HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
kk_1.jsonl.tsv	3/22/2024	Kazakh (kk)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
406,351	51,693,572	15,387,737 (29.77 %)	612M	6.1 GB	

Top 10 domains

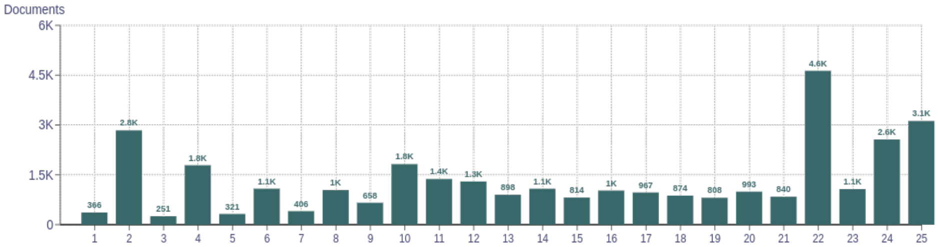
Domain	Docs	% of total
kodeksy-kz.com	18K	4.34
nur.kz	14K	3.46
game-game.kz	8.4K	2.07
wikipedia.org	8.3K	2.05
bilimdiler.kz	7.9K	1.94
zan.kz	6.2K	1.53
alashainasy.kz	5.9K	1.45
lektsii.org	5.9K	1.44
qazaquni.kz	4.9K	1.21
mylektsii.ru	4.3K	1.06

Top 10 TLDs

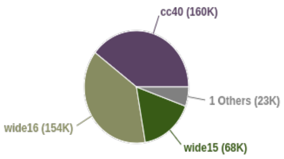
Domain	Docs	% of total
kz	268K	65.98
com	42K	10.24
org	32K	7.91
ru	19K	4.66
gov.kz	11K	2.79
info	9.1K	2.23
net	6.9K	1.70
ua	2.8K	0.69
edu.kz	2.3K	0.55
news	1.9K	0.46

Documents size (in segments)

<= 25 segments **8.14%** (33K documents)
> 25 segments **91.86%** (371K documents)

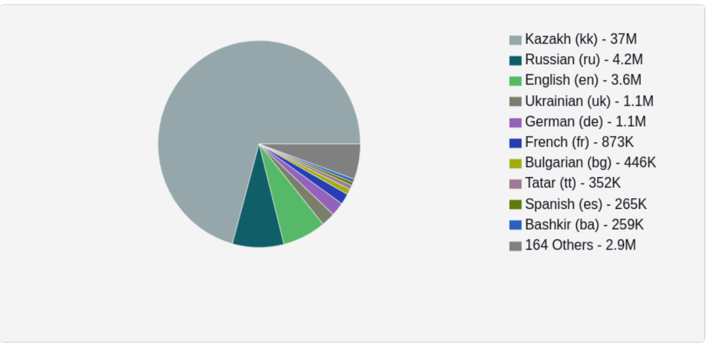


Documents by collection

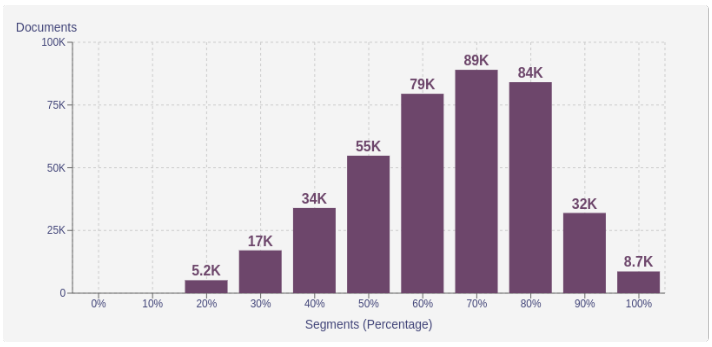


Language Distribution

Number of segments

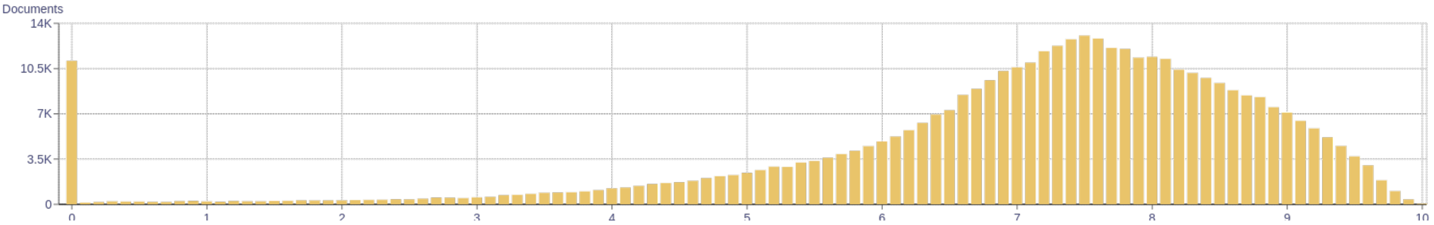


Percentage of segments in Kazakh (kk) inside documents



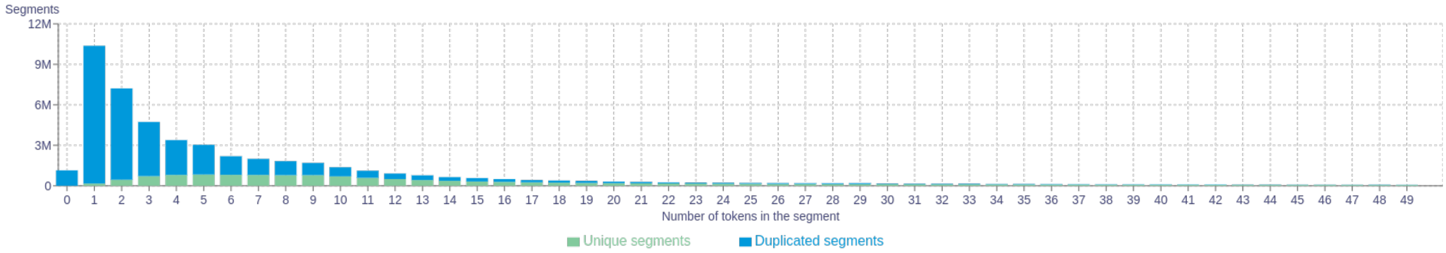
Distribution of documents by document score

score < 5 - **11.04%** (45K documents)
score >= 5 - **88.96%** (362K documents)

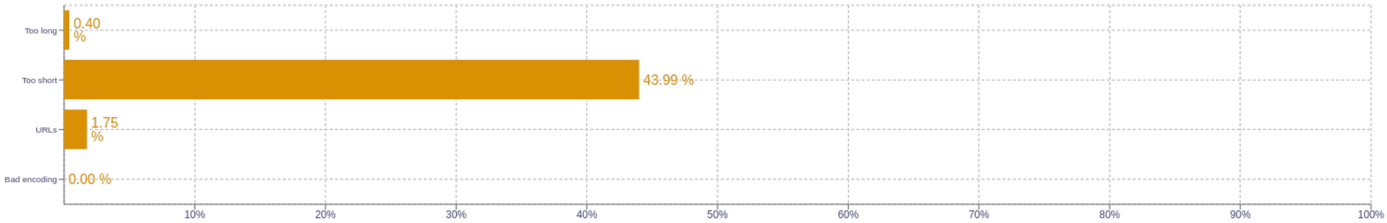


Segment length distribution by token

<= 49 tokens = **13M** segments | **36M** duplicates
> 50 tokens = **2.4M** segments | **425K** duplicates



Segment noise distribution



Frequent n-grams

<table>
<tr><th>Size</th><th>n-grams</th></tr>
<tr><td>1</td><td>және | 5609164
қазақстан | 1833616
да | 1401038
бойынша | 1400474
бір | 1366271</td></tr>
<tr><td>2</td><td>қазақстан республикасының | 556431
қазақстан республикасы | 385886
білім беру | 268447
болып табылады | 248914
басқа да | 155028</td></tr>
<tr><td>3</td><td>сыбайлас жемқорлыққа қарсы | 63128
өткен соң қолданысқа | 51531
және басқа да | 51024
білім және ғылым | 50872
қазақстан республикасы үкіметінің | 47140</td></tr>
<tr><td>4</td><td>күн өткен соң қолданысқа | 48564
өткен соң қолданысқа енгізіледі | 39894
алғашқы ресми жарияланған күнінен | 35181
күнтізбелік он күн өткен | 34256
тарих және география пәнінен | 29632</td></tr>
<tr><td>5</td><td>он күн өткен соң қолданысқа | 43772
күн өткен соң қолданысқа енгізіледі | 37735
тарих және география пәнінен үзд | 29621
ресми жарияланған күнінен кейін күнтізбелік | 23392
күнінен кейін күнтізбелік он күн | 20478</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