HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
bho_Deva.jsonl.tsv	10/3/2024	Bhojpuri (bho)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
28,643	458,258	223,732 (48.82 %)	15M	68,219,632	163.08 MB

Top 10 domains

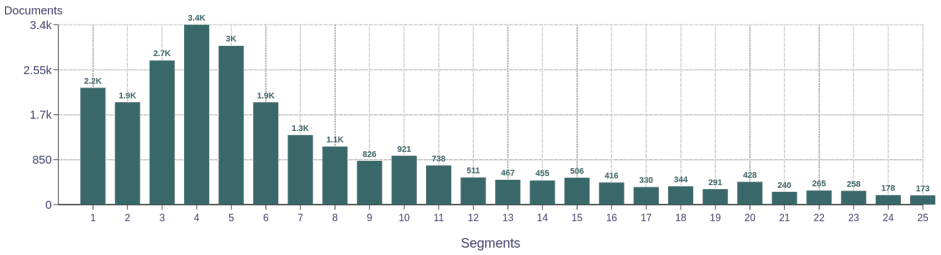
Domain	Docs	% of total
anjoria.com	12K	41.35
wikipedia.org	3.3K	11.55
bhojpurisamay.com	1K	3.64
bhojpuriaa.com	978	3.41
khabarbhojpuri.com	917	3.20
jogira.com	721	2.52
maina.co.in	564	1.97
blogspot.com	362	1.26
humbhojpuria.com	302	1.05
bhojpuria.com	301	1.05

Top 10 TLDs

Domain	Docs	% of total
com	21K	74.22
org	3.8K	13.14
in	820	2.86
co.in	714	2.49
net	691	2.41
space	279	0.97
top	258	0.90
xyz	223	0.78
cyou	216	0.75
page	142	0.50

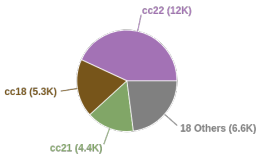
Documents size (in segments)

<= 25 segments **87.11%** (25K documents)
> 25 segments **12.89%** (3.7K documents)



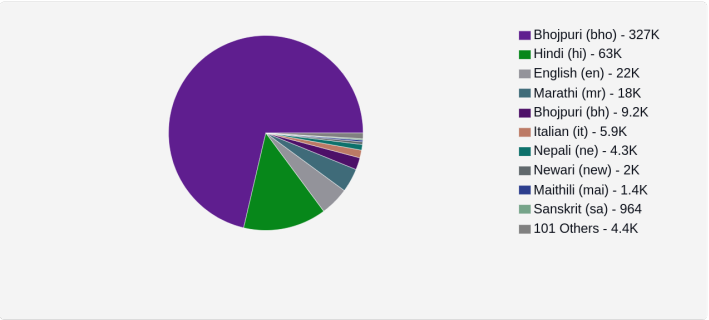
Documents by collection

CC = 86.25%
IA = 13.75%

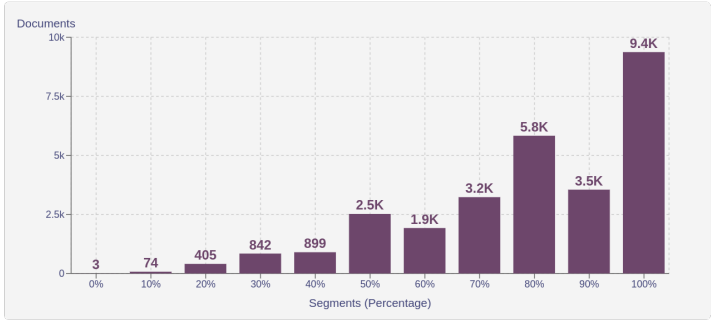


Language Distribution

Number of segments in the Bhojpuri (bho) corpus

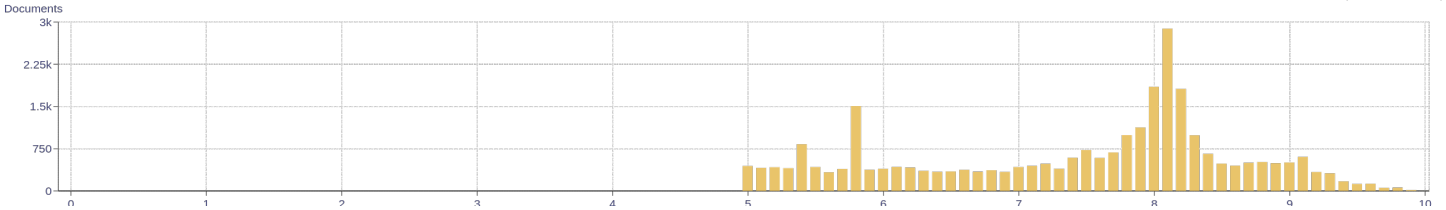


Percentage of segments in Bhojpuri (bho) inside documents



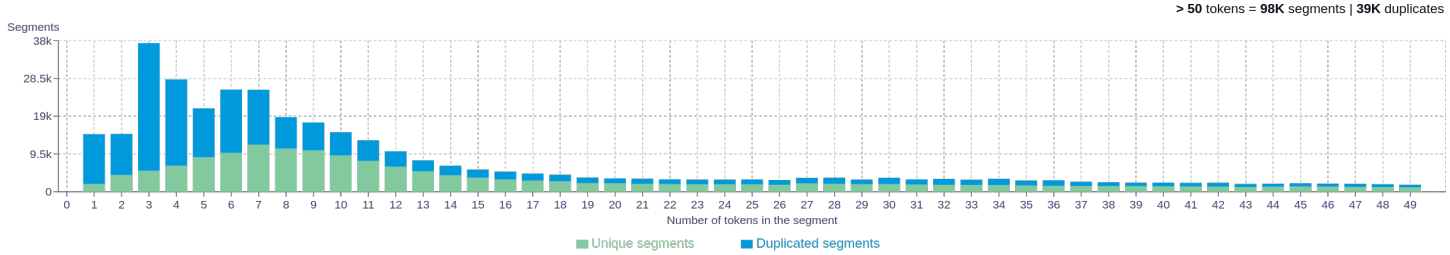
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (29K documents)

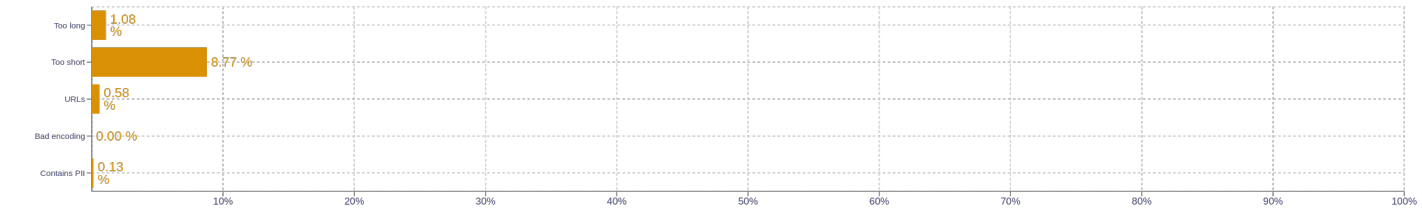


Segment length distribution by token

<= 49 tokens = **164K** segments | **196K** duplicates
> 50 tokens = **98K** segments | **39K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	भोजपुरी | 56378 · सिंह | 44921 · फिल्म | 39167 · सेक्सी | 37002 · हिंदी | 26246
2	सेक्सी मूवी | 8812 · सेक्सी फिल्म | 8121 · लोक कवि | 5230 · सेक्सी वीडियो | 4318 · भोजपुरी फिल्म | 4106
3	सेक्सी फिल्म हिंदी | 2187 · सेक्सी मूवी हिंदी | 1991 · हिंदी में फुल | 1973 · सेक्सी फिल्म फुल | 1889 · हिंदी में सेक्सी | 1706
4	सेक्सी फिल्म फुल एचडी | 1228 · हिंदी में फुल सेक्सी | 894 · सेक्सी फिल्म हिंदी फुल | 852 · फिल्म हिंदी फुल एचडी | 814 · हिंदी में फुल सेक्स | 562
5	सेक्सी फिल्म हिंदी फुल एचडी | 813 · हिंदी में फुल सेक्सी फिल्म | 667 · हिंदी में फुल सेक्स मूवी | 523 · बनारसीदास चतुर्वेदी के नाम पत्र | 423 · सेक्सी फिल्म हिंदी में फुल | 392

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