HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
HPLT-v2-ukr_Cyrl.tsv	9/24/2024	Ukrainian (uk)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
47,395,787	1,169,038,099			306.5 GB	181,746,006,077

Top 10 domains

Domain	Docs	% of total
wikipedia.org	858K	1.81
unian.ua	787K	1.66
uapatents.com	448K	0.95
korrespondent.net	425K	0.90
radiosvoboda.org	367K	0.77
co.ua	350K	0.74
24tv.ua	334K	0.70
pp.ua	325K	0.69
rada.gov.ua	301K	0.64
referatcentral.org.ua	298K	0.63

Top 10 TLDs

Domain	Docs	% of total
com	8.4M	17.75
ua	8.1M	17.04
com.ua	7.6M	16.04
org.ua	2.8M	5.88
org	2.8M	5.88
gov.ua	2.7M	5.76
net	1.9M	4.03
in.ua	1.7M	3.61
info	1.5M	3.14
ru	1.2M	2.57

Documents size (in segments)
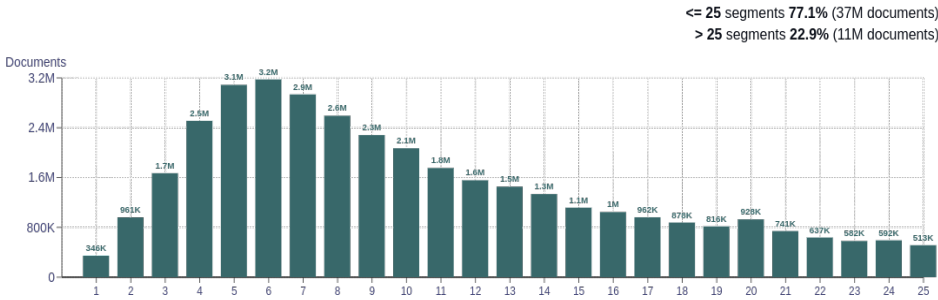
<= 25 segments **77.1%** (37M documents)
> 25 segments **22.9%** (11M documents)

Documents by collection

cc18 (9.5M)
cc22 (11M)
cc21 (5.4M)
18 Others (22M)
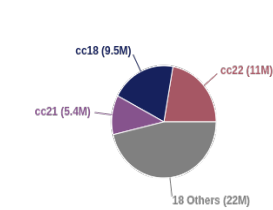
Language Distribution

Number of segments

- Ukrainian (uk) - 1.1B
- Russian (ru) - 36M
- Italian (it) - 17M
- English (en) - 17M
- German (de) - 5M
- French (fr) - 4.7M
- Serbian (sr) - 3M
- Belarusian (be) - 2.3M
- Macedonian (mk) - 1.4M
- Bulgarian (bg) - 1.2M
- 165 Others - 12M
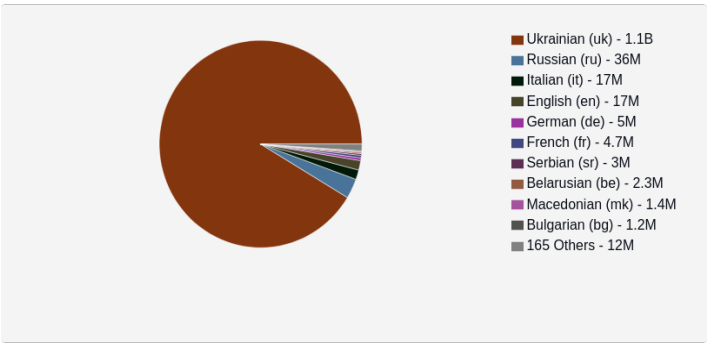
Percentage of segments in Ukrainian (uk) inside documents
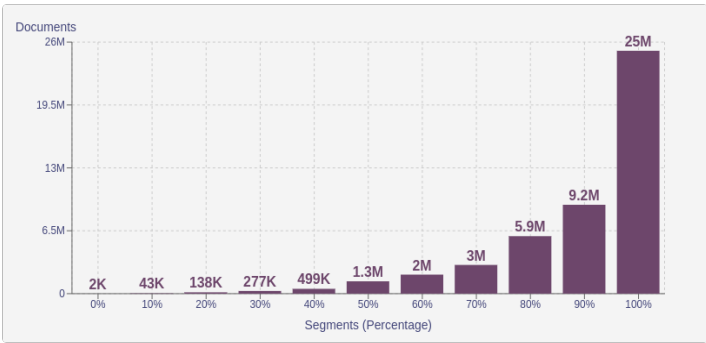
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (47M documents)

Segment noise distribution

Too long	0.73 %
Too short	12.36 %
URLs	0.91 %
Bad encoding	0.00 %
Contains PII	0.14 %
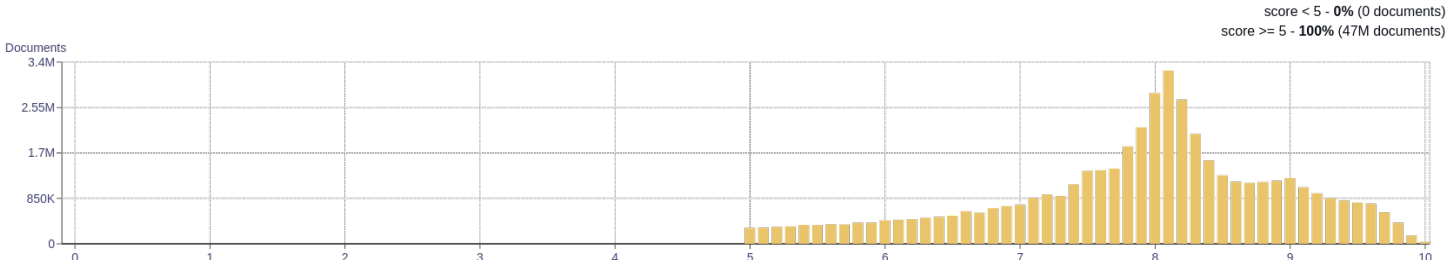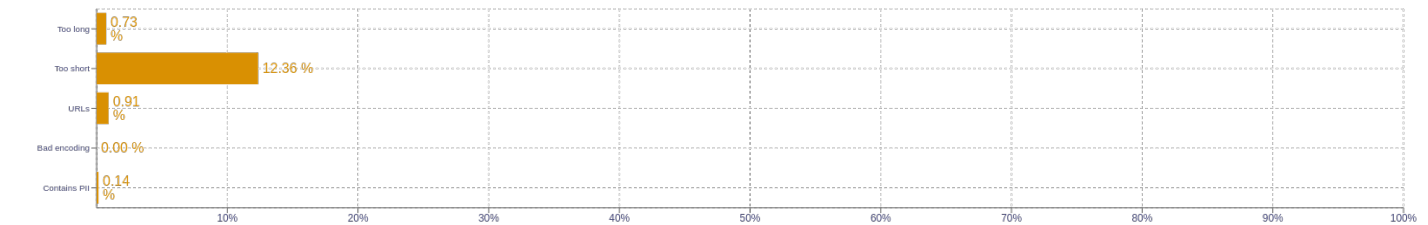
About HPLT Analytics

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