HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
ukr_Cyrl.jsonl.tsv	6/26/2025	Ukrainian (uk)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
47,395,779	1,168,793,792	537,002,359 (45.95 %)	32B	181,746,006,077	306.5 GB

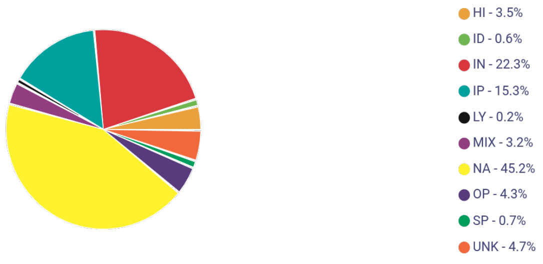
Top 10 domains

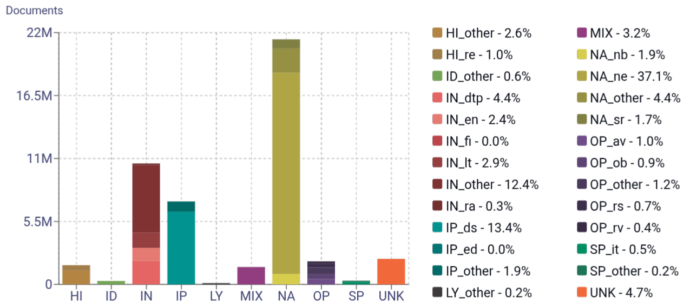
Domain	Docs	% of total
wikipedia.org	858K	1.81%
unian.ua	787K	1.66%
uapatents.com	448K	0.95%
korrespondent.net	425K	0.90%
radiosvoboda.org	367K	0.77%
co.ua	350K	0.74%
24tv.ua	334K	0.70%
pp.ua	325K	0.69%
rada.gov.ua	301K	0.64%
referatcentral....	298K	0.63%

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%

Register labels

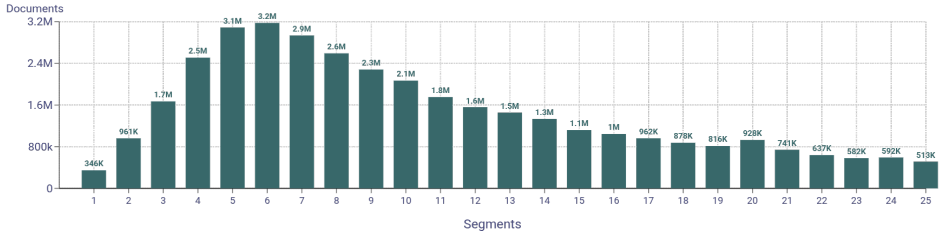




🤖 **MT**:1.2% | 573K Documents

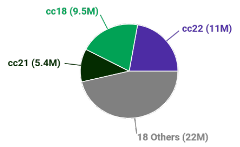
Documents size (in segments)

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



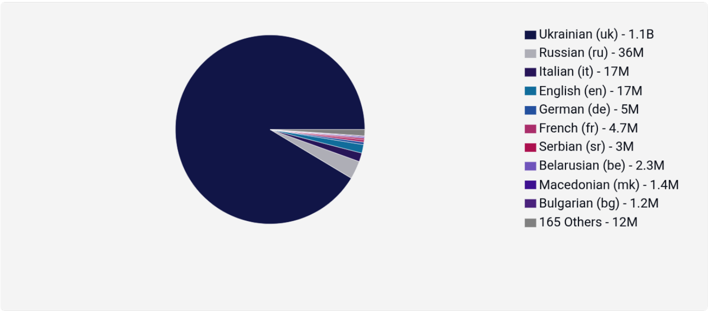
Documents by collection

CC = 65.40%
IA = 34.60%

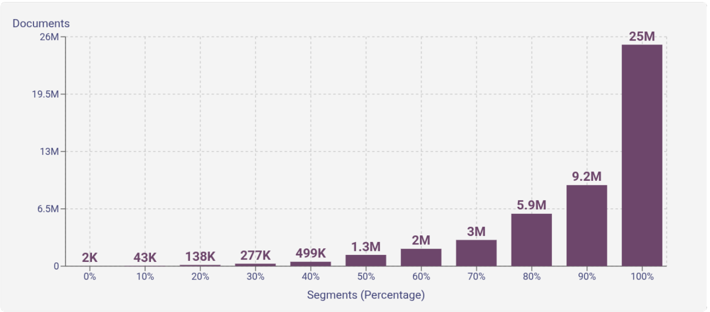


Language Distribution

Number of segments in the Ukrainian (uk) corpus

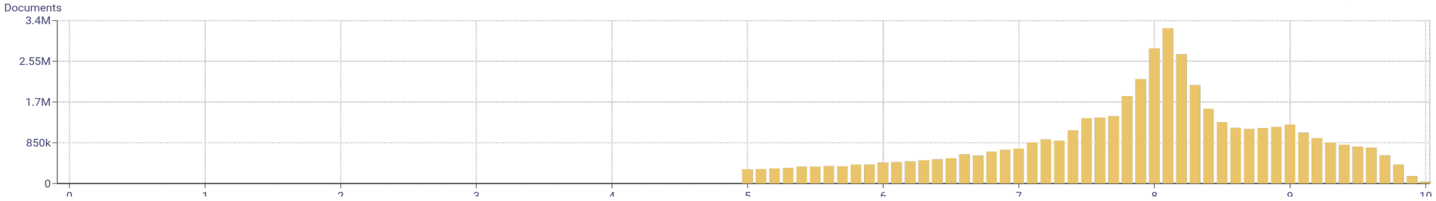


Percentage of segments in Ukrainian (uk) inside documents



Distribution of documents by document score

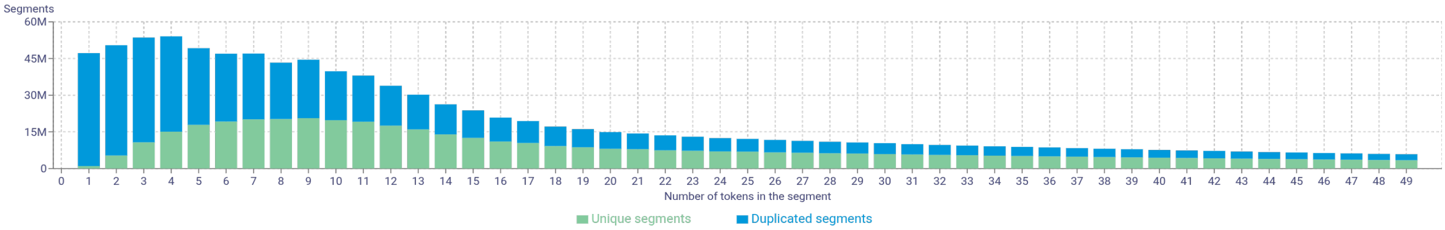
score < 5 - **0%** (0 documents)
score >= 5 - **100%** (47M documents)



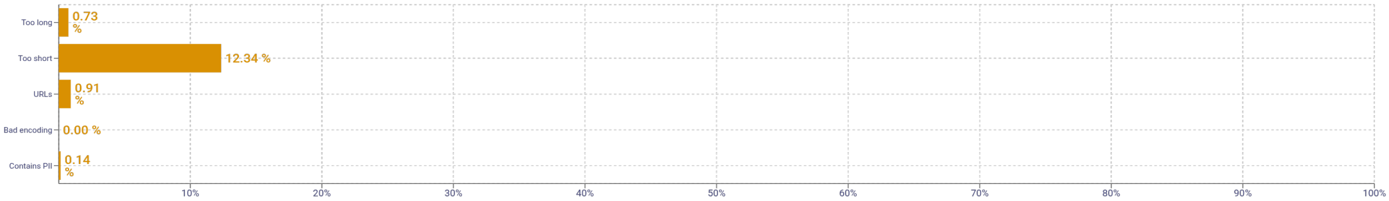
Segment length distribution by token

≤ 49 tokens = **990M** segments | **556M** duplicates

> 50 tokens = **179M** segments | **78M** duplicates



Segment noise distribution



Frequent n-grams

<table>
<tr><th>Size</th><th>n-grams</th></tr>
<tr><td>1</td><td>україни | 73192772
його | 60796919
при | 50653118
із | 47680333
він | 43800126</td></tr>
<tr><td>2</td><td>під час | 16668196
може бути | 8655692
при цьому | 8407996
крім того | 5494569
можуть бути | 4196695</td></tr>
<tr><td>3</td><td>верховної ради україни | 1532593
відповідно до закону | 657530
включає в себе | 626415
другої світової війни | 595833
автономної республіки крим | 497519</td></tr>
<tr><td>4</td><td>внесення змін до деяких | 396244
відповідно до закону україни | 332998
внесення змін до закону | 213279
вітаю тебе з днем | 208832
внесеними згідно із законом | 180985</td></tr>
<tr><td>5</td><td>внесення змін до деяких законодавчих | 250845
деяких законодавчих актів україни щодо | 191895
внесення змін до закону україни | 187395
вітаю тебе з днем народження | 174670
верховної ради україни з питань | 149134</td></tr>
</table>

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

Register labels

Name	Abbr.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	**HI**
Recipe	re
Informational persuasion	**IP**
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	**IN**
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	**OP**
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av