

General overview

Corpus	Analytics date	Language
mag_Deva.jsonl.tsv	12/5/2024	Magahi (mag)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
328	19,288	10,798 (55.98 %)	1M	10.07 MB	4,263,812

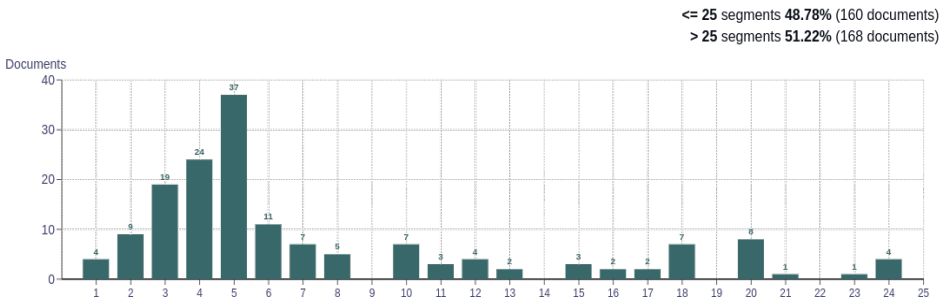
Top 10 domains

Domain	Docs	% of total
blogspot.com	109	33.23
khabarlahariya.org	93	28.35
blogspot.in	49	14.94
jagranjunction.com	29	8.84
angika.com	11	3.35
unicode.org	6	1.83
gotquestions.org	6	1.83
cutway.net	6	1.83
astrotap.net	5	1.52
wikimedia.org	2	0.61

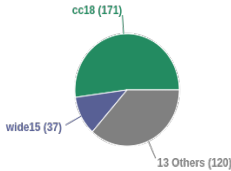
Top 10 TLDs

Domain	Docs	% of total
com	157	47.87
org	110	33.54
in	50	15.24
net	11	3.35

Documents size (in segments)

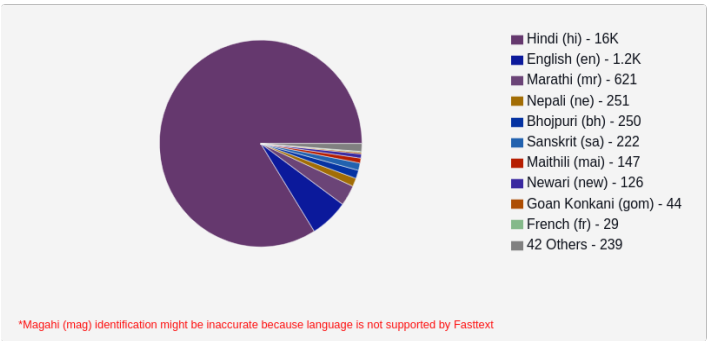


Documents by collection

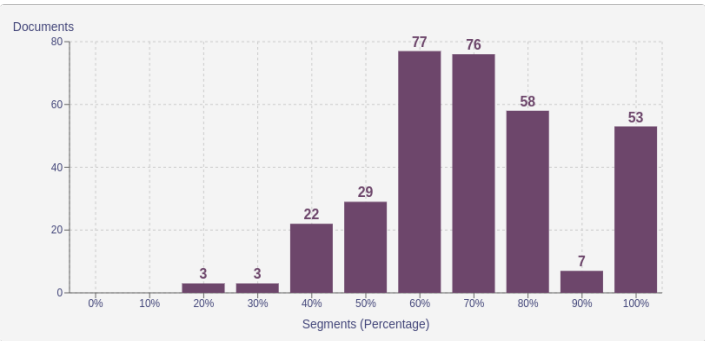


Language Distribution

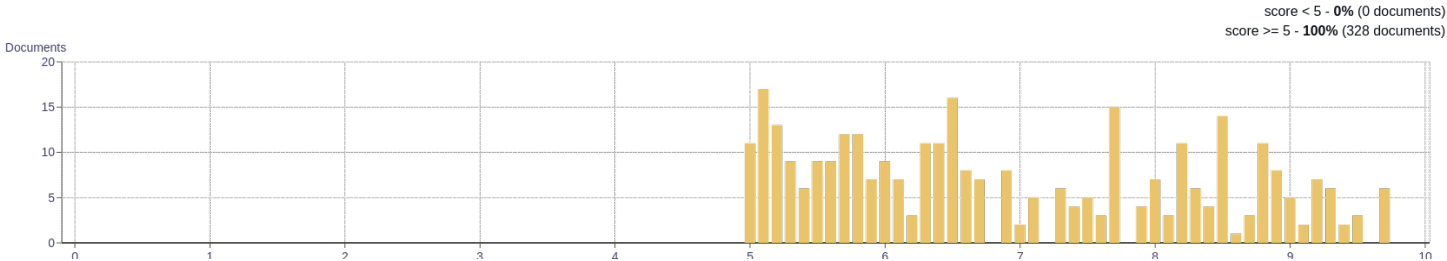
Number of segments



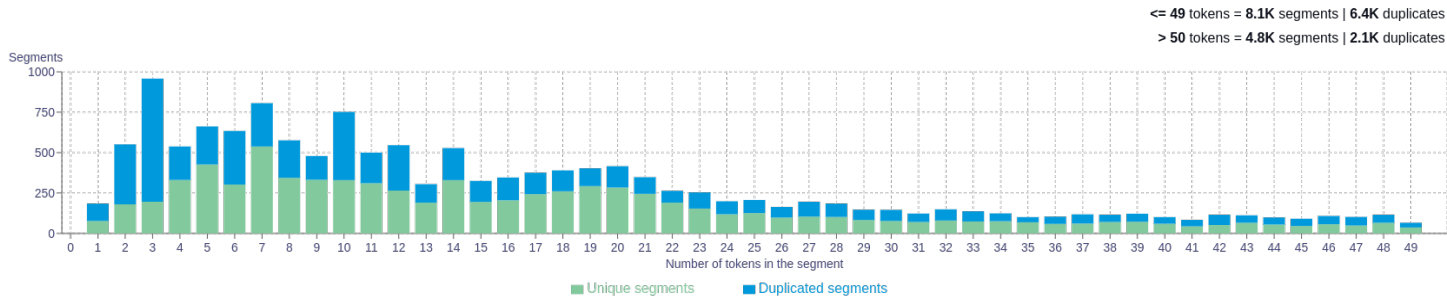
Percentage of segments in Magahi (mag) inside documents



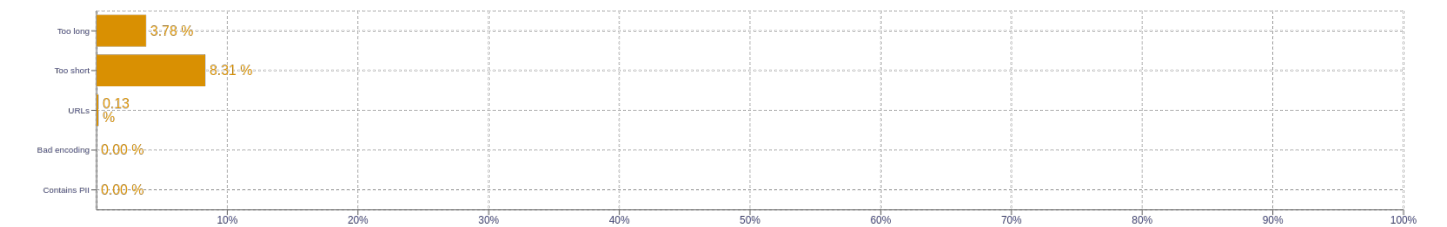
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	आउ 23504 में 19572 हइ 15022 ऊ 11847 नर्व 11445
2	रहले हल 2185 जा हइ 1461 कि ऊ 1220 ई बात 1163 करऽ हलइ 982
3	जा सकऽ हइ 437 नारायण प्रसाद at 234 posted by नारायण 234 by नारायण प्रसाद 234 करे के चाही 189
4	posted by नारायण प्रसाद 234 by नारायण प्रसाद at 234 links to this post 164 कइल जा सकऽ हइ 124 सैन एंटोनियो डे ला 101
5	posted by नारायण प्रसाद at 234 सैन एंटोनियो डे ला सीएल 101 सामान्य जीव विज्ञान प्रश्न उत्तर 57 दूरी सैन एंटोनियो डे ला 50 जीव विज्ञान प्रश्न उत्तर part 49

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.stinfo.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>