HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
quy_Latn.jsonl.tsv	12/9/2024	Ayacucho Quechua (quy)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
36,940	494,253	213,654 (43.23 %)	23M	138.87 MB	142,953,215

Top 10 domains

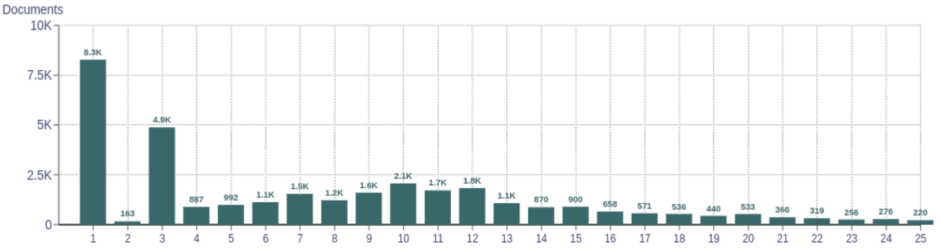
Domain	Docs	% of total
wikipedia.org	14K	38.67
bible.is	13K	34.10
jw.org	3.6K	9.85
mndigital.org	1.3K	3.58
ebible.org	667	1.81
bibles.org	415	1.12
biblegateway.com	388	1.05
wikimedia.org	291	0.79
bible.com	172	0.47
biblica.com	169	0.46

Top 10 TLDs

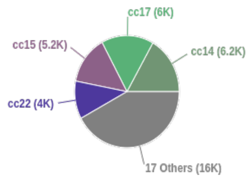
Domain	Docs	% of total
org	21K	57.89
is	13K	34.10
com	1.6K	4.33
gob.ec	412	1.12
net	246	0.67
pe	202	0.55
com.ar	70	0.19
gob.pe	48	0.13
support	46	0.12
ec	38	0.10

Documents size (in segments)

<= 25 segments **90.23%** (33K documents)
> 25 segments **9.77%** (3.6K documents)

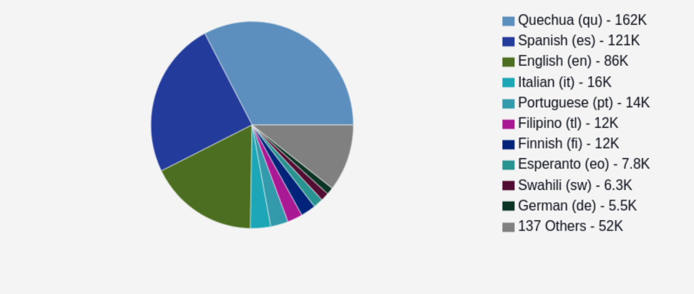


Documents by collection



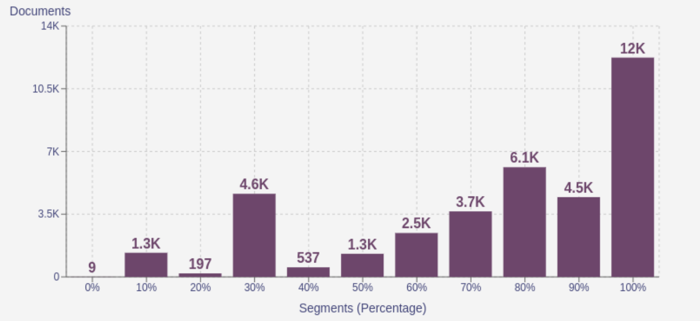
Language Distribution

Number of segments



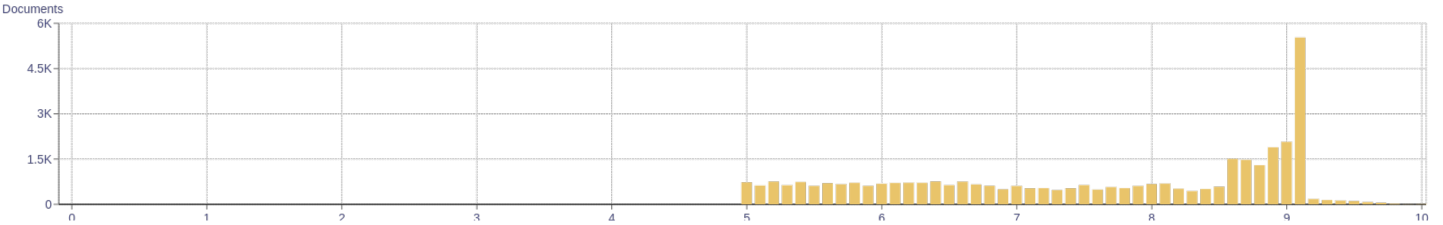
*Ayacucho Quechua (quy) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Ayacucho Quechua (quy) inside documents



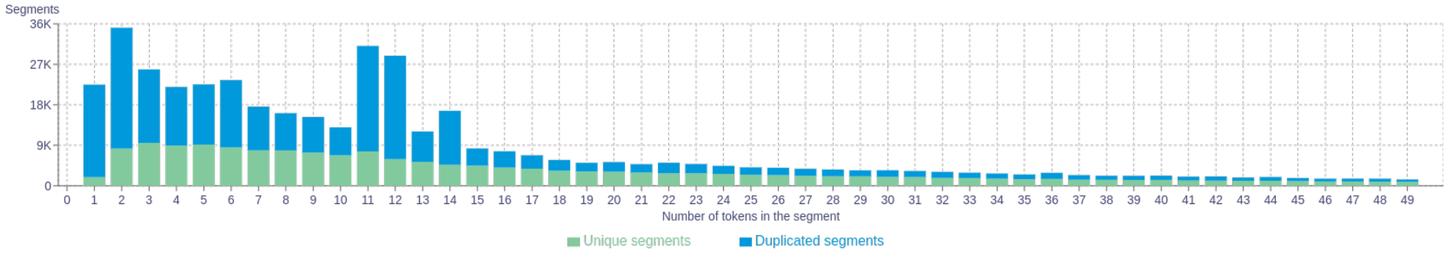
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (37K documents)

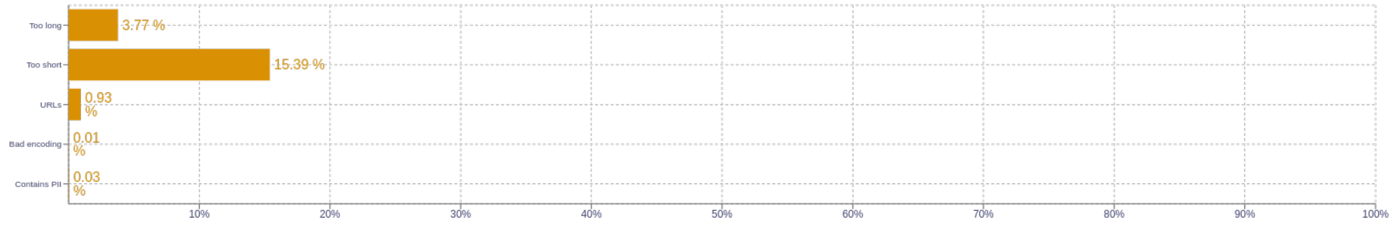


Segment length distribution by token

<= 49 tokens = **171K** segments | **254K** duplicates
> 50 tokens = **69K** segments | **27K** duplicates



Segment noise distribution



Frequent n-grams

<table>
<tr><th>Size</th><th>n-grams</th></tr>
<tr><td>1</td><td>pi | 111878
llamk | 92886
apuy | 90929
kin | 88290
ra | 77059</td></tr>
<tr><td>2</td><td>pukyuta llamk | 43849
mi manchi | 12481
pikachu pikachu | 11594
kin he | 11424
pi pi | 11381</td></tr>
<tr><td>3</td><td>pikachu pikachu pikachu | 10946
hinata hinata hinata | 5710
nisqaqa multimidya kapuyninkunayuqmi | 5408
kapuyninkunayuqmi kay hawa | 5408
commons nisqaqa multimidya | 5408</td></tr>
<tr><td>4</td><td>pikachu pikachu pikachu pikachu | 10869
hinata hinata hinata hinata | 5705
multimidya kapuyninkunayuqmi kay hawa | 5408
commons nisqaqa multimidya kapuyninkunayuqmi | 5408
spamspamspamspamspam spamspamspamspamspam spamspamspamspamspam spamspamspamspamspam | 2930</td></tr>
<tr><td>5</td><td>pikachu pikachu pikachu pikachu pikachu | 10827
hinata hinata hinata hinata hinata | 5700
nisqaqa multimidya kapuyninkunayuqmi kay hawa | 5408
spamspamspamspamspam spamspamspamspamspam spamspamspamspamspam spamspamspamspamspam spamspamspamspamspam | 2921
nisqapi suyukunata uyarinakunatapas tarinki kaymantam | 2681</td></tr>
</table>

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