HPLT Analytics report

General overview

Corpus	Analytics date	Language
HPLT-docslite.sv.tsv	6/14/2024	Swedish (sv)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
13,668,868	1,787,683,997	380,972,579 (21.31 %)	20B	102.36 GB	

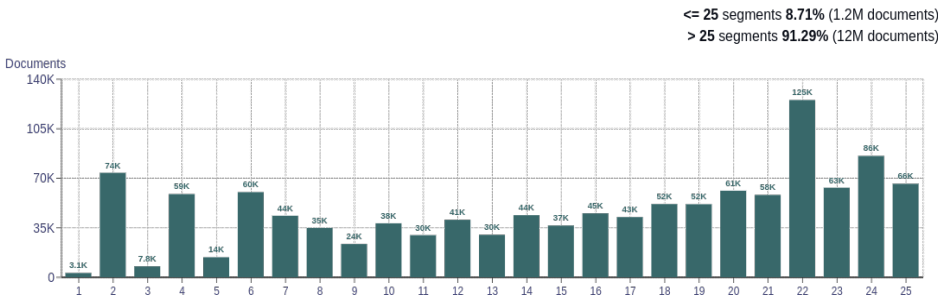## Top 10 domains

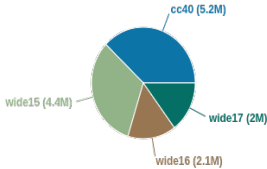| Domain	Docs	% of total
blogspot.se	1.2M	8.84
web.app	484K	3.54
blogspot.com	231K	1.69
wordpress.com	225K	1.64
blogspot.fi	200K	1.46
docplayer.se	181K	1.33
blogg.se	118K	0.87
firebaseapp.com	104K	0.76
blogspot.no	102K	0.75
abc-annons.se	83K	0.61

Top 10 TLDs

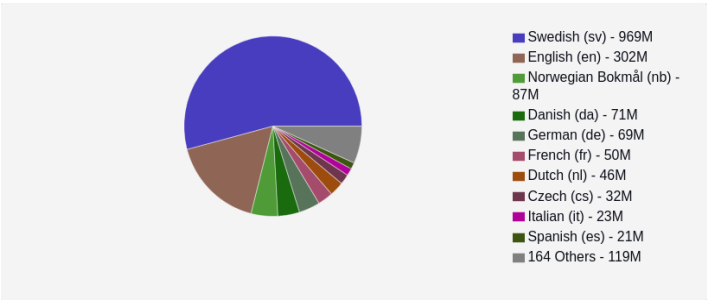| Domain	Docs	% of total
se	8.4M	61.66
com	2.6M	19.06
app	485K	3.55
nu	450K	3.29
fi	335K	2.45
org	253K	1.85
net	148K	1.08
eu	146K	1.07
no	108K	0.79
de	59K	0.43

Documents size (in segments)
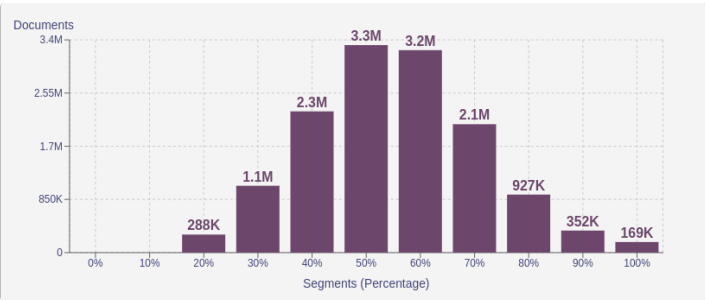
<= 25 segments **8.71%** (1.2M documents)
> 25 segments **91.29%** (12M documents)

Documents by collection

cc40 (5.2M), wide15 (4.4M), wide17 (2M), wide16 (2.1M)

Language Distribution

Number of segments

- Swedish (sv) - 969M
- English (en) - 302M
- Norwegian Bokmål (nb) - 87M
- Danish (da) - 71M
- German (de) - 69M
- French (fr) - 50M
- Dutch (nl) - 46M
- Czech (cs) - 32M
- Italian (it) - 23M
- Spanish (es) - 21M
- 164 Others - 119M

Percentage of segments in Swedish (sv) inside documents

Distribution of documents by document score

score <= 5 - **28.56%** (3.9M documents)
score > 5 - **71.44%** (9.8M documents)

Segment length distribution by token

<= 49 tokens = **322M** segments | **1.4B** duplicates
> 50 tokens = **87M** segments | **27M** duplicates

- Unique segments
- Duplicated segments
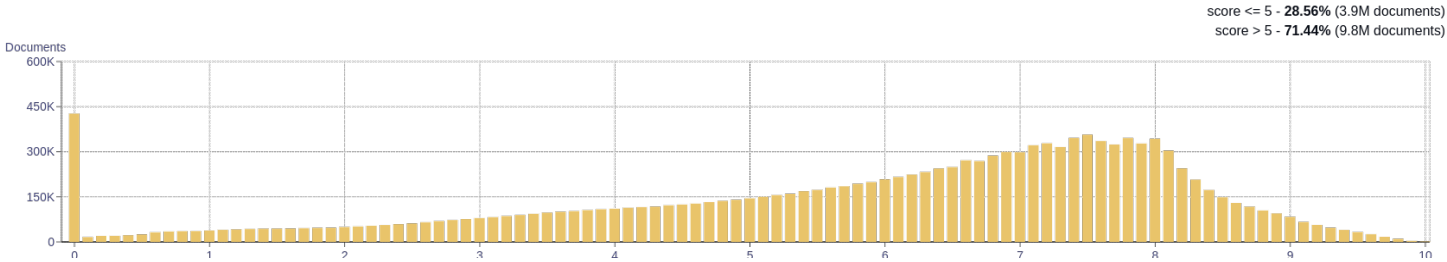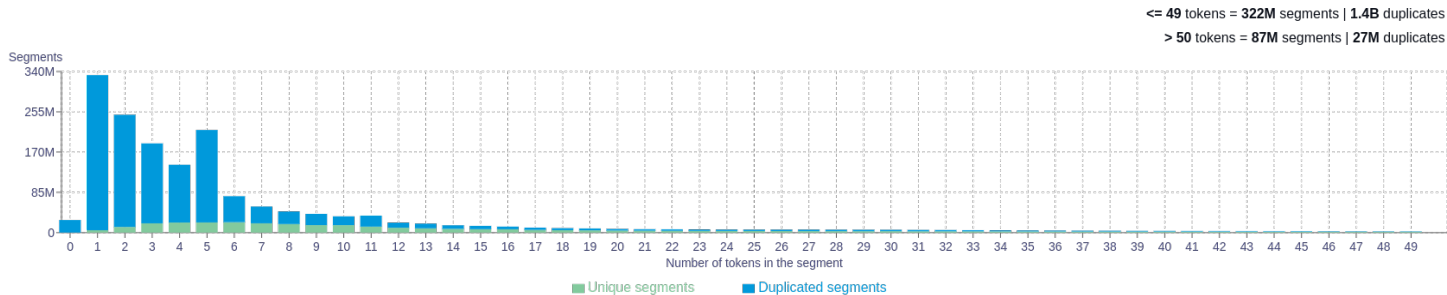
About HPLT Analytics

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