

General overview

Corpus	Analytics date	Language
isl_Latn.jsonl.tsv	9/19/2024	Icelandic (is)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
2,840,735	69,643,257	28,868,018 (41.45 %)	1.7B	9.8 GB	9,526,444,446

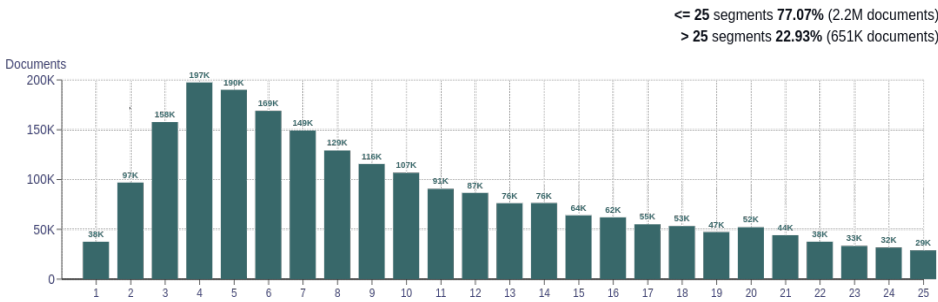
Top 10 domains

Domain	Docs	% of total
hotels.com	131K	4.62
wikipedia.org	118K	4.14
visir.is	71K	2.50
mbli.is	70K	2.46
blog.is	61K	2.14
blogspot.com	53K	1.87
althingi.is	32K	1.13
ruv.is	27K	0.97
dv.is	26K	0.93
skessuhom.is	26K	0.91

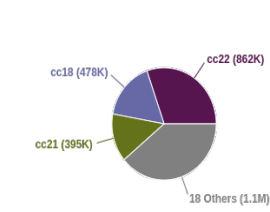
Top 10 TLDs

Domain	Docs	% of total
is	2.1M	72.84
com	478K	16.81
org	164K	5.77
net	56K	1.98
eu	9.4K	0.33
info	8.2K	0.29
dk	3.7K	0.13
no	3.4K	0.12
blog	2.9K	0.10
co.uk	2.9K	0.10

Documents size (in segments)

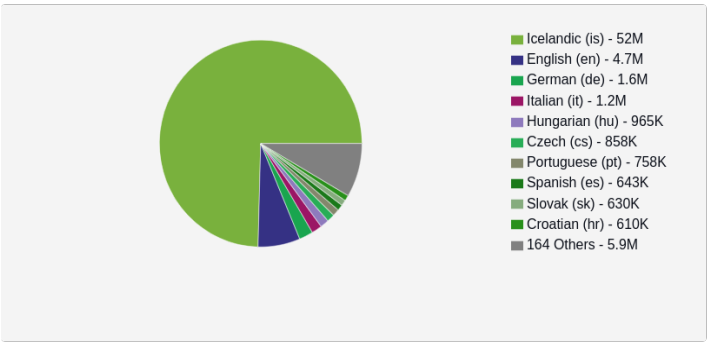


Documents by collection

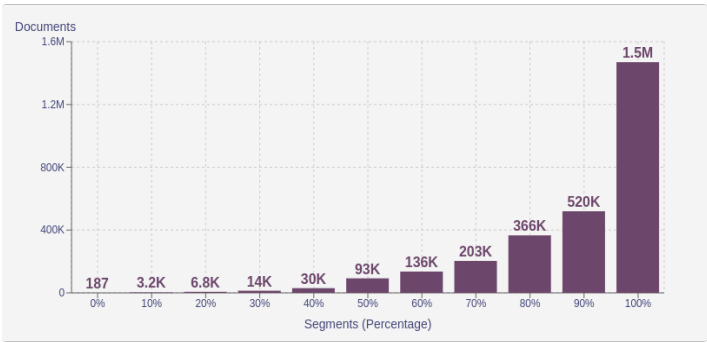


Language Distribution

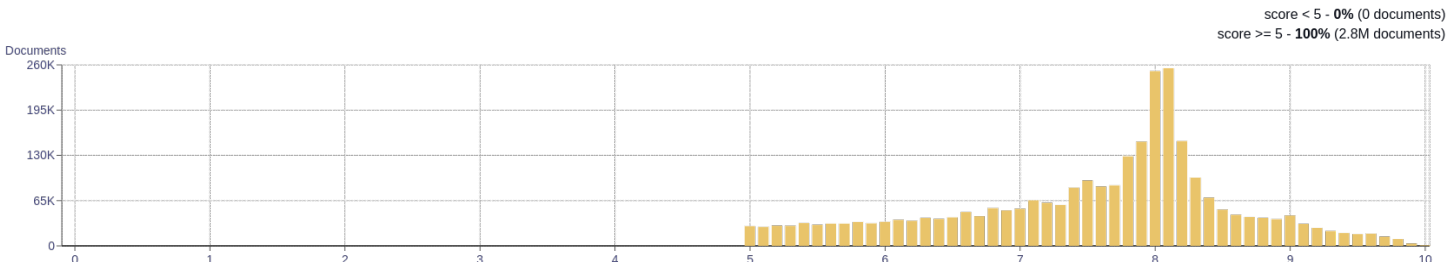
Number of segments



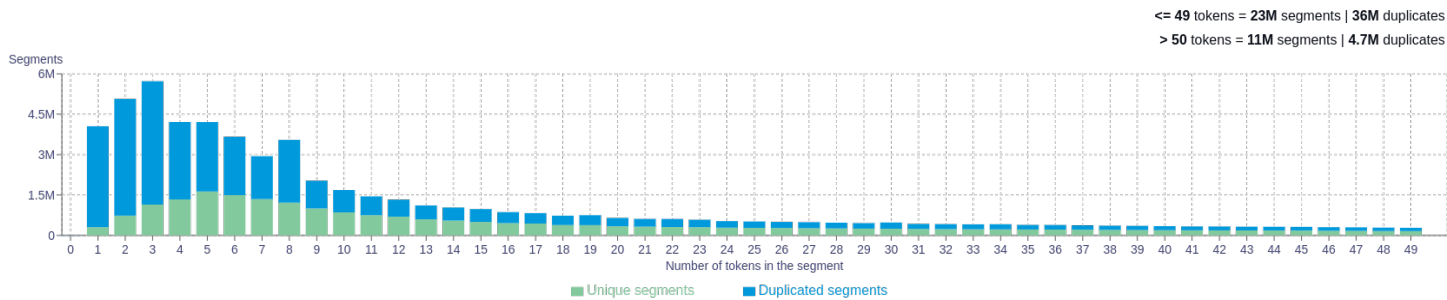
Percentage of segments in Icelandic (is) inside documents



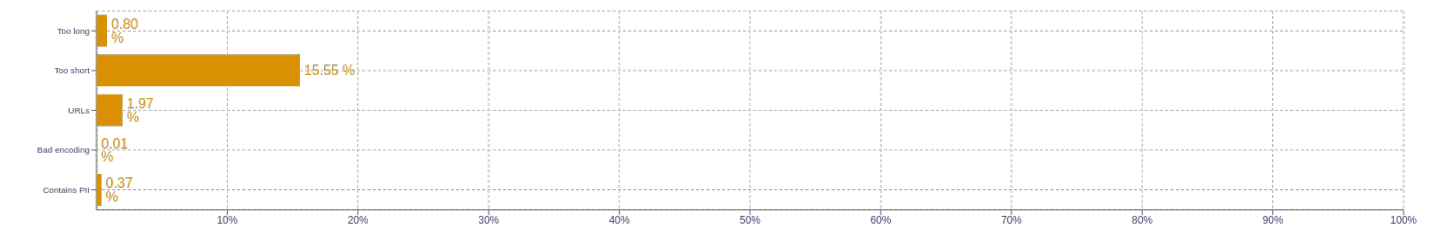
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>var   9361168</div> <div>hafa   4308729</div> <div>verið   3803958</div> <div>sé   2701146</div> <div>vera   2588016</div>
2	<div>siðustu klukkustund   691945</div> <div>einstaklingar skoðuðu   691844</div> <div>hafi verið   432232</div> <div>hafa verið   420583</div> <div>lesa meira   199970</div>
3	<div>hótel á siðustu   691845</div> <div>skoðuðu þetta hótel   691842</div> <div>hér á landi   242118</div> <div>gr. laga nr.   98705</div> <div>koma í veg   98421</div>
4	<div>hótel á siðustu klukkustund   691842</div> <div>einstaklingar skoðuðu þetta hótel   691842</div> <div>gestur hefur gefið umsögn   74119</div> <div>hótel og önnur gisting   39004</div> <div>geta gjöld verið breytileg   32854</div>
5	<div>skoðuðu þetta hótel á siðustu   691842</div> <div>gefið er upp í bókunarstaðfestingunni   35676</div> <div>upplýst okkur um eru innifalin   32763</div> <div>vegar geta gjöld verið breytileg   32761</div> <div>gjöld verið breytileg og farið   32761</div>

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 of 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>