HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
mk_1.jsonl.tsv	3/22/2024	Macedonian (mk)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
734,687	86,189,730	14,489,281 (16.81 %)	845M	7.59 GB	

Top 10 domains

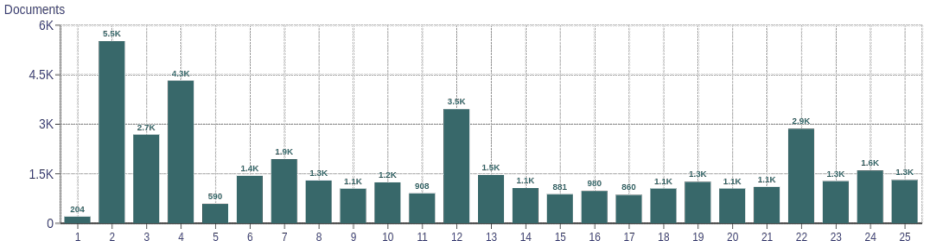
Domain	Docs	% of total
sport.com.mk	23K	3.19
daily.mk	20K	2.78
meta.mk	15K	2.10
kurir.mk	15K	2.07
wikipedia.org	15K	1.98
rbth.com	13K	1.81
voanews.com	12K	1.58
mia.mk	9.9K	1.35
nmd.mk	7.6K	1.03
press24.mk	7.1K	0.97

Top 10 TLDs

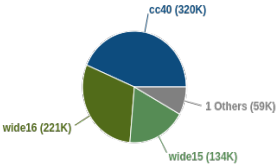
Domain	Docs	% of total
mk	428K	58.24
com	111K	15.14
com.mk	93K	12.62
org	26K	3.58
org.mk	21K	2.87
gov.mk	16K	2.12
net	7.7K	1.05
edu.mk	7.6K	1.03
news	3.3K	0.45
info	2.4K	0.32

Documents size (in segments)

<= 25 segments **5.64%** (41K documents)
> 25 segments **94.36%** (693K documents)

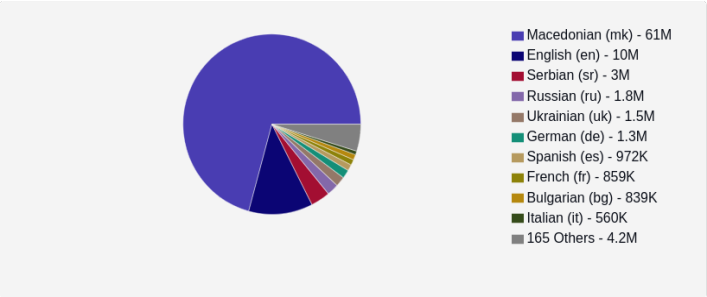


Documents by collection

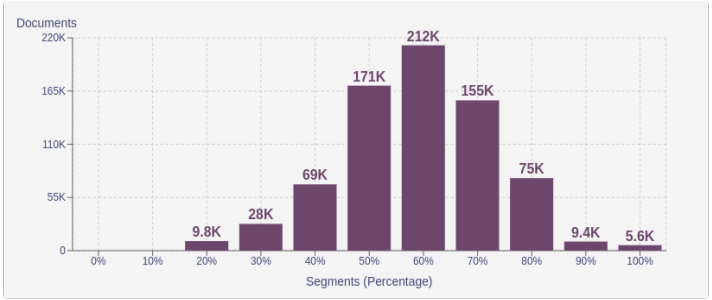


Language Distribution

Number of segments

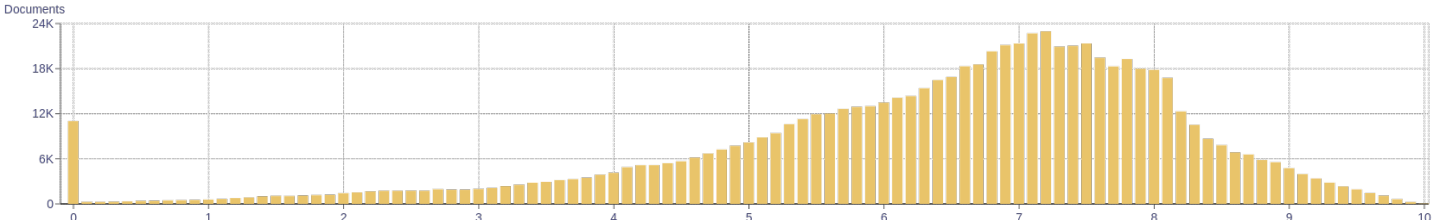


Percentage of segments in Macedonian (mk) inside documents



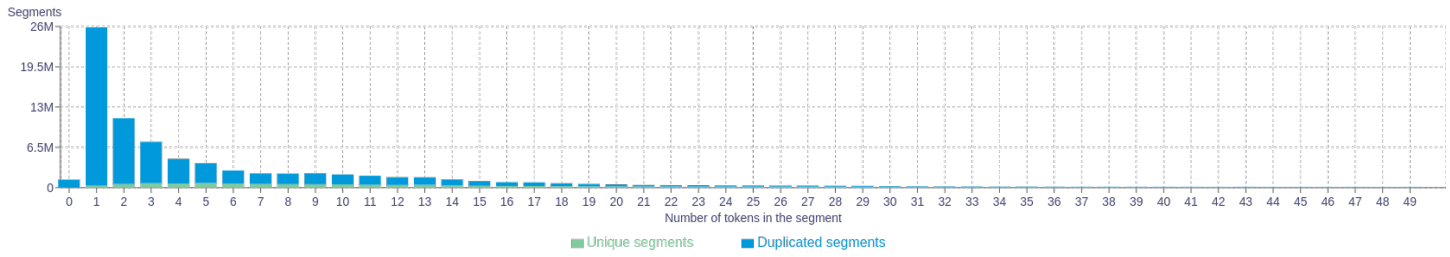
Distribution of documents by document score

score <= 5 - **17.45%** (128K documents)
score > 5 - **82.55%** (606K documents)

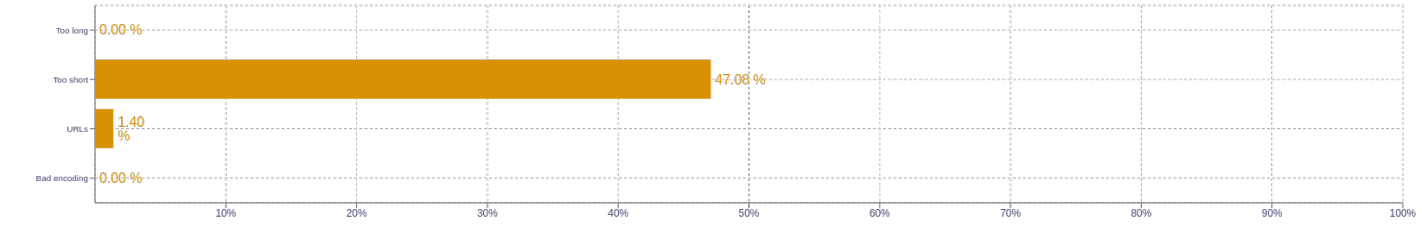


Segment length distribution by token

<= 49 tokens = **12M** segments | **71M** duplicates
> 50 tokens = **3M** segments | **810K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	македонија | 2546789 · вести | 1408817 · година | 1313749 · скопје | 1187774 · видео | 1138123
2	република македонија | 237582 · најнови вести | 171533 · of the | 142448 · милиони евра | 124160 · read more | 122921
3	права се задржани | 127603 · услови за користење | 120153 · љубов и секс | 96559 · all rights reserved | 83602 · skip to content | 68997
4	cookie is set by | 47536 · | 40187 · the user consent for | 40018 · user consent for the | 39935 · consent for the cookies | 39935
5	user consent for the cookies | 39935 · the user consent for the | 39935 · for the cookies in the | 39933 · consent for the cookies in | 39933 · the cookies in the category | 38982

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