HPLT Analytics report

General overview

Corpus	Analytics date	Language
HPLT-v2-deu_Latn.tsv	10/14/2024	German (de)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
482,053,407	11,127,739,434			1.64 TB	1,771,482,213,502

Top 10 domains

Domain	Docs	% of total
blogspot.com	4.7M	0.97
wikipedia.org	4.5M	0.92
wordpress.com	4M	0.83
blogspot.de	2.1M	0.44
derstandard.at	1.3M	0.27
webwiki.de	1.1M	0.22
gutefrage.net	1M	0.21
blogspot.co.at	996K	0.21
welt.de	945K	0.20
spiegel.de	885K	0.18

Top 10 TLDs

Domain	Docs	% of total
de	290M	60.19
com	80M	16.56
at	26M	5.49
ch	23M	4.71
org	15M	3.05
net	13M	2.71
eu	6.8M	1.40
info	5.5M	1.14
biz	1.3M	0.27
co.at	1.2M	0.25

Documents size (in segments)

<= 25 segments **76.36%** (368M documents)
> 25 segments **23.64%** (114M documents)
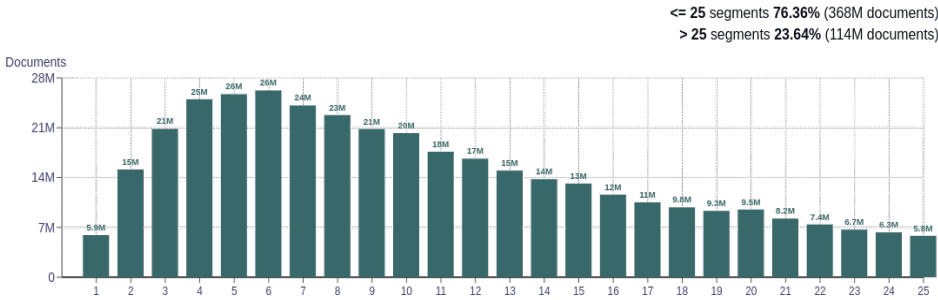
Documents by collection

cc18 (92M) · cc22 (118M) · cc21 (58M) · 18 Others (213M)
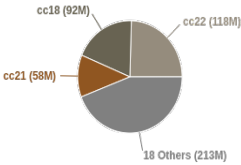
Language Distribution

Number of segments

- German (de) - 9.5B
- English (en) - 704M
- Italian (it) - 306M
- French (fr) - 136M
- Dutch (nl) - 83M
- Spanish (es) - 53M
- Swedish (sv) - 43M
- Norwegian Bokmål (nb) - 24M
- Polish (pl) - 22M
- Portuguese (pt) - 22M
- 165 Others - 239M
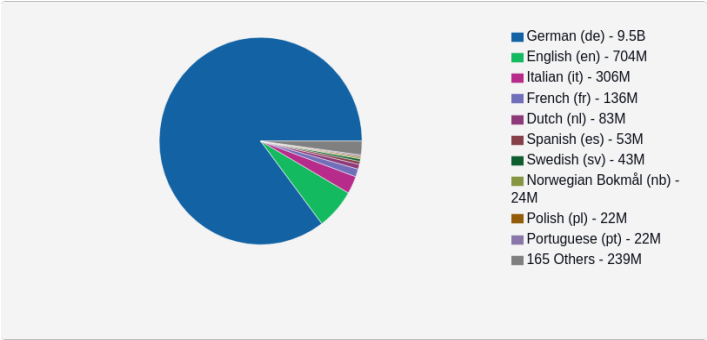
Percentage of segments in German (de) inside documents
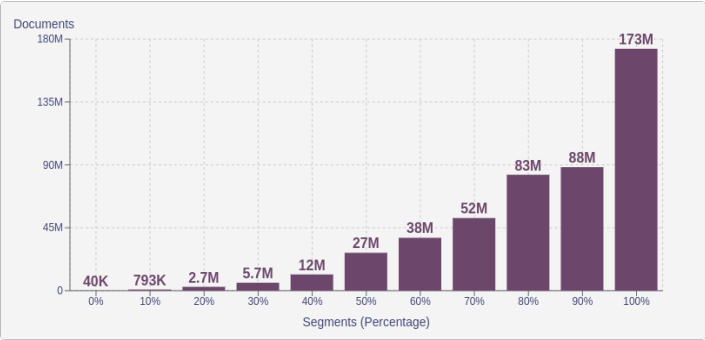
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (482M documents)

Segment noise distribution

- Too long — 1.10 %
- Too short — 18.65 %
- URLs — 2.56 %
- Bad encoding — 0.02 %
- Contains PII — 0.72 %
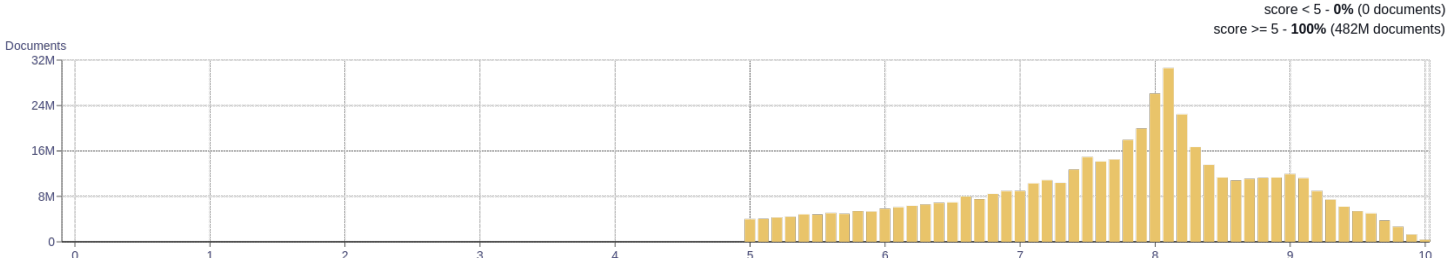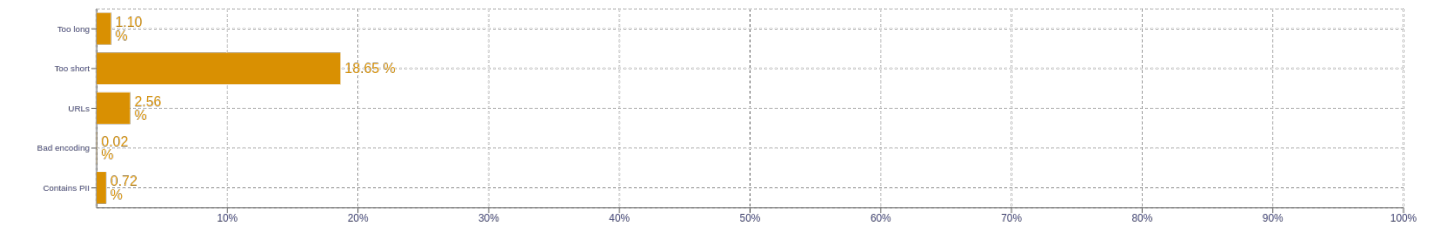
About HPLT Analytics

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