HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
nno_Latn.jsonl.tsv	9/21/2024	Norwegian Nynorsk (nn)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
1,423,143	34,603,343	12,530,321 (36.21 %)	983M	5.12 GB	5,371,043,855

Top 10 domains

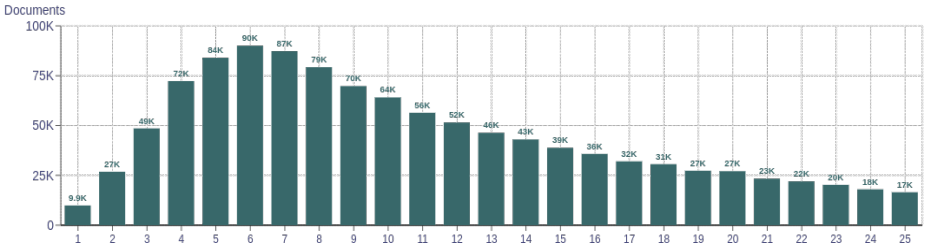
Domain	Docs	% of total
wikipedia.org	325K	22.84
blogspot.com	80K	5.60
blogspot.no	49K	3.44
nrk.no	36K	2.52
docplayer.me	31K	2.20
ndla.no	25K	1.76
blogg.no	20K	1.38
wordpress.com	19K	1.36
framtida.no	15K	1.06
allkunne.no	9.7K	0.68

Top 10 TLDs

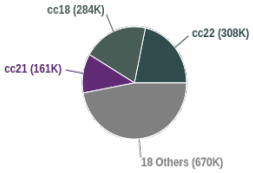
Domain	Docs	% of total
no	758K	53.29
org	345K	24.28
com	185K	13.03
kommune.no	35K	2.43
me	32K	2.21
net	19K	1.36
info	9.5K	0.67
eu	5K	0.35
vgs.no	4.3K	0.30
dk	1.9K	0.13

Documents size (in segments)

<= 25 segments **78.81%** (1.1M documents)
> 25 segments **21.19%** (302K documents)

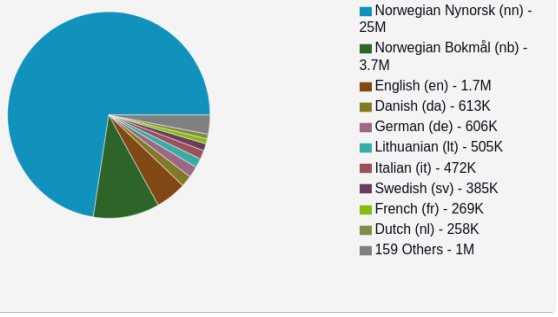


Documents by collection

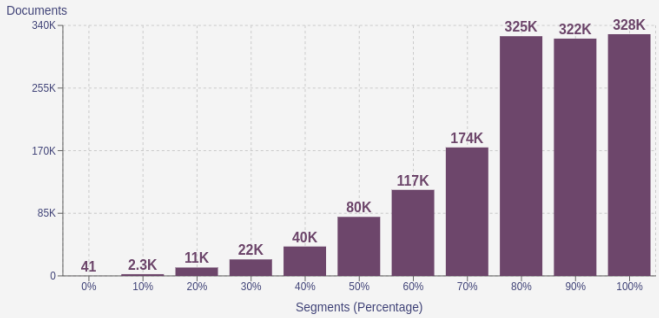


Language Distribution

Number of segments

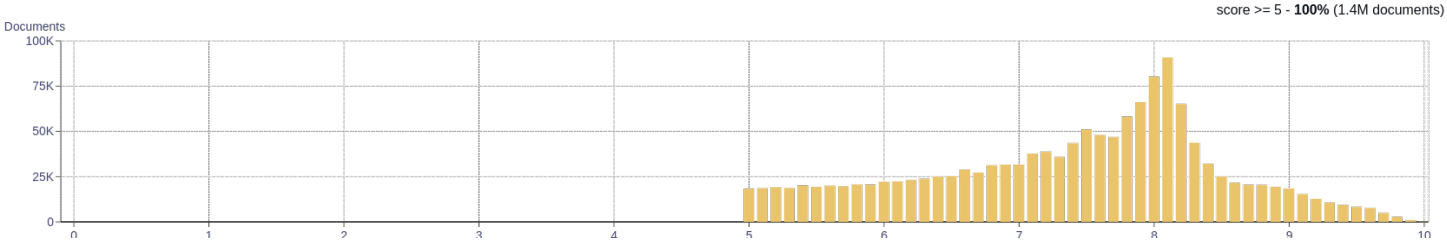


Percentage of segments in Norwegian Nynorsk (nn) inside documents



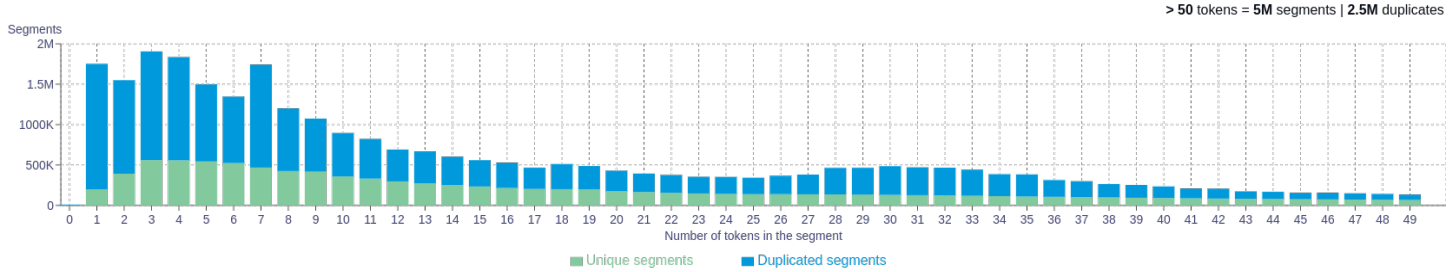
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (1.4M documents)

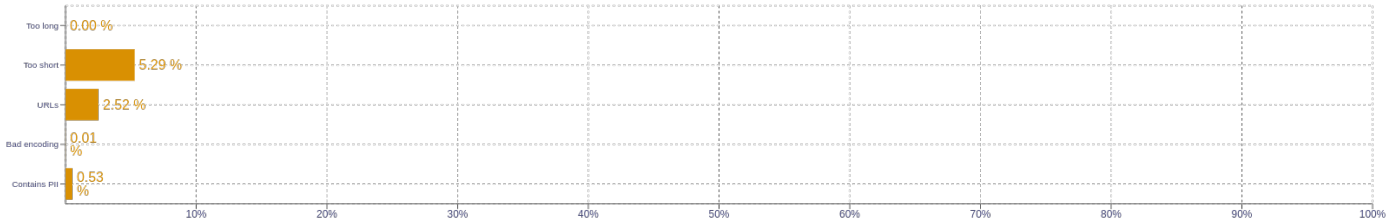


Segment length distribution by token

<= 49 tokens = **10M** segments | **20M** duplicates
> 50 tokens = **5M** segments | **2.5M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	kommune | 1946572 · endre | 1707517 · år | 1426019 · må | 1266078 · vert | 1193946
2	endre wikiteksten | 777797 · kommune møteprotokoll | 223642 · null null | 140176 · utval møtedato | 126664 · artikkelen bygger | 126297
3	sogn og fjordane | 262276 · møre og romsdal | 230336 · null null null | 139713 · wikipedia på engelsk | 108166 · navn funksjon representerer | 67805
4	null null null null | 139310 · møre og romsdal fylkeskommune | 45201 · nynorsk/bokmål nynorsk eksamensinformasjon eksamenstid | 44853 · følgjande faste medlemmer møtte | 43823 · medlemmer var til stades | 42274
5	null null null null null | 138938 · faste medlemmer var til stades | 42266 · wikipedia på engelsk oppgav desse | 32322 · innkalling til møtet vart gjort | 30583 · møtet vart gjort i samsvar | 28997

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