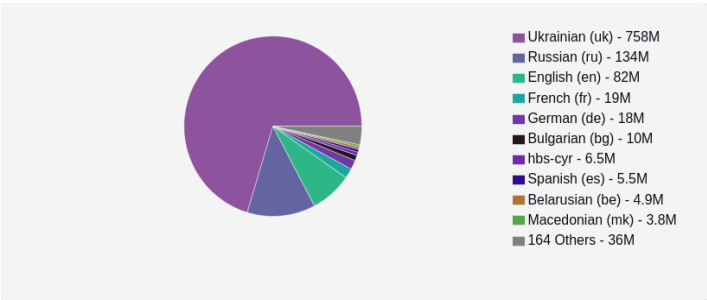| Domain	Docs	% of total
com	1.8M	19.72
ua	1.3M	13.91
com.ua	1.1M	12.34
org.ua	652K	7.01
org	441K	4.74
ru	430K	4.62
in.ua	403K	4.33
net	366K	3.94
info	361K	3.88
gov.ua	344K	3.69

Documents size (in segments)

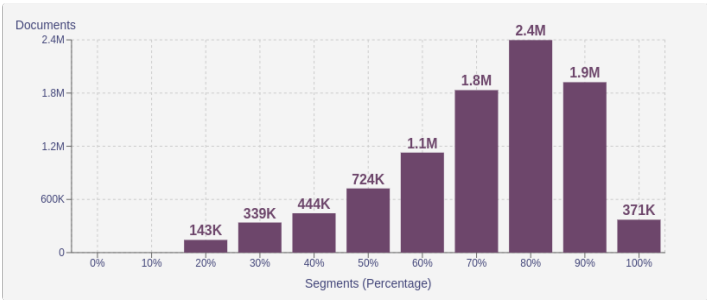<= 25 segments **9.29%** (864K documents)
> 25 segments **90.71%** (8.4M documents)

Documents by collection

wide15 (2.9M)
wide16 (2.8M)
1 Others (828K)
cc40 (2.7M)

Language Distribution

Number of segments

- Ukrainian (uk) - 758M
- Russian (ru) - 134M
- English (en) - 82M
- French (fr) - 19M
- German (de) - 18M
- Bulgarian (bg) - 10M
- hbs-cyr - 6.5M
- Spanish (es) - 5.5M
- Belarusian (be) - 4.9M
- Macedonian (mk) - 3.8M
- 164 Others - 36M

Percentage of segments in Ukrainian (uk) inside documents

Distribution of documents by document score

score <= 5 - **13.26%** (1.2M documents)
score > 5 - **86.74%** (8.1M documents)

Segment length distribution by token

<= 49 tokens = **226M** segments | **793M** duplicates
> 50 tokens = **59M** segments | **13M** duplicates

- Unique segments
- Duplicated segments
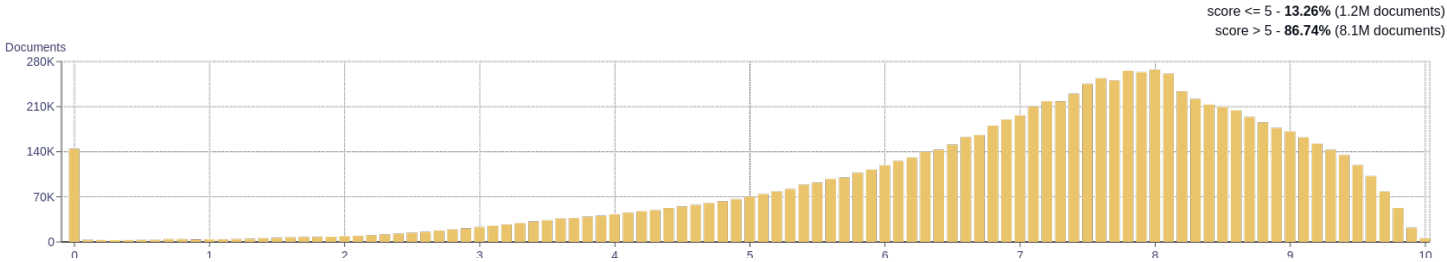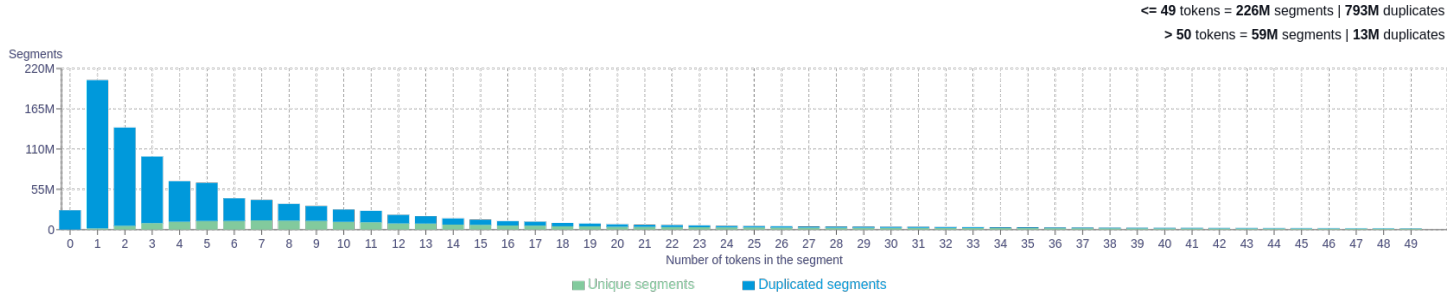
About HPLT Analytics

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