HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
HPLT-v2-nob_Latn.tsv	9/23/2024	Norwegian Bokmål (nb)

Volumes

Docs	Segments	Size	Characters
27,053,845	675,970,248	132,594,741,063	126.24 GB

Top 10 domains

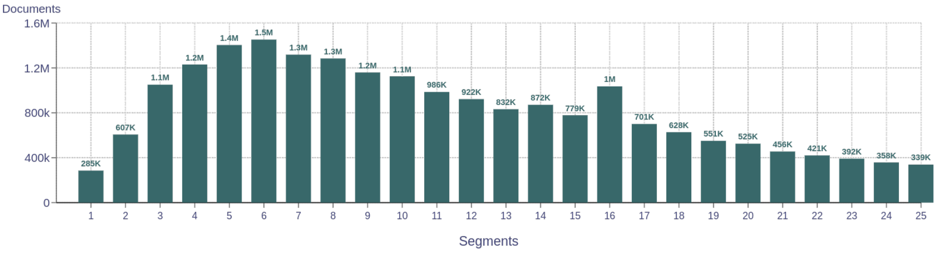
Domain	Docs	% of total
blogspot.com	1.2M	4.37
blogg.no	1.1M	3.90
blogspot.no	711K	2.63
wikipedia.org	594K	2.20
aftenposten.no	386K	1.43
docplayer.me	376K	1.39
dagbladet.no	330K	1.22
wordpress.com	267K	0.99
tripadvisor.com	261K	0.96
nrk.no	222K	0.82

Top 10 TLDs

Domain	Docs	% of total
no	17M	64.44
com	5.8M	21.34
org	997K	3.68
net	465K	1.72
me	385K	1.42
eu	385K	1.42
kommune.no	189K	0.70
info	179K	0.66
ru	94K	0.35
se	83K	0.31

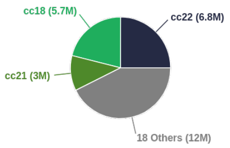
Documents size (in segments)

<= 25 segments **76.57%** (21M documents)
> 25 segments **23.43%** (6.3M documents)



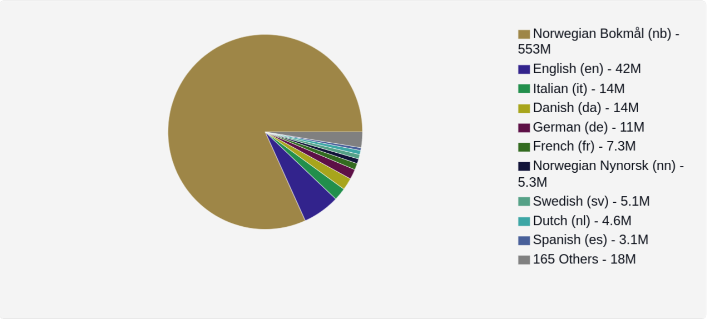
Documents by collection

CC = 69.54%
IA = 30.46%

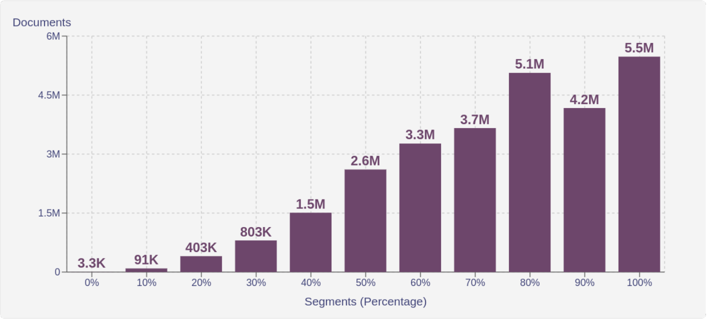


Language Distribution

Number of segments in the Norwegian Bokmål (nb) corpus

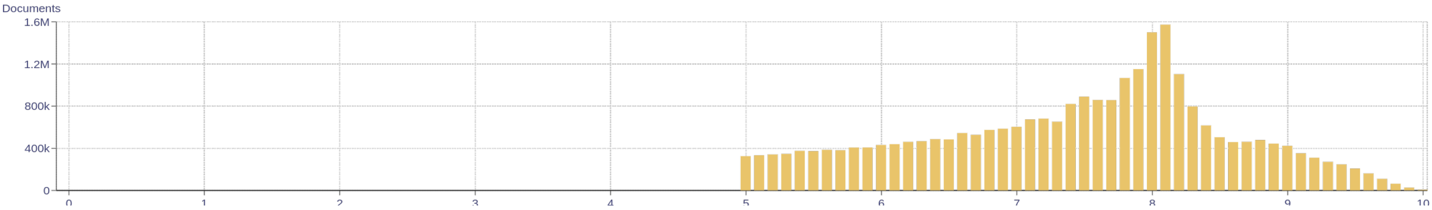


Percentage of segments in Norwegian Bokmål (nb) inside documents

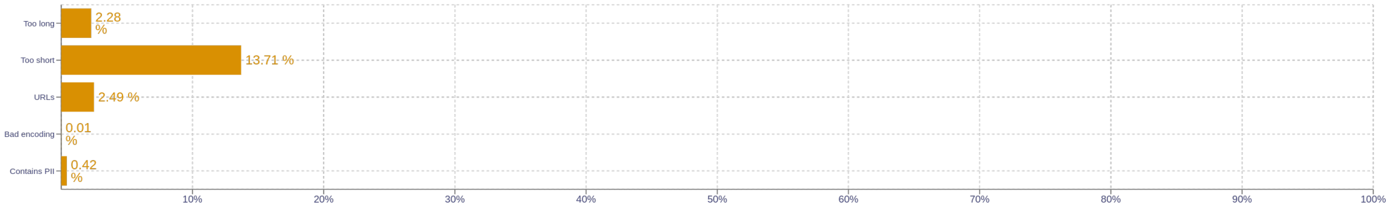


Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (27M documents)



Segment noise distribution



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