HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
mkd_Cyrl.jsonl.tsv	9/24/2024	Macedonian (mk)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
3,565,647	57,008,331	27,635,192 (48.48 %)	1.7B	15.64 GB	9,386,182,083

Top 10 domains

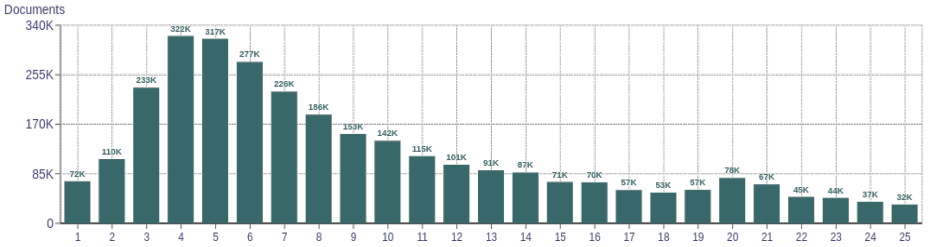
Domain	Docs	% of total
wikipedia.org	180K	5.05
slobodnaevropa.mk	79K	2.23
voanews.com	57K	1.61
kurir.mk	55K	1.53
daily.mk	53K	1.48
netpress.com.mk	42K	1.18
rbth.com	39K	1.10
hitportal.com.mk	39K	1.08
republika.mk	36K	1.02
kafepauza.mk	34K	0.94

Top 10 TLDs

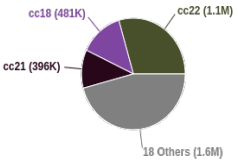
Domain	Docs	% of total
mk	2M	55.09
com	544K	15.27
com.mk	433K	12.14
org	263K	7.37
gov.mk	86K	2.40
org.mk	80K	2.25
edu.mk	39K	1.08
net	32K	0.90
info	14K	0.40
news	13K	0.36

Documents size (in segments)

<= 25 segments **85.34%** (3M documents)
> 25 segments **14.66%** (523K documents)

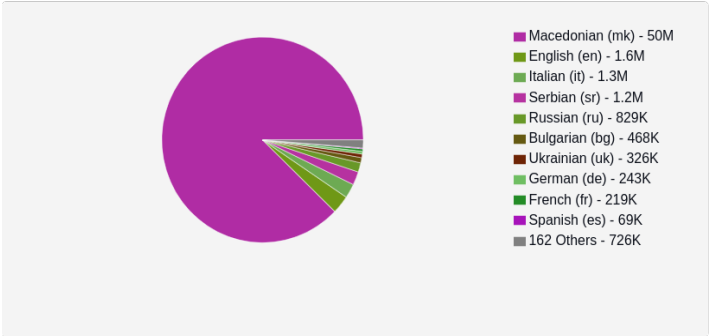


Documents by collection

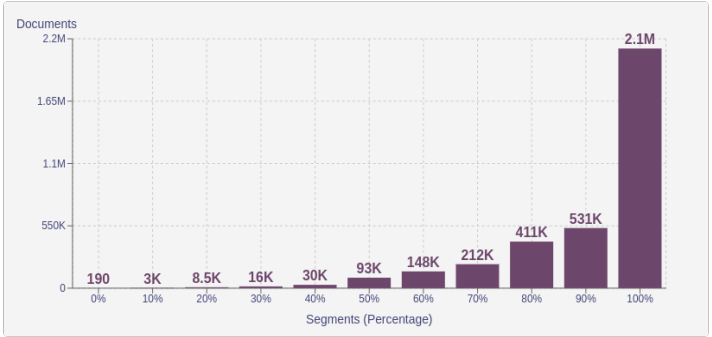


Language Distribution

Number of segments

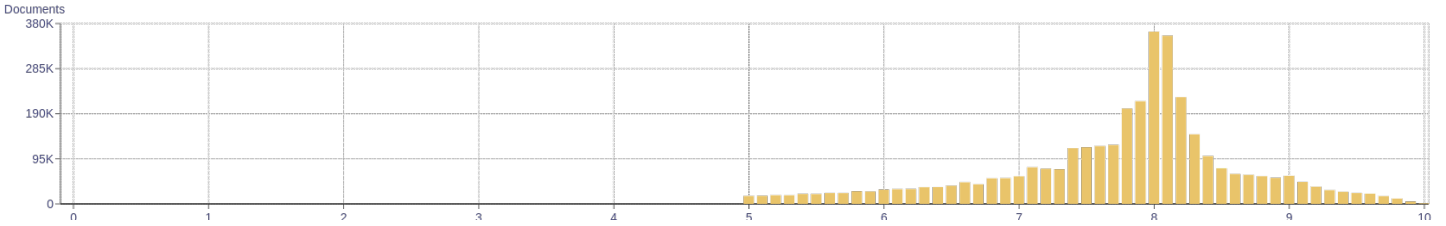


Percentage of segments in Macedonian (mk) inside documents



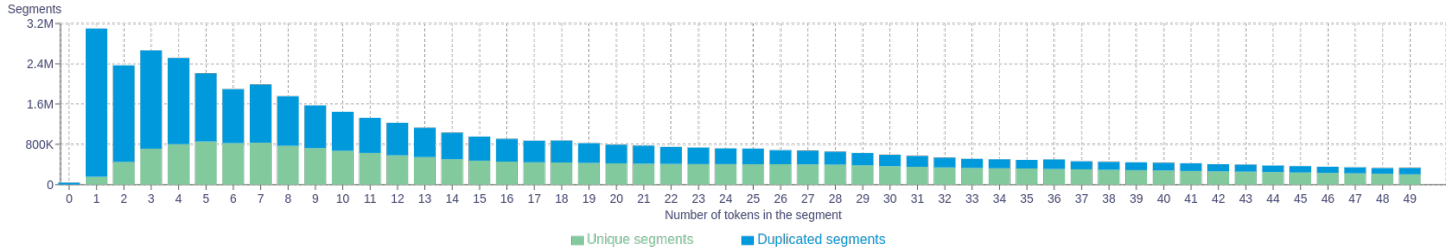
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (3.6M documents)

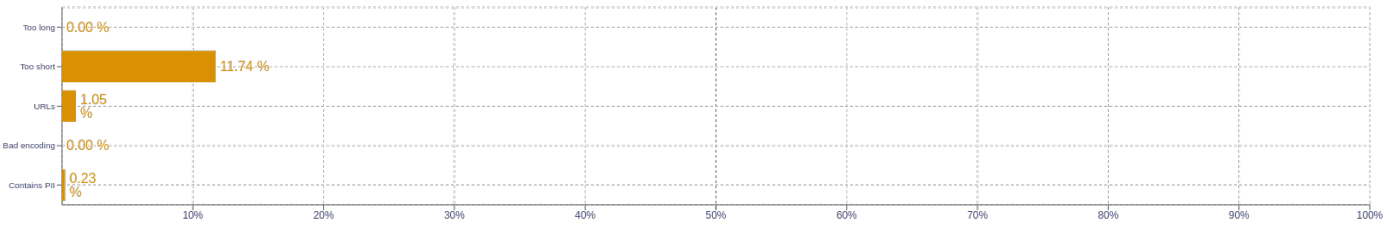


Segment length distribution by token

<= 49 tokens = **21M** segments | **26M** duplicates
> 50 tokens = **10M** segments | **3.8M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	година | 3930388 · македонија | 2807325 · време | 2074517 · години | 1767968 · дел | 1730470
2	република македонија | 539687 · уреди извор | 506624 · станува збор | 277149 · ве молиме | 245983 · голем број | 245318
3	република северна македонија | 88737 · можат да бидат | 85258 · лигата на шампионите | 60021 · втората светска војна | 59704 · владата на република | 57927
4	владата на република македонија | 41668 · министерството за внатрешни работи | 40710 · труд и социјална политика | 37252 · собранието на република македонија | 35902 · авторски текстови е казниво | 30930
5	текстови е казниво со закон | 30976 · неделата директно во вашето сандаче | 28528 · најдобрите стории на неделата директно | 28528 · дистрибуира во каква било форма | 25188 · писмена дозвола од македонската информативна | 25171

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