HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
uz_1.jsonl.tsv	3/20/2024	Uzbek (uz)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
290,289	37,680,934	10,728,950 (28.47 %)	435M	3.83 GB	

Top 10 domains

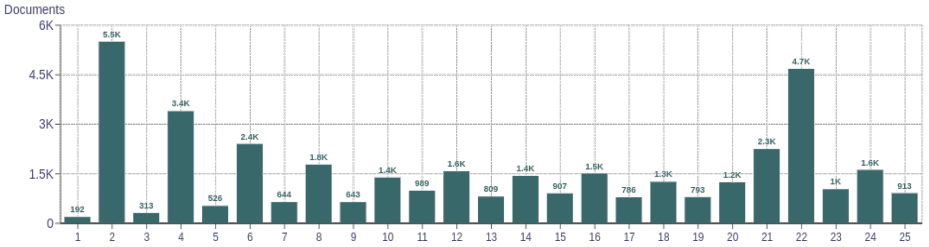
Domain	Docs	% of total
ziyouz.com	7.1K	2.45
fikr.uz	7.1K	2.43
wikipedia.org	6.3K	2.16
kun.uz	4.2K	1.43
shejot.com	4.1K	1.42
lex.uz	4.1K	1.41
game-game.uz	3.8K	1.31
kh-davron.uz	3.6K	1.24
gazeta.uz	3.1K	1.06
uzxalqharakati.com	2.8K	0.97

Top 10 TLDs

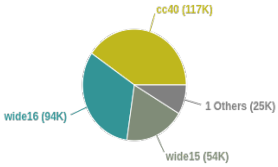
Domain	Docs	% of total
uz	174K	60.06
com	56K	19.32
org	17K	5.98
net	11K	3.63
ru	9.3K	3.21
info	2.8K	0.95
biz	1.6K	0.54
su	1.5K	0.53
katowice.pl	1.2K	0.40
asia	1.1K	0.39

Documents size (in segments)

<= 25 segments **13.3%** (39K documents)
> 25 segments **86.7%** (252K documents)

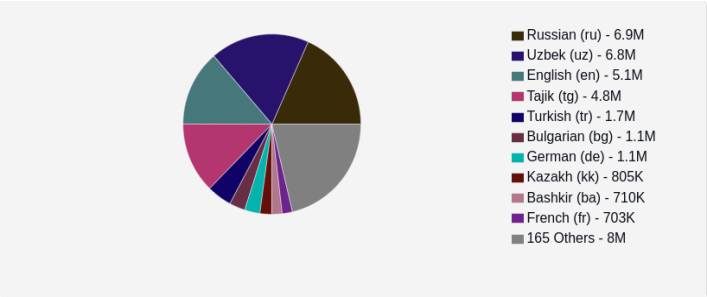


Documents by collection

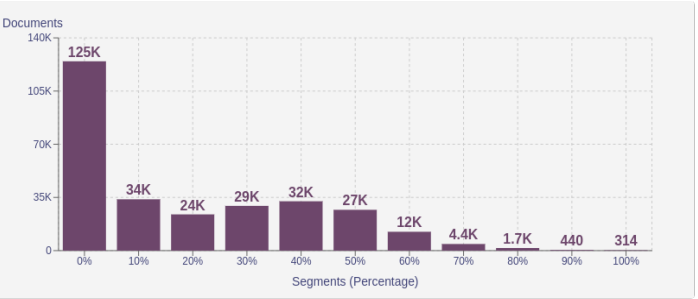


Language Distribution

Number of segments

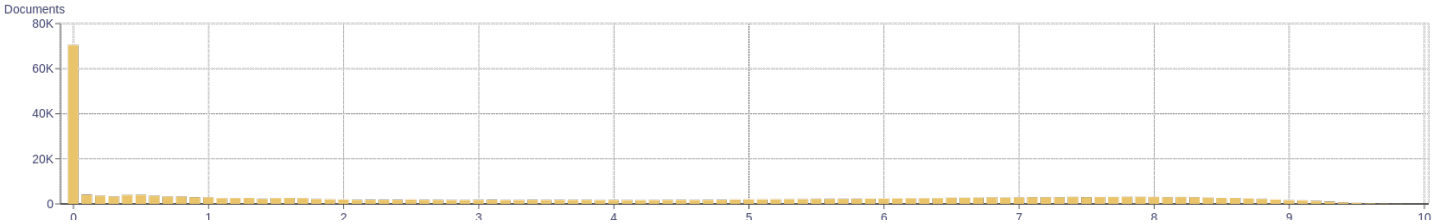


Percentage of segments in Uzbek (uz) inside documents



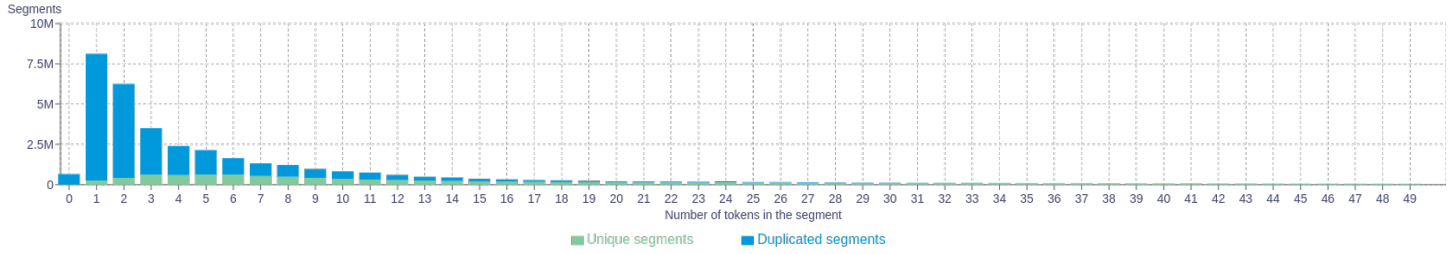
Distribution of documents by document score

score <= 5 - **61.72%** (179K documents)
score > 5 - **38.28%** (111K documents)

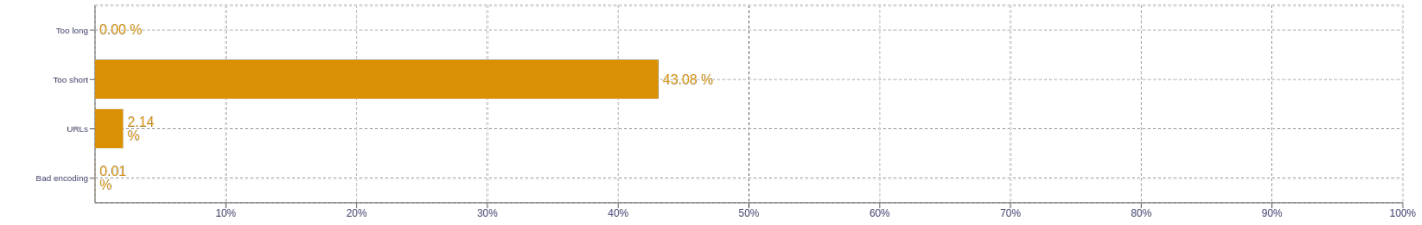


Segment length distribution by token

<= 49 tokens = **9.4M** segments | **27M** duplicates
> 50 tokens = **1.6M** segments | **266K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	ҳам | 1014840 · бир | 901803 · ўзбекистон | 766960 · у | 516526 · br | 489531
2	ўзбекистон республикаси | 351304 · марта ўқилган | 140347 · o‘zbekiston respublikasi | 93956 · ҳар бир | 93308 · bosh sahifa | 86715
3	ўзбекистон республикаси олий | 50967 · ўзбекистон республикаси вазирлар | 41753 · ўзбекистон республикаси президентининг | 40526 · ўзбекистон республикаси президенти | 30614 · республикаси вазирлар маҳкамасининг | 28360
4	ўзбекистон республикаси вазирлар маҳкамасининг | 28192 · ўзбекистон республикаси олий мажлиси | 22104 · соллаллоҳу алайҳи ва саллам | 18367 · бесплатные анимационные смайлики для | 14972 · анимационные смайлики для одноклассники.ру | 14947
5	бесплатные анимационные смайлики для одноклассники.ру | 14946 · турецкие сериалы онлайн на русском | 11705 · смотреть сериал великолепный век все | 11444 · сериал великолепный век все серии | 11444 · все серии на русском языке | 11444

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