HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
knc_Arab.jsonl.tsv	12/13/2024	Kanuri (knc)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
245	10,828	7,799 (72.03 %)	872K	2.22 MB	1,291,532

Top 10 domains

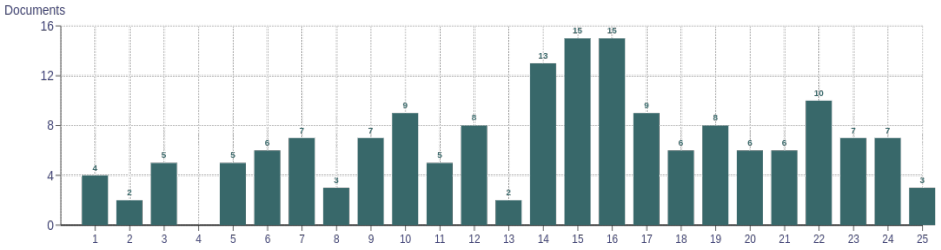
Domain	Docs	% of total
ebible.org	49	20.00
girlscoool.com	18	7.35
breakeveryyoke.com	14	5.71
blogfa.com	12	4.90
scribd.com	10	4.08
fiisci-saafi.com	8	3.27
7olm.org	7	2.86
mam9.com	6	2.45
hackps.com	5	2.04
blogspot.com	5	2.04

Top 10 TLDs

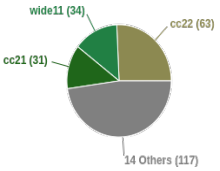
Domain	Docs	% of total
com	141	57.55
org	68	27.76
net	15	6.12
ir	7	2.86
ae	3	1.22
id	2	0.82
info	1	0.41
biz	1	0.41
town	1	0.41
ru	1	0.41

Documents size (in segments)

<= 25 segments **68.57%** (168 documents)
> 25 segments **31.43%** (77 documents)

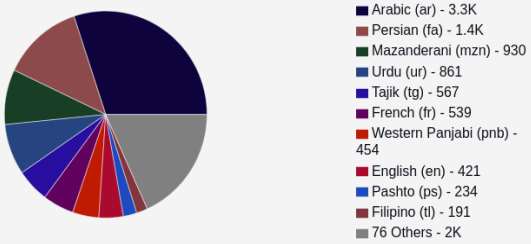


Documents by collection



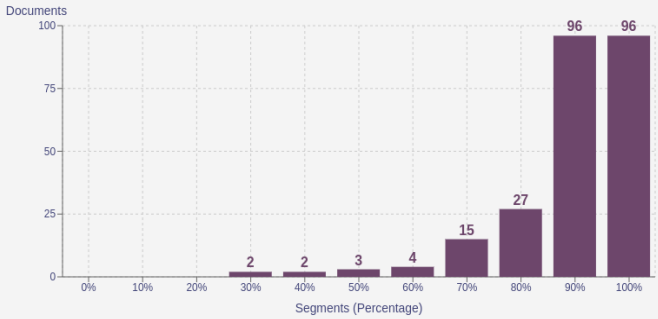
Language Distribution

Number of segments



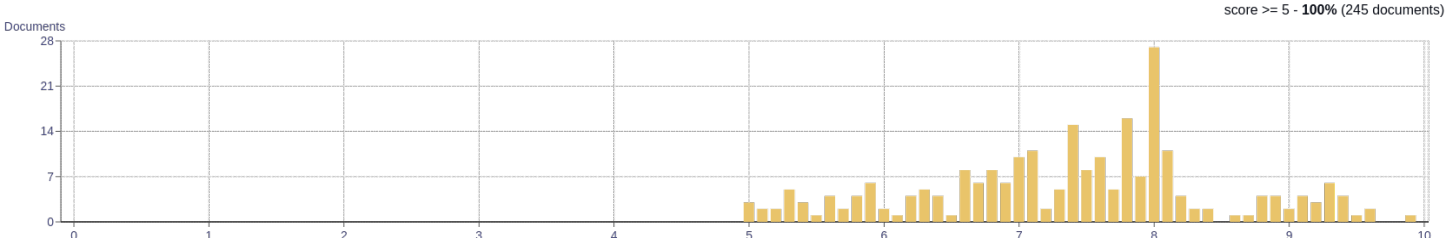
*Kanuri (knc) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Kanuri (knc) inside documents



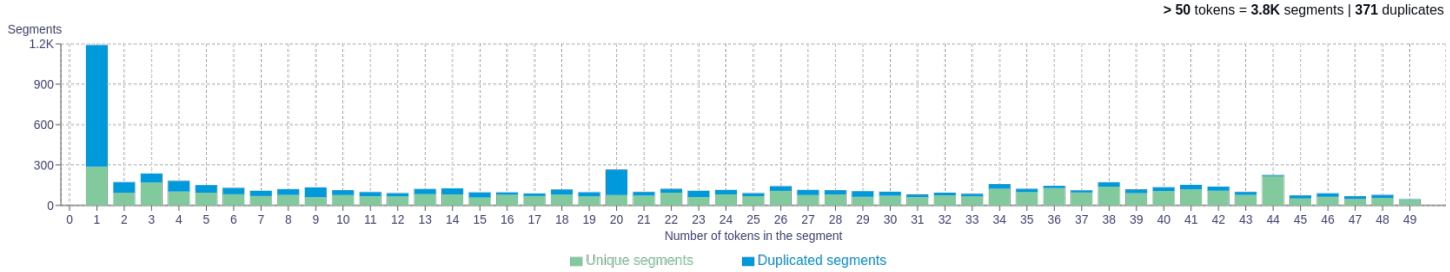
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (245 documents)

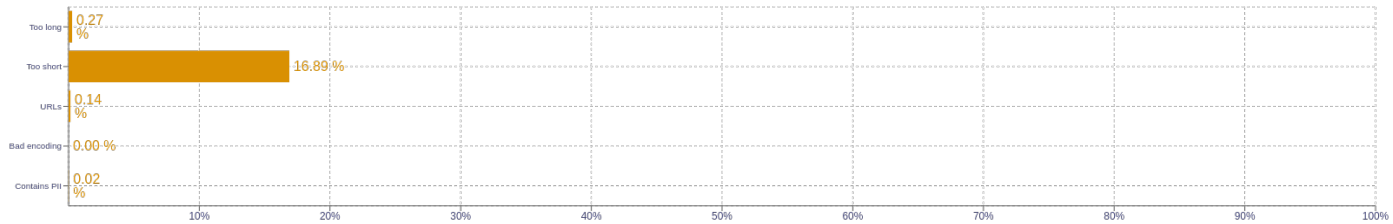


Segment length distribution by token

<= 49 tokens = **4.4K** segments | **2.7K** duplicates
> 50 tokens = **3.8K** segments | **371** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	ي | 17331 ه | 7551 غ | 7495 ى | 4888 ق | 4321
2	ت خ | 148 ن ب | 143 ّ ت | 131 ي ق | 88 ى ث | 87
3	جامع المتون للحفظ | 18 ج ن ث | 10 ب ط خ | 10 ـ ﷺ | 10 ﻫ ْ م | 9
4	ـ ﷺ ـ ْمي | 5 ونبضي حتى في وقت | 5 وقت عذابي مع قلبي | 5 ث ط ر ث | 4 لست شاعر ولست للحزن | 4
5	ونبضي حتى في وقت عذابي | 5 مشاعري ونبضي حتى في وقت | 5 غرفتي هي إحدى غرف المشفى | 4 غرف المشفى الذي كنت به | 4 أنا لست شاعر ولست للحزن | 4

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