

General overview

Corpus	Analytics date	Language
oci_Latn.jsonl.tsv	12/5/2024	Occitan (oc)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
189,906	4,194,906	1,267,789 (30.22 %)	133M	621.05 MB	631,394,992

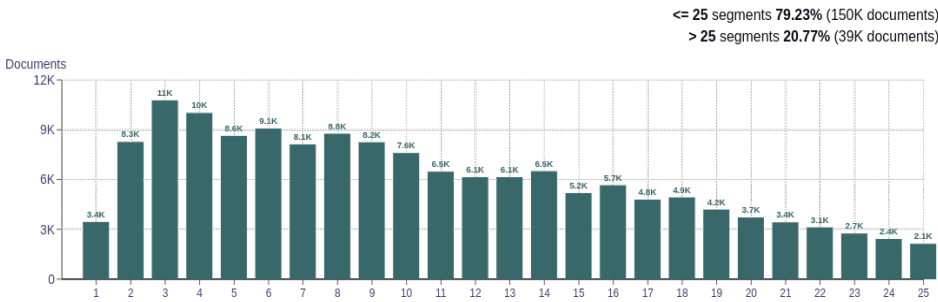
Top 10 domains

Domain	Docs	% of total
wikipedia.org	118K	62.25
jornalet.com	12K	6.16
blogspot.com	4.1K	2.14
vincent-lefrancois.com	3.1K	1.62
occitanparis.com	2.3K	1.21
aquodaqui.info	1.8K	0.94
blogspot.be	1.6K	0.82
sudouest.fr	1.5K	0.81
dordogneilibre.fr	1.2K	0.63
espaci-occitan.com	1.1K	0.60

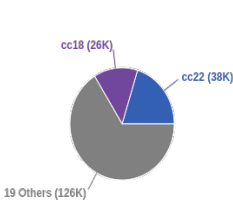
Top 10 TLDs

Domain	Docs	% of total
org	128K	67.17
com	36K	18.95
fr	11K	5.77
eu	4.6K	2.42
cat	2.8K	1.48
info	1.9K	1.02
be	1.6K	0.83
net	1.4K	0.75
com.mx	868	0.46
es	414	0.22

Documents size (in segments)

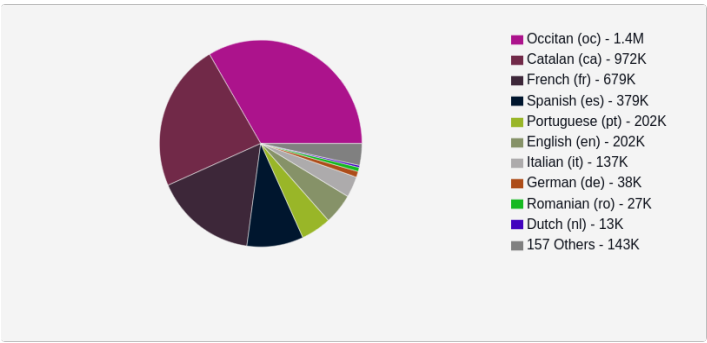


Documents by collection

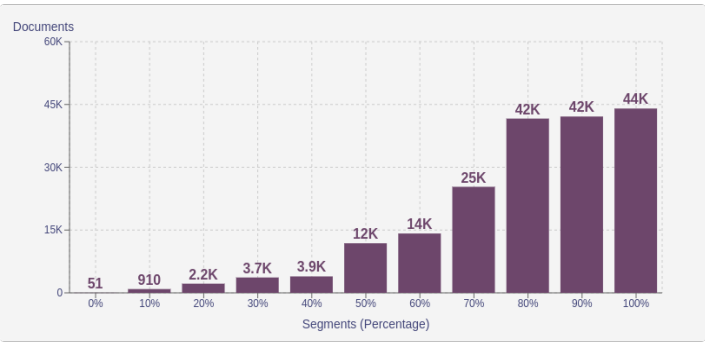


Language Distribution

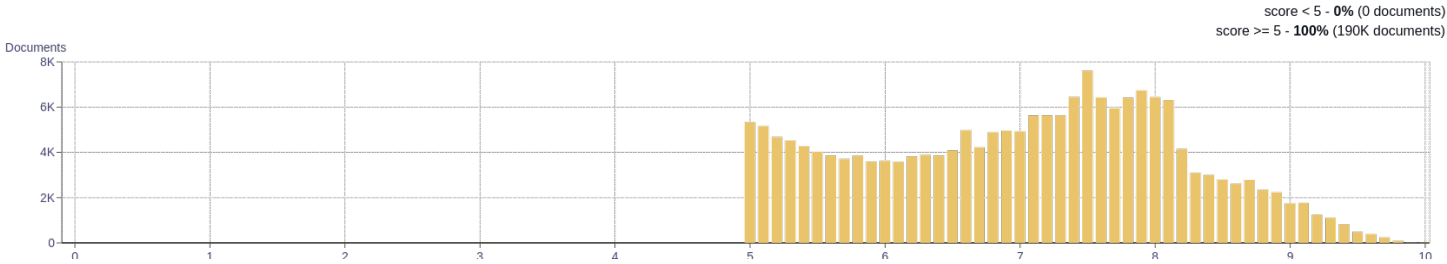
Number of segments



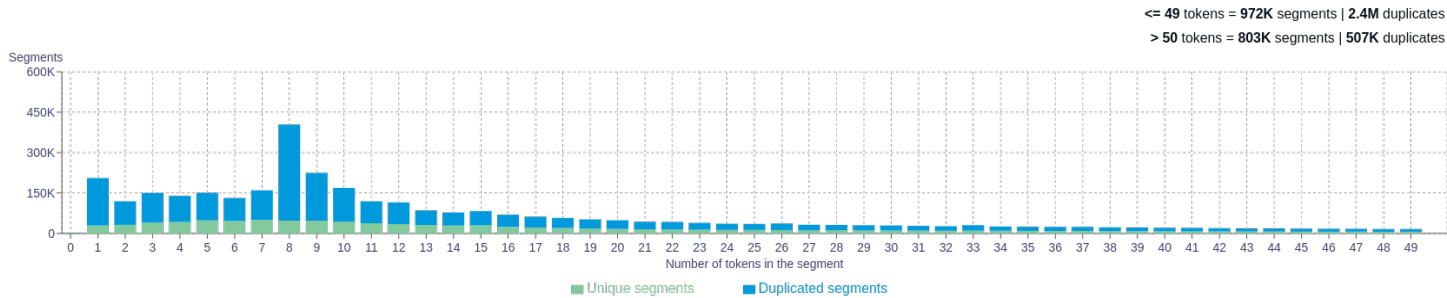
Percentage of segments in Occitan (oc) inside documents



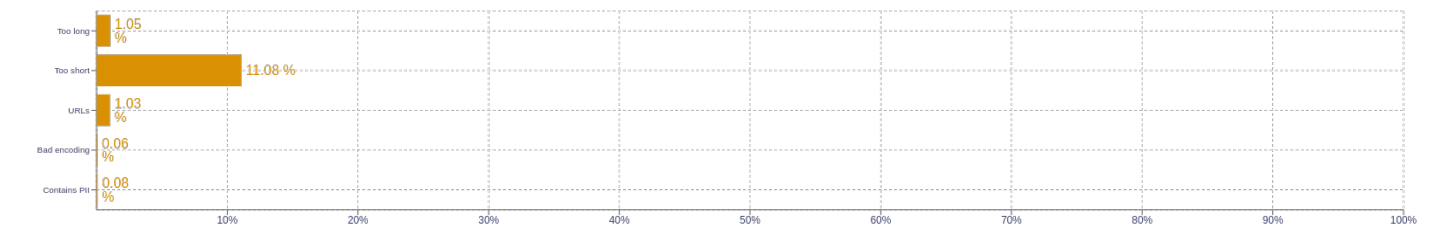
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>e 2490409</div> <div>l 2036095</div> <div>d 1965628</div> <div>en 1652824</div> <div>modificar 1171034</div>
2	<div>e l 65497</div> <div>e d 61088</div> <div>en francés 35377</div> <div>e en 31816</div> <div>en occitan 30570</div>
3	<div>modificar la font 568442</div> <div>e la region 26962</div> <div>luòcs e monuments 15517</div> <div>e de l 14941</div> <div>amb la comuna 14525</div>
4	<div>situada dins lo departament 22448</div> <div>ligadas amb la comuna 14392</div> <div>mail et un site 10245</div> <div>et un site internet 10245</div> <div>aquesta pagina concernís l 8190</div>
5	<div>personalitats ligadas amb la comuna 14381</div> <div>mail et un site internet 10245</div> <div>situada dins lo departament d 5805</div> <div>des noms de lieux en 5259</div> <div>noms de lieux en france 5255</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>