

General overview

Corpus	Analytics date	Language
kmr_Latn.jsonl.tsv	9/24/2024	Kurdish (kmr)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
364,347	7,147,414	4,165,409 (58.28 %)	228M	1.15 GB	1,116,200,141

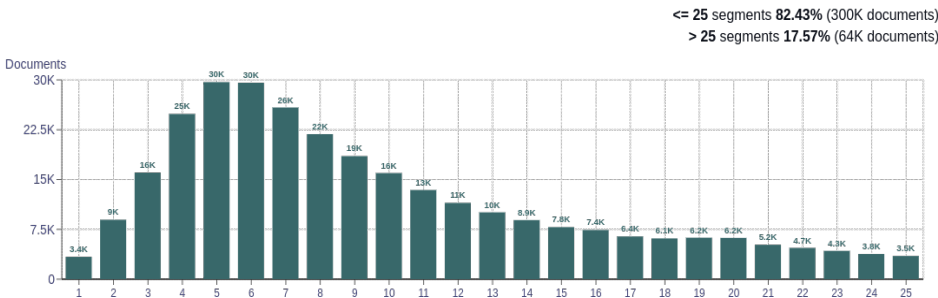
Top 10 domains

Domain	Docs	% of total
wikipedia.org	40K	11.05
dengeamerika.com	16K	4.25
ronahi.tv	13K	3.48
hk-mg.net	12K	3.27
trtnuce.com	9.1K	2.49
lotikxane.com	9K	2.46
armradio.am	8.4K	2.32
denge-welat.org	8.3K	2.29
rojevakurd.com	7.5K	2.05
sputniknews.com	5.9K	1.61

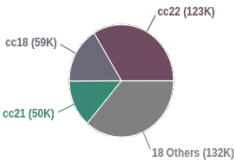
Top 10 TLDs

Domain	Docs	% of total
com	186K	51.02
org	76K	20.74
net	43K	11.72
tv	16K	4.48
am	8.5K	2.32
com.tr	6.6K	1.81
info	5.8K	1.60
se	2.3K	0.62
ir	1.9K	0.53
de	1.8K	0.49

Documents size (in segments)

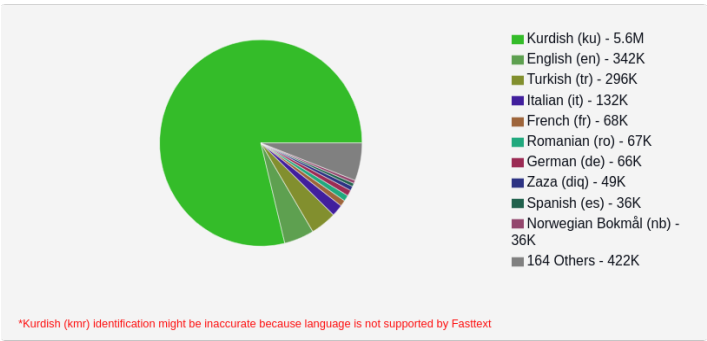


Documents by collection

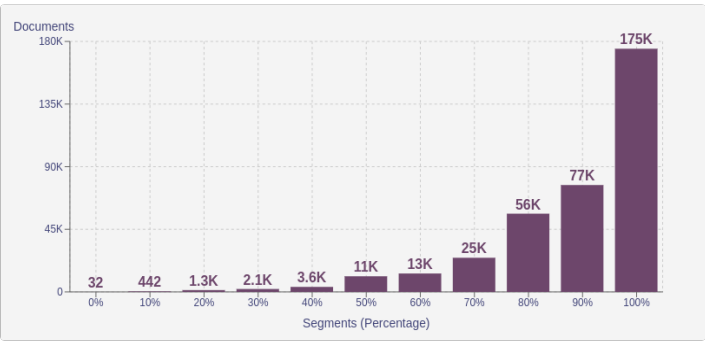


Language Distribution

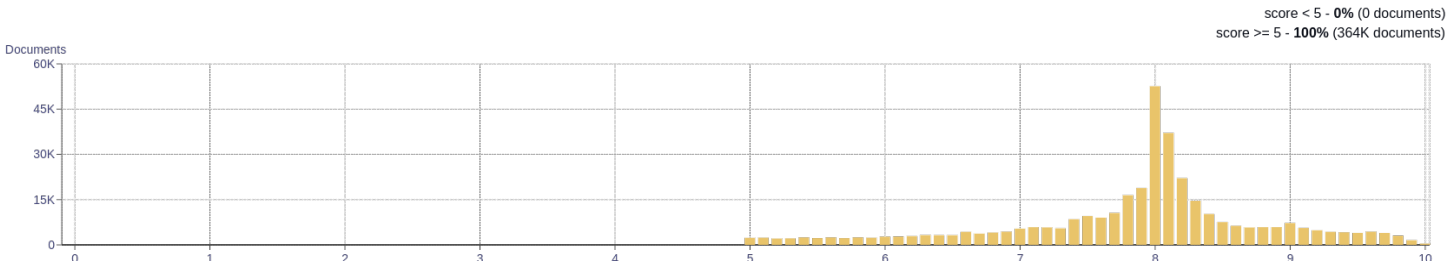
Number of segments



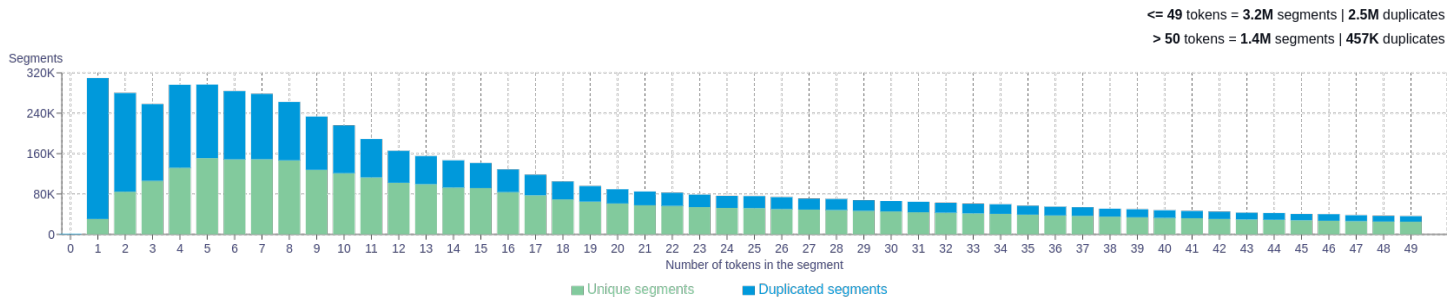
Percentage of segments in Kurdish (kmr) inside documents



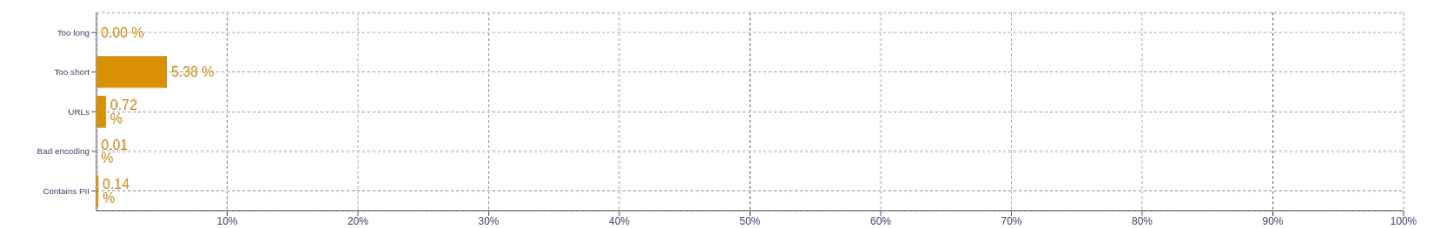
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>kurd 573759</div> <div>kirin 563589</div> <div>dike 508695</div> <div>bikin 483100</div> <div>kurdistanê 482595</div>
2	<div>i i 94449</div> <div>herêma kurdistanê 77450</div> <div>dewleta tirk 74919</div> <div>zimanê kurdî 60463</div> <div>başûrê kurdistanê 38284</div>
3	<div>i i i 93943</div> <div>tirk a dagirker 14187</div> <div>dest pê dike 13023</div> <div>rêberê gelê kurd 11200</div> <div>şert û mercên 10564</div>
4	<div>i i i i 93487</div> <div>jiyana xwe ji dest 28423</div> <div>rêberê gelê kurd abduallah 8937</div> <div>dewleta tirk a dagirker 7872</div> <div>artêşa tirk a dagirker 5731</div>
5	<div>i i i i i 93054</div> <div>rêberê gelê kurd abduallah ocalan 8777</div> <div>jiyana xwe ji dest dan 7512</div> <div>jiyana xwe ji dest dane 4254</div> <div>hawar net servîsa nûçeyên rojevê 3758</div>

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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 of 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>, , , 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>