HPLT Analytics report

HPLTAnalytics

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

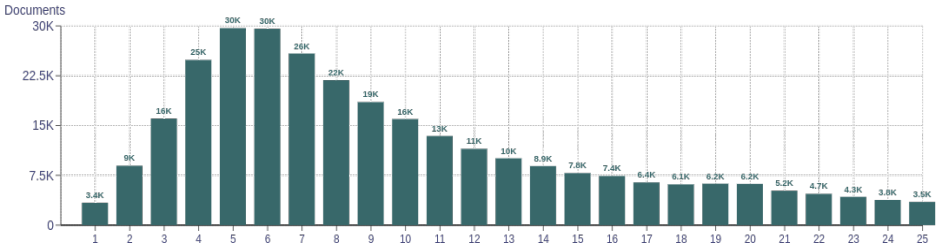
Domain	Docs	% of total
wikipedia.org	40K	11.05
dengeamerika.com	16K	4.25
ronahi.tv	13K	3.48
hk-mg.net	12K	3.27
trtnuce.com	9.1K	2.49
lotikxane.com	9K	2.46
armradio.am	8.4K	2.32
denge-welat.org	8.3K	2.29
rojevakurd.com	7.5K	2.05
sputniknews.com	5.9K	1.61

Top 10 TLDs

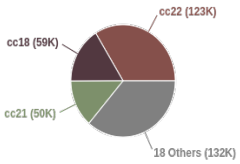
Domain	Docs	% of total
com	186K	51.02
org	76K	20.74
net	43K	11.72
tv	16K	4.48
am	8.5K	2.32
com.tr	6.6K	1.81
info	5.8K	1.60
se	2.3K	0.62
ir	1.9K	0.53
de	1.8K	0.49

Documents size (in segments)

<= 25 segments **82.43%** (300K documents)
> 25 segments **17.57%** (64K documents)

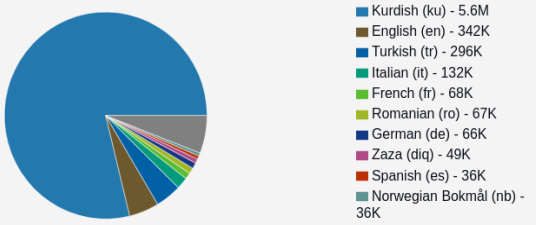


Documents by collection



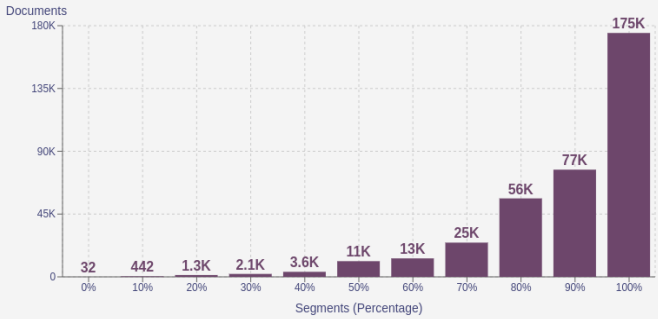
Language Distribution

Number of segments



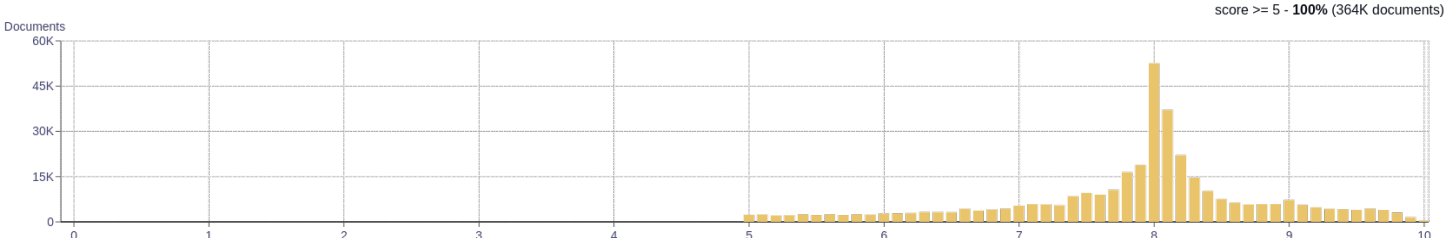
*Kurdish (kmr) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Kurdish (kmr) inside documents



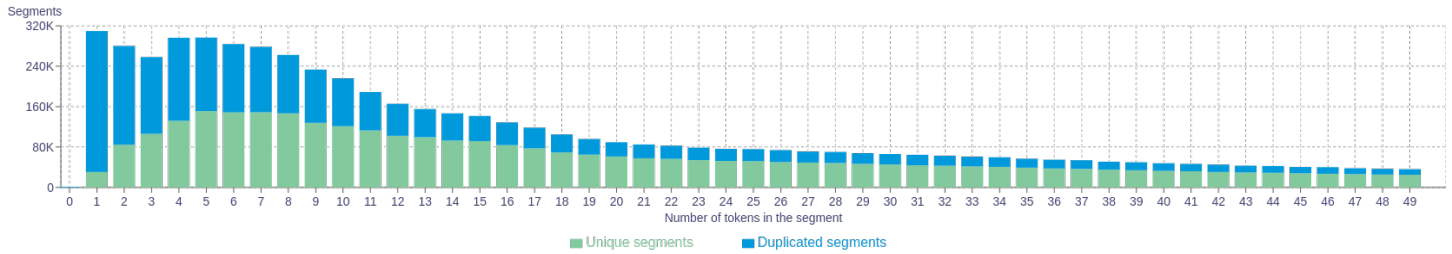
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (364K documents)

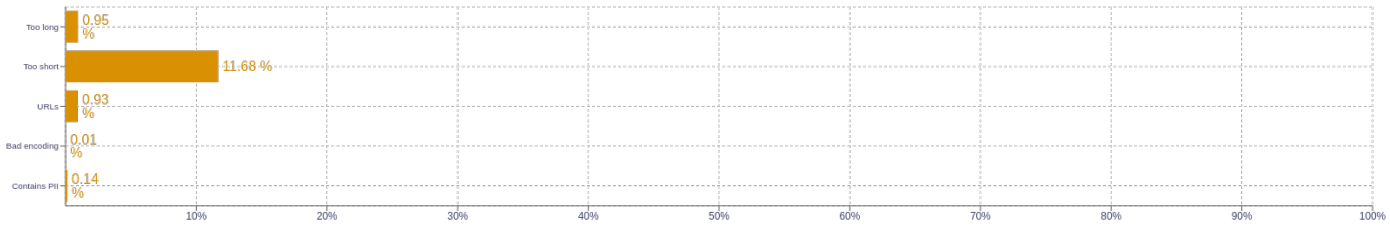


Segment length distribution by token

<= 49 tokens = **3.2M** segments | **2.5M** duplicates
> 50 tokens = **1.4M** segments | **457K** duplicates



Segment noise distribution



Frequent n-grams

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

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