

General overview

Corpus	Analytics date	Language
scn_Latn.jsonl.tsv	11/27/2024	Sicilian (scn)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
81,970	1,650,375	735,503 (44.57 %)	53M	246.97 MB	250,748,924

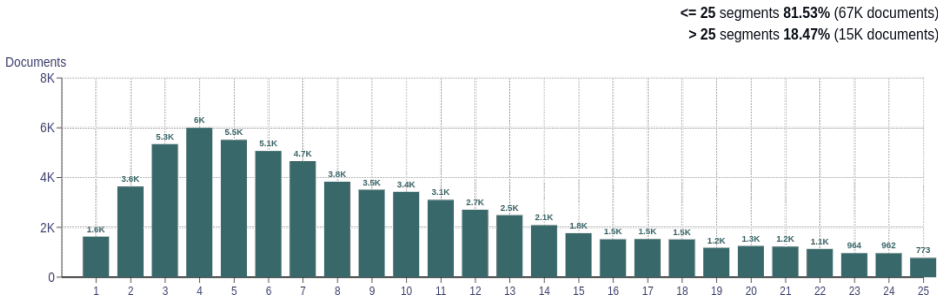
Top 10 domains

Domain	Docs	% of total
wikipedia.org	45K	54.61
vsaduidoma.com	1.6K	1.90
apiazzetta.com	1.5K	1.88
tempicorsica.com	1.1K	1.34
interomania.com	700	0.85
julinse.com	679	0.83
eodishasamachar.com	630	0.77
blogspot.com	541	0.66
arriti.corsica	526	0.64
educationbro.com	485	0.59

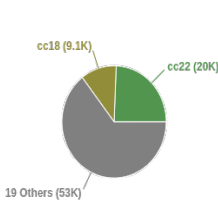
Top 10 TLDs

Domain	Docs	% of total
org	47K	57.82
com	24K	28.97
it	2.8K	3.44
corsica	1.7K	2.12
net	1.3K	1.60
fr	773	0.94
pt	510	0.62
de	330	0.40
eu	325	0.40
zone	294	0.36

Documents size (in segments)

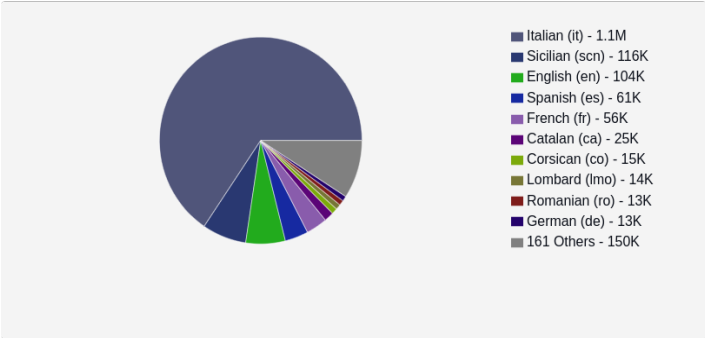


Documents by collection

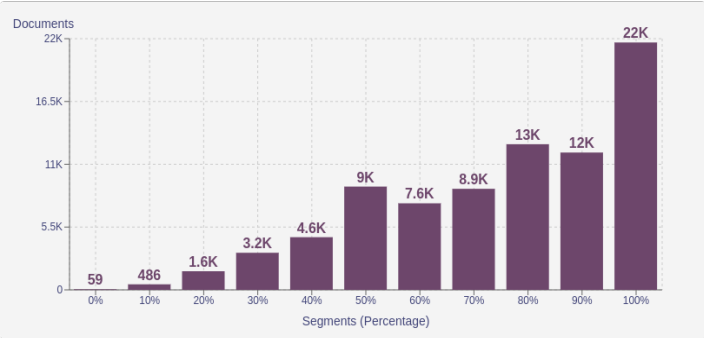


Language Distribution

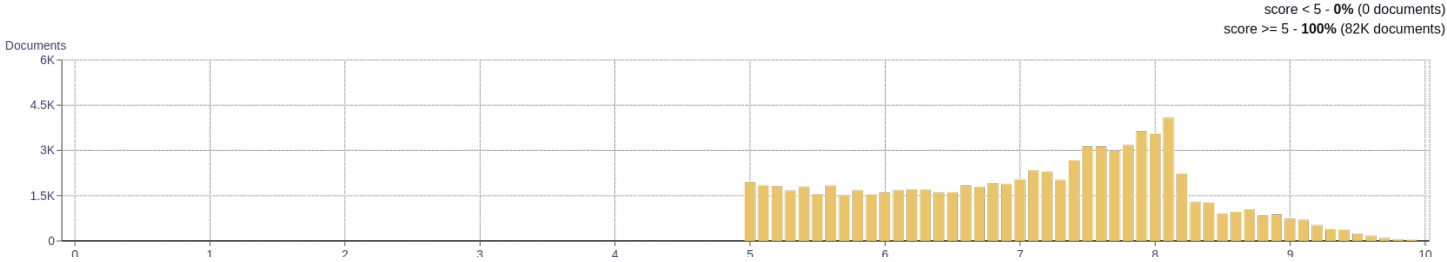
Number of segments



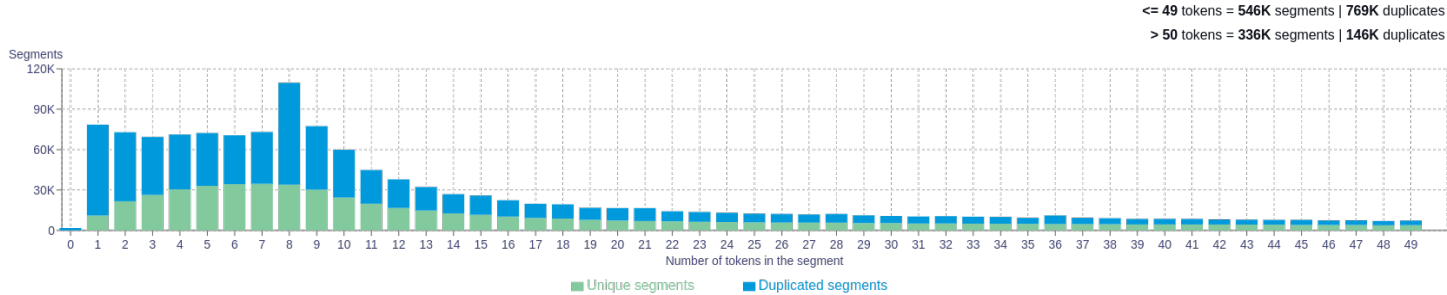
Percentage of segments in Sicilian (scn) inside documents



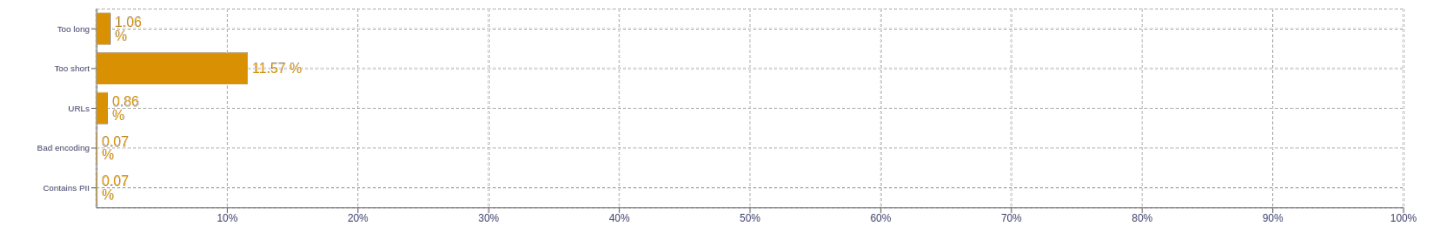
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>à   619243</div> <div>hè   343180</div> <div>chì   330121</div> <div>d   219744</div> <div>cancia   145839</div>
2	<div>nantu à   58728</div> <div>ùn hè   20509</div> <div>chi hè   18890</div> <div>ciò chi   16421</div> <div>hè micca   16307</div>
3	<div>cancia la surgenti   67926</div> <div>edità a fonte   20671</div> <div>ùn hè micca   13615</div> <div>ùn sò micca   6030</div> <div>ùn ci hè   5137</div>
4	<div>ùn ci hè micca   1820</div> <div>seminali e altru outzioni   1692</div> <div>quì ci sò rivista   1692</div> <div>qualità e chemica seminali   1692</div> <div>ordua qualità e chemica   1692</div>
5	<div>senza un ochju è vi   1692</div> <div>outzioni per casi e cottages   1692</div> <div>ordua qualità e chemica seminali   1692</div> <div>ochju è vi maravigghiate cum   1692</div> <div>chemica seminali e altru outzioni   1692</div>

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/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>