

General overview

Corpus	Analytics date	Language
vec_Latn.jsonl.tsv	11/7/2024	Venetian (vec)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
84,805	1,578,812	706,548 (44.75 %)	45M	210.64 MB	216,477,126

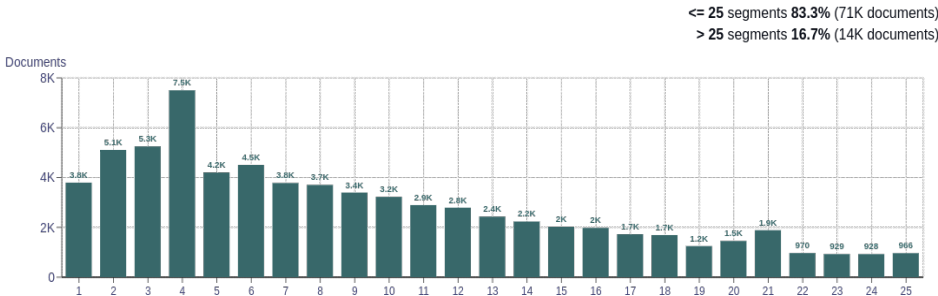
Top 10 domains

Domain	Docs	% of total
wikipedia.org	35K	41.56
meteo-world.com	1.8K	2.11
2night.it	784	0.92
brasitalian.com	770	0.91
verbling.com	571	0.67
gelocal.it	564	0.67
larenadomila.it	555	0.65
sonorika.com	516	0.61
blogspot.com	469	0.55
ilovehomemade.nl	436	0.51

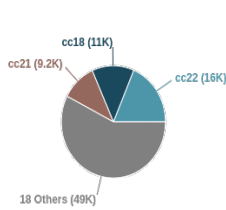
Top 10 TLDs

Domain	Docs	% of total
org	38K	44.93
it	21K	25.10
com	16K	19.41
net	1.7K	2.04
eu	1.2K	1.40
nl	939	1.11
info	435	0.51
de	400	0.47
com.br	394	0.46
la	244	0.29

Documents size (in segments)

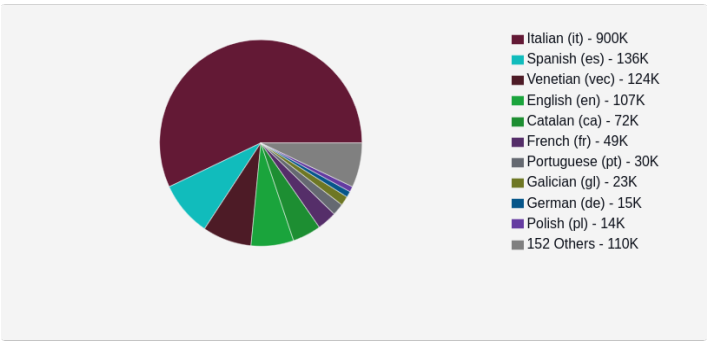


Documents by collection

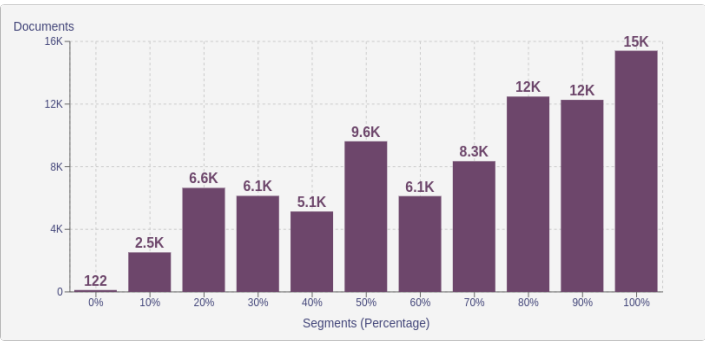


Language Distribution

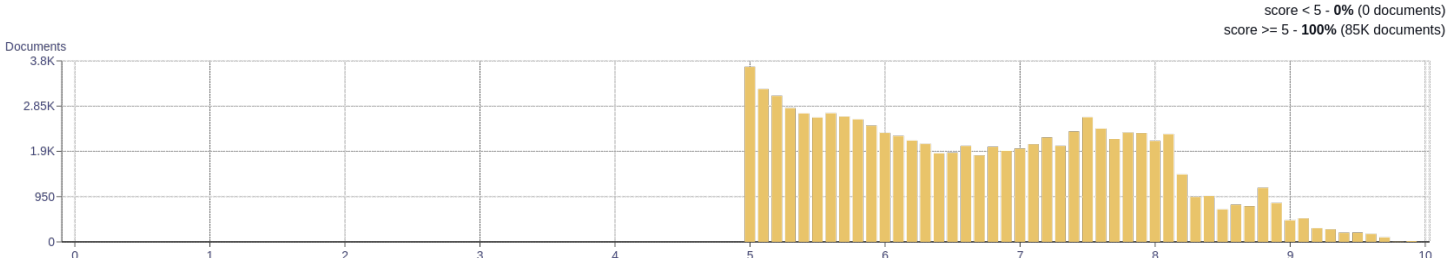
Number of segments



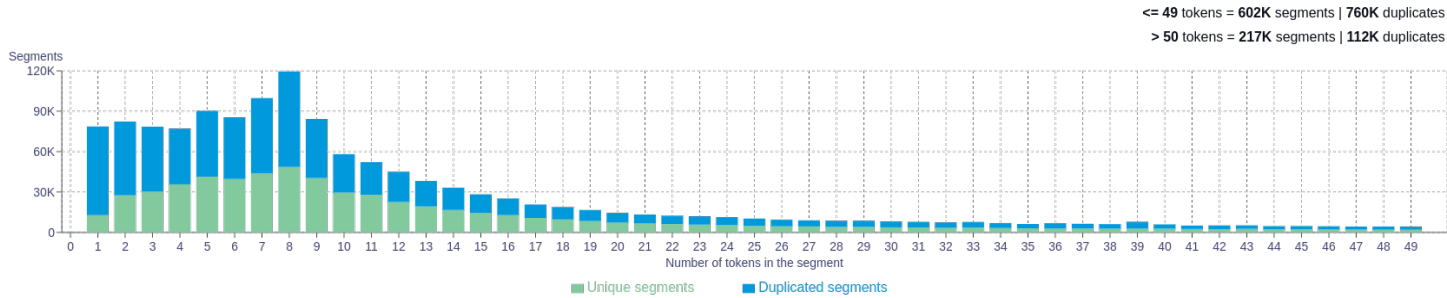
Percentage of segments in Venetian (vec) inside documents



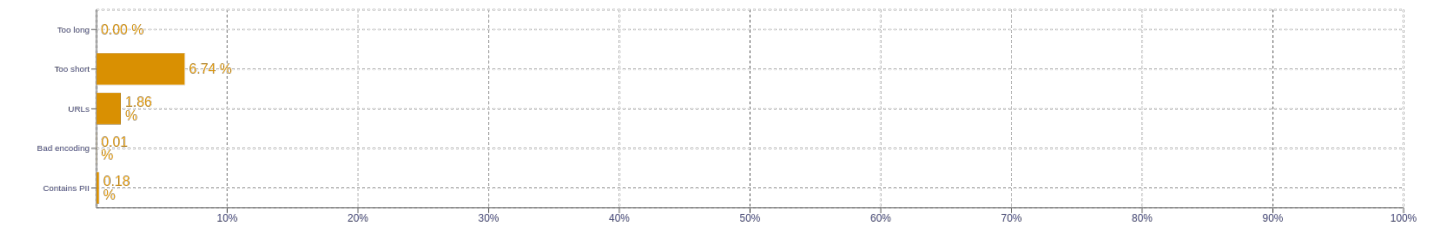
Distribution of documents by document score



Segment length distribution by token



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 of types (uniques)/number of tokens\*, after removing punctuation (<https://www.stinfo.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/mbanoni/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/pabiop16n/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>