

General overview

Corpus	Analytics date	Language
nn_1.jsonl.tsv	3/19/2024	Norwegian Nynorsk (nn)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
228,480	28,787,245	7,014,173 (24.37 %)	350M	1.77 GB	

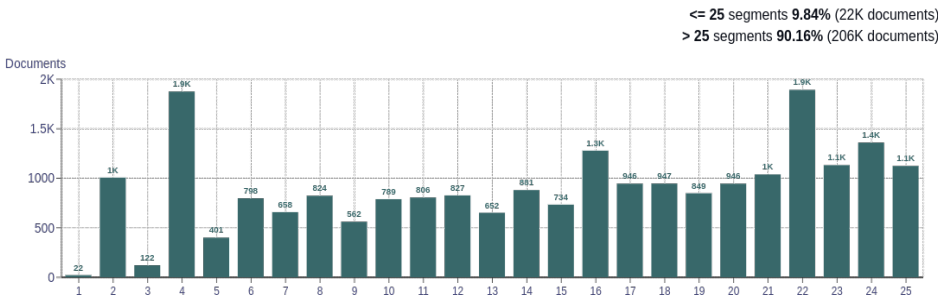
Top 10 domains

Domain	Docs	% of total
blogspot.no	26K	11.50
wikipedia.org	14K	6.15
docplayer.me	9.9K	4.31
ndla.no	9.1K	3.99
blogspot.com	8.8K	3.84
framida.no	3.7K	1.64
lokalhistoriewiki.no	2.8K	1.21
wordpress.com	2.2K	0.94
midtsiden.no	1.8K	0.77
uib.no	1.6K	0.68

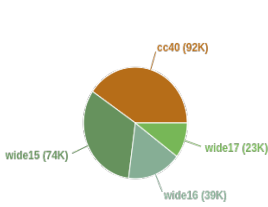
Top 10 TLDs

Domain	Docs	% of total
no	147K	64.30
com	31K	13.68
org	19K	8.45
me	9.9K	4.32
kommune.no	5.2K	2.27
net	3.5K	1.54
info	2K	0.86
vgs.no	1.1K	0.49
fr	871	0.38
se	772	0.34

Documents size (in segments)

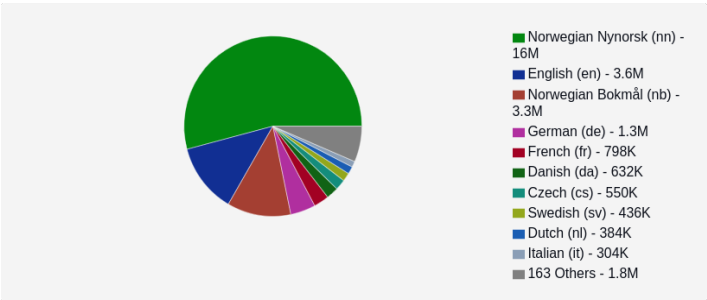


Documents by collection

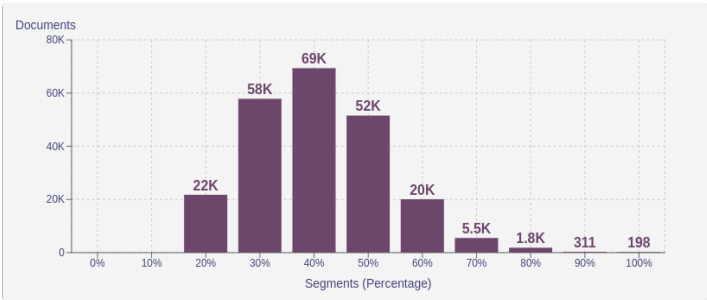


Language Distribution

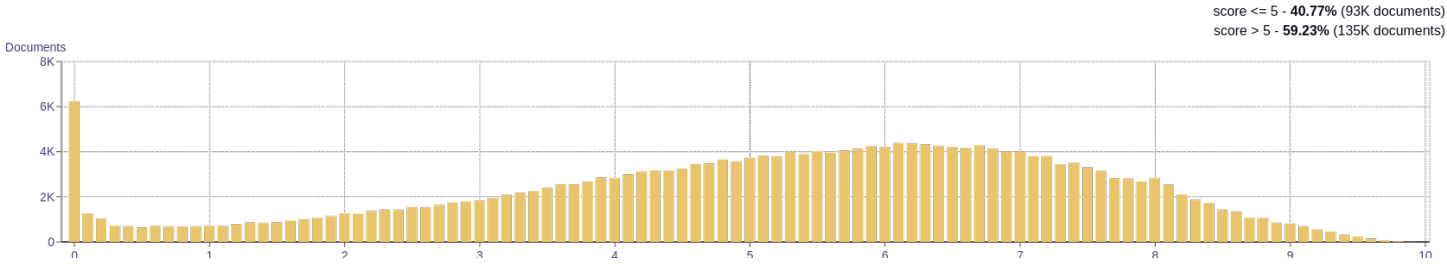
Number of segments



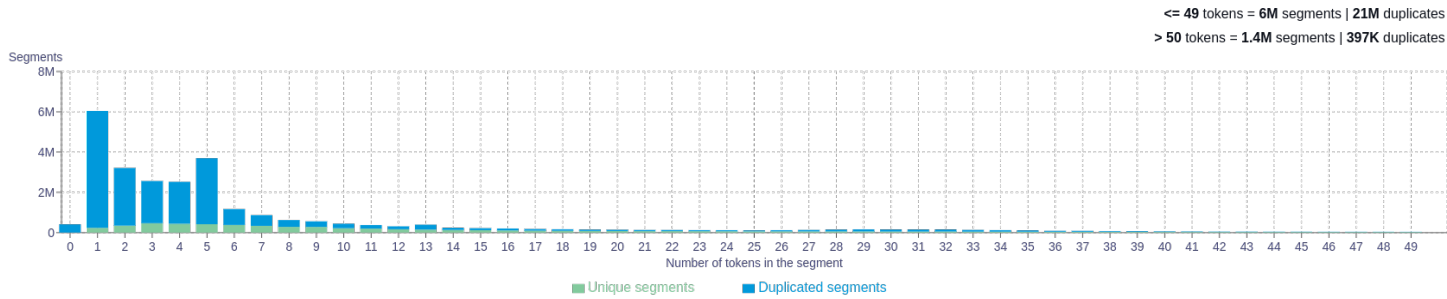
Percentage of segments in Norwegian Nynorsk (nn) inside documents



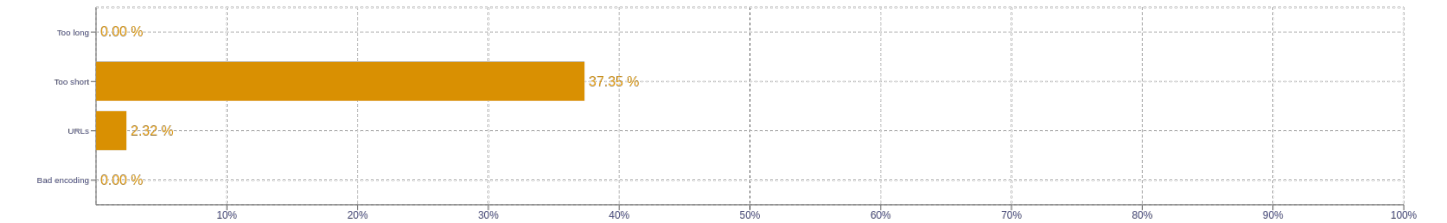
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>detaljer 895215</div> <div>kommune 733749</div> <div>år 692325</div> <div>to 514673</div> <div>the 456217</div>
2	<div>kommune møteprotokoll 110302</div> <div>via e-postblogg 91549</div> <div>funksjon representerer 63615</div> <div>møteprotokoll utval 59567</div> <div>utval møtedato 52739</div>
3	<div>send dette via 91644</div> <div>del på twitterdel 91550</div> <div>facebookdel på pinterest 91313</div> <div>twitterdel på facebookdel 91312</div> <div>sogn og fjordane 77264</div>
4	<div>send dette via e-postblogg 91549</div> <div>nyere innlegg eldre innlegg 33711</div> <div>innlegg eldre innlegg start 25492</div> <div>nynorsk/bokmål nynorsk eksamensinformasjon eksamenstid 22900</div> <div>følgjande faste medlemmer møte 21141</div>
5	<div>twitterdel på facebookdel på pinterest 91312</div> <div>del på twitterdel på facebookdel 91312</div> <div>nyere innlegg eldre innlegg start 25492</div> <div>faste medlemmer var til stades 14925</div> <div>ronk ronk ronk ronk ronk 13092</div>

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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>, , , 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>