HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
fij_Latn.jsonl.tsv	11/27/2024	Fijian (fj)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
8,914	178,924	117,544 (65.69 %)	8.4M	36.05 MB	37,520,766

Top 10 domains

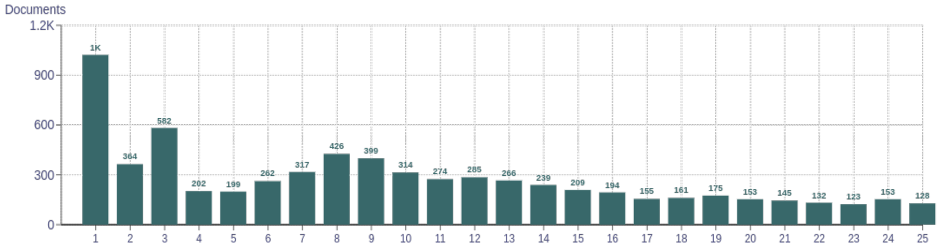
Domain	Docs	% of total
jw.org	3.2K	35.83
fijitimes.com	773	8.67
wikipedia.org	646	7.25
bible.is	626	7.02
vitifm.com.fj	500	5.61
wordproject.org	308	3.46
fijilive.com	290	3.25
fijianlyrics.com	222	2.49
lds.org	173	1.94
matavuvale.com	154	1.73

Top 10 TLDs

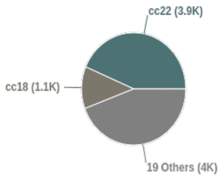
Domain	Docs	% of total
org	4.8K	53.81
com	2.4K	27.38
com.fj	686	7.70
is	626	7.02
net	63	0.71
govt.nz	60	0.67
ru	17	0.19
fr	14	0.16
bible	14	0.16
co.nz	12	0.13

Documents size (in segments)

<= 25 segments **77.17%** (6.9K documents)
> 25 segments **22.83%** (2K documents)

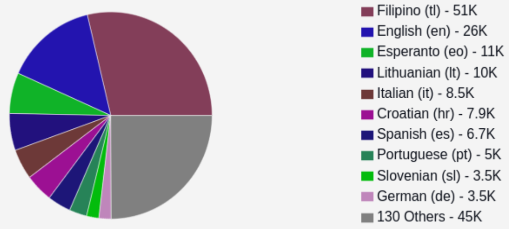


Documents by collection



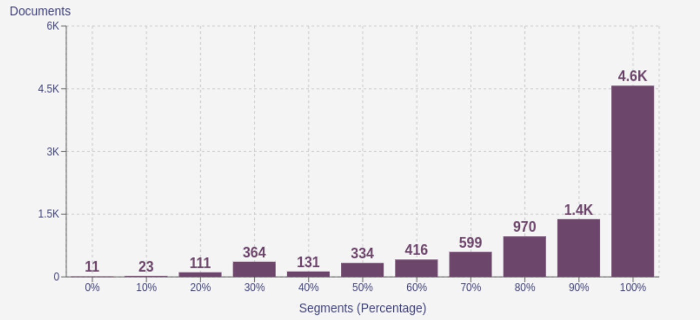
Language Distribution

Number of segments



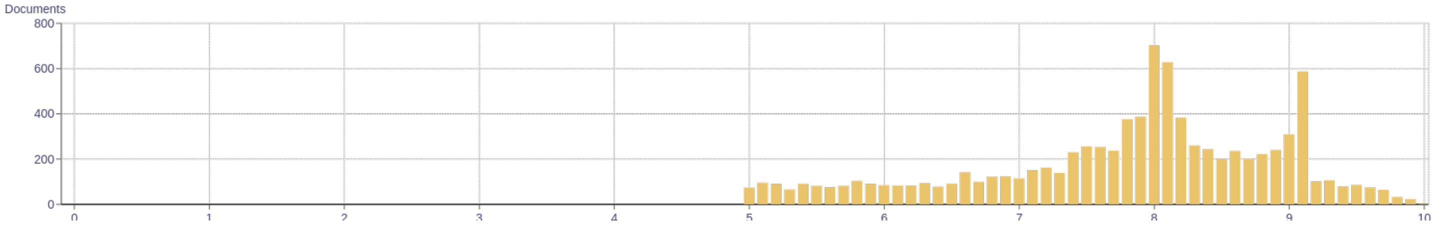
*Fijian (fj) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Fijian (fj) inside documents



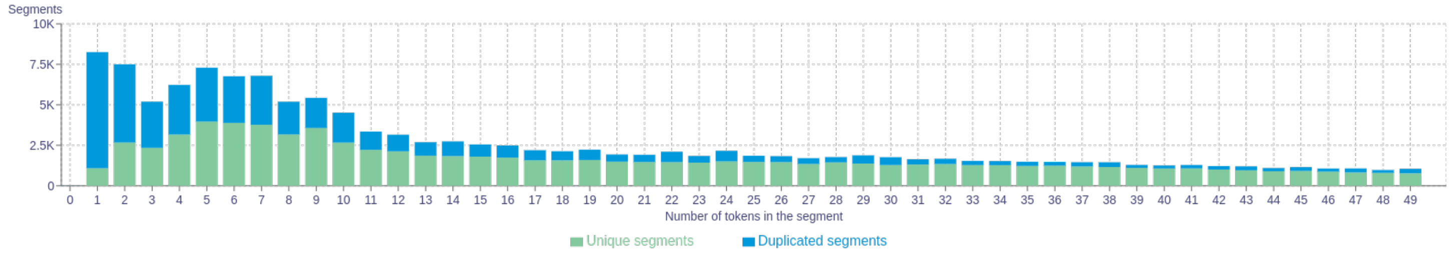
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (8.9K documents)

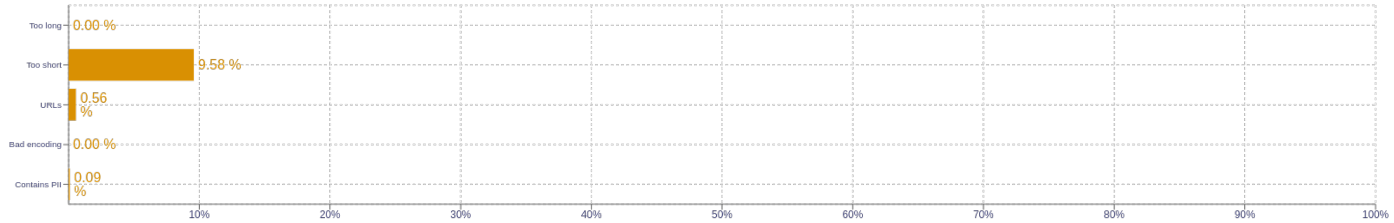


Segment length distribution by token

<= 49 tokens = **83K** segments | **50K** duplicates
> 50 tokens = **47K** segments | **12K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	ena | 145862 mai | 80863 ga | 73762 kina | 72439 ira | 68035
2	tale ga | 16686 ena gauna | 8373 i jiova | 6024 sara ga | 5826 ena dua | 5286
3	ira na tamata | 2930 kina e dua | 2049 matanitu ni kalou | 1858 mai vei ira | 1837 dua na gauna | 1830
4	cake cake cake cake | 1527 ena dua na gauna | 876 ena so na gauna | 728 mai vua na kalou | 694 ira na nona tamata | 596
5	cake cake cake cake cake | 1503 qnimate qnimate qnimate qnimate qnimate | 497 mai na vosa ni kalou | 321 local and international rugby news | 224 local and international footbal news | 224

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