HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
mri_Latn.jsonl.tsv	12/6/2024	Māori (mi)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
108,256	2,795,092	1,562,664 (55.91 %)	100M	410.55 MB	421,607,601

Top 10 domains

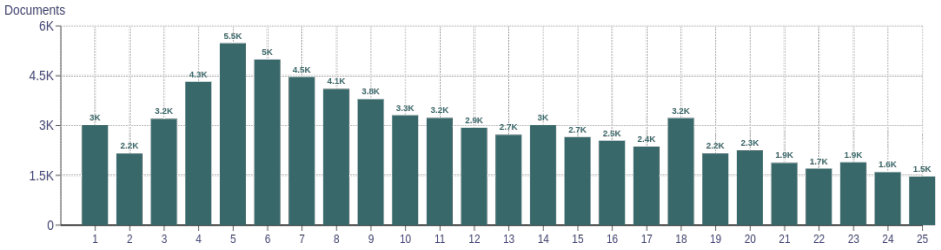
Domain	Docs	% of total
wondershare.com	5.9K	5.45
maoritelevision.com	5.8K	5.37
jw.org	3.9K	3.61
teaomaori.news	3.3K	3.06
vessoft.com	2.9K	2.65
teara.govt.nz	2.4K	2.25
bibliaonline.com.br	2K	1.89
wikipedia.org	1.9K	1.77
biblegateway.com	1.9K	1.76
vsaduidoma.com	1.8K	1.68

Top 10 TLDs

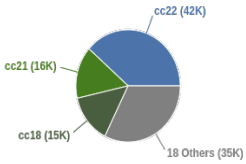
Domain	Docs	% of total
com	64K	58.88
org	13K	12.33
news	5.3K	4.91
govt.nz	5.2K	4.76
org.nz	2.7K	2.48
com.br	2.6K	2.42
co.nz	2.1K	1.98
net	1.7K	1.61
ac.nz	1.5K	1.34
zone	1.1K	1.05

Documents size (in segments)

<= 25 segments **68.78%** (74K documents)
> 25 segments **31.22%** (34K documents)

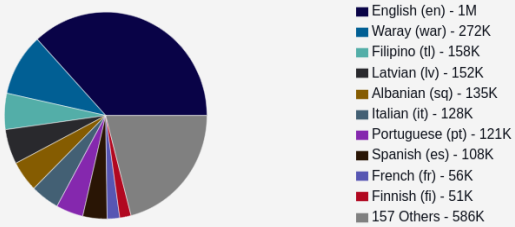


Documents by collection



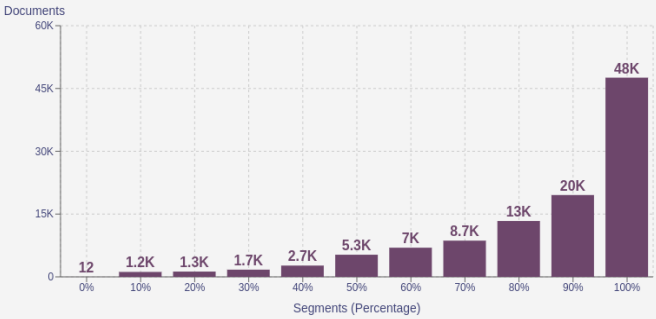
Language Distribution

Number of segments



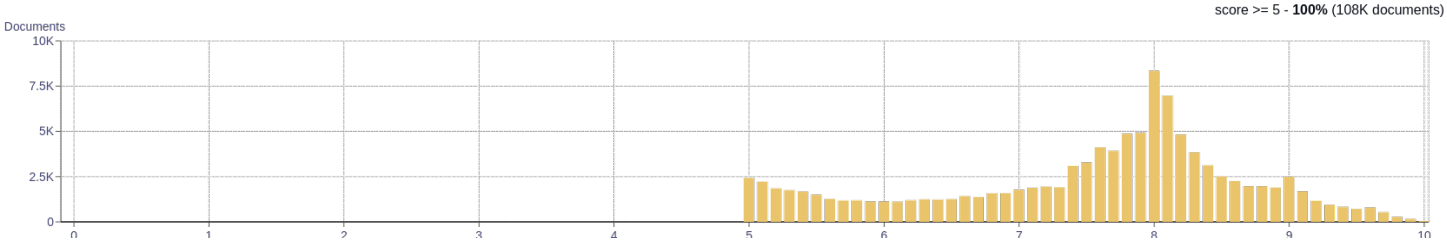
*Māori (mi) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Māori (mi) inside documents



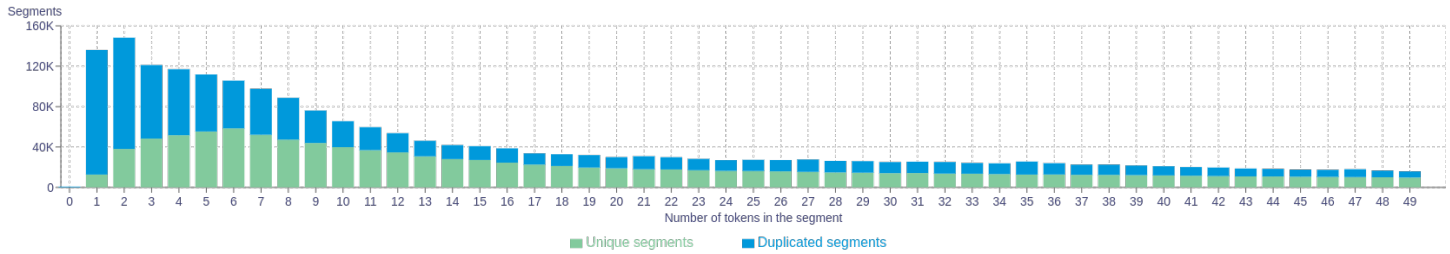
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (108K documents)

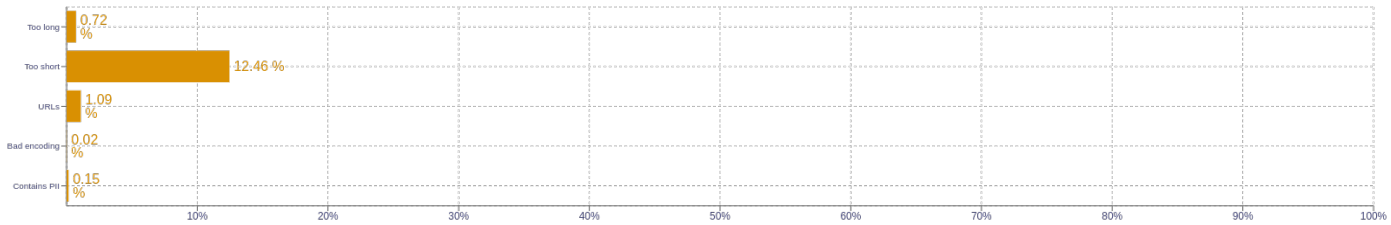


Segment length distribution by token

<= 49 tokens = **1.1M** segments | **1.1M** duplicates
> 50 tokens = **641K** segments | **177K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	te | 8145406 · o | 2387942 · ki | 2297521 · me | 1405939 · nga | 1371395
2	ki te | 1230834 · o te | 855505 · me te | 627626 · o nga | 203162 · mo te | 186438
3	roto i te | 271499 · runga i te | 104961 · roto i nga | 39979 · ki a koutou | 30861 · te nuinga o | 28606
4	taea e koe te | 52702 · te whakamahi i te | 22339 · neke atu i te | 13933 · hiahia ana koe ki | 13924 · pā ana ki te | 13033
5	hiahia ana koe ki te | 13524 · te hiahia koe ki te | 7443 · te pūmanawa e āhei ki | 7057 · ki te hiahia koe ki | 5348 · taea e koe te whakamahi | 4665

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