

General overview

Corpus	Date	Language
vec_Latn.jsonl.tsv	11/27/2024	Venetian (vec)

Volumes

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

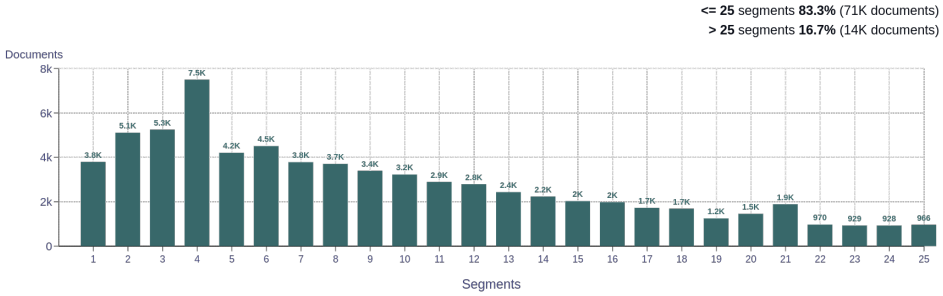
Top 10 domains

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

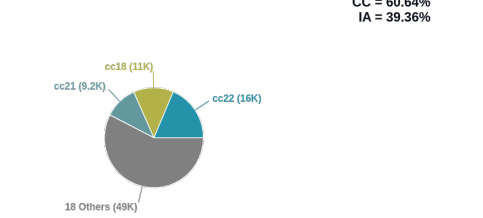
Domain	Docs	% of total
org	38K	44.93
it	21K	25.10
com	16K	19.41
net	1.7K	2.04
eu	1.2K	1.40
nl	939	1.11
info	435	0.51
de	400	0.47
com.br	394	0.46
la	244	0.29

Documents size (in segments)



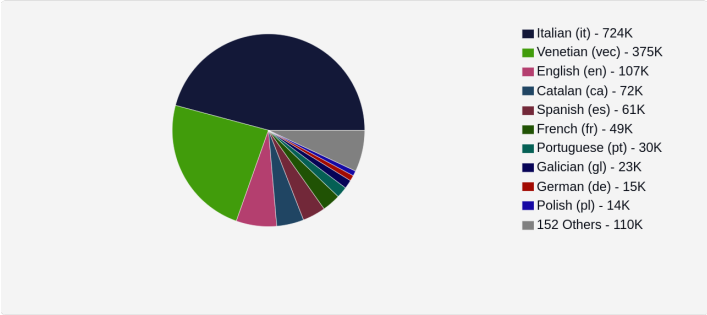
<= 25 segments **83.3%** (71K documents)
> 25 segments **16.7%** (14K documents)

Documents by collection

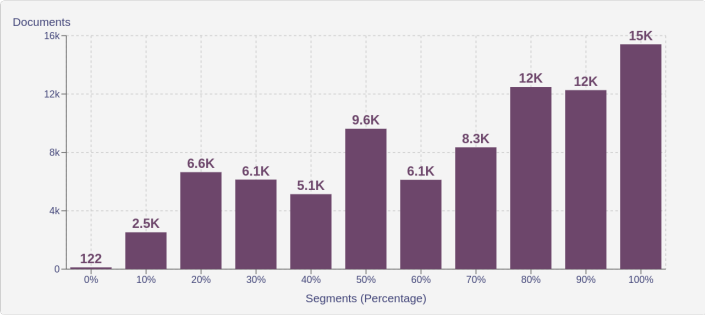


Language Distribution

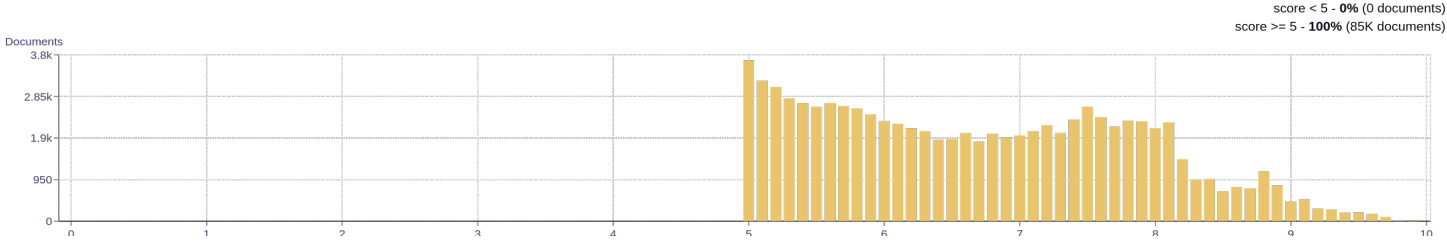
Number of segments in the Venetian (vec) corpus



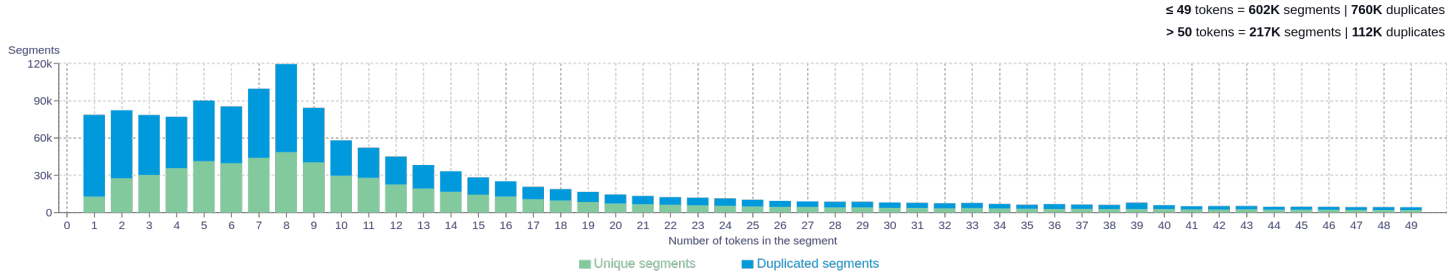
Percentage of segments in Venetian (vec) inside documents



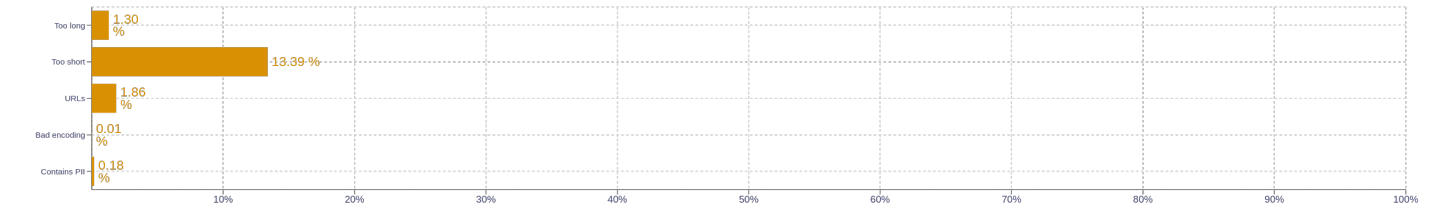
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>canbia 195282</div> <div>d 86576</div> <div>hotel 70986</div> <div>maria 68062</div> <div>roma 58738</div>
2	<div>tutori inglesi 53907</div> <div>canbia sorxente 38745</div> <div>plagio plagio 30804</div> <div>santa maria 21004</div> <div>stemma famiglia 15737</div>
3	<div>plagio plagio plagio 30794</div> <div>canbia el còdaxe 28343</div> <div>canbia el còdexe 24890</div> <div>araldica e stemma 15165</div> <div>case in vendita 10581</div>
4	<div>plagio plagio plagio plagio 30784</div> <div>araldica e stemma famiglia 15165</div> <div>wikimedia commons el detien 6506</div> <div>commons el detien imàjini 6256</div> <div>imàjini o altri file 6252</div>
5	<div>plagio plagio plagio plagio plagio 30777</div> <div>wikimedia commons el detien imàjini 6256</div> <div>detien imàjini o altri file 6252</div> <div>oro oro oro oro oro 3670</div> <div>marmota marmota marmota marmota marmota 3070</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>