HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
cy_1.jsonl.tsv	3/16/2024	Welsh (cy)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
111,254	12,687,508	3,487,441 (27.49 %)	151M	733.16 MB	

Top 10 domains

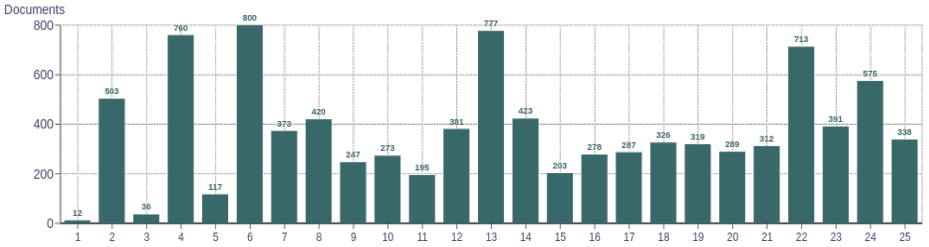
Domain	Docs	% of total
wikipedia.org	7.9K	7.12
sgames.org	3.2K	2.88
golwg360.cymru	2.2K	1.97
llyfrgell.cymru	2.1K	1.89
southwales.ac.uk	1.8K	1.59
testunau.org	1.6K	1.44
blogspot.co.uk	1.5K	1.39
whatdotheyknow.com	1.3K	1.21
ofunnygames.com	1.3K	1.15
llyw.cymru	1.2K	1.10

Top 10 TLDs

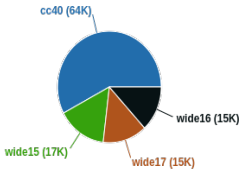
Domain	Docs	% of total
com	27K	24.37
cymru	25K	22.35
org	21K	19.27
gov.uk	8.9K	8.01
ac.uk	8.3K	7.50
co.uk	7.8K	7.01
org.uk	3.7K	3.32
wales	2.8K	2.56
net	1.4K	1.23
news	676	0.61

Documents size (in segments)

<= 25 segments **8.41%** (9.3K documents)
> 25 segments **91.59%** (102K documents)

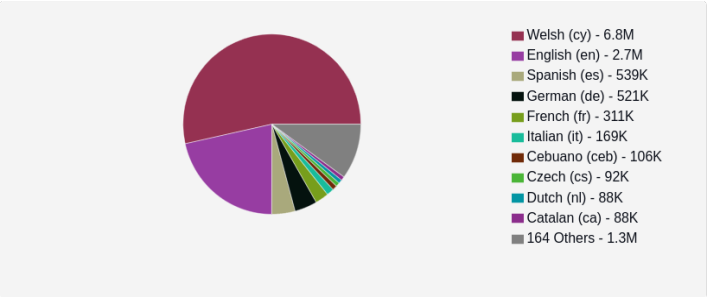


Documents by collection

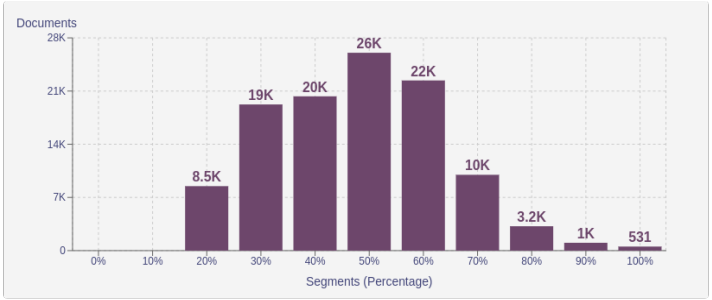


Language Distribution

Number of segments

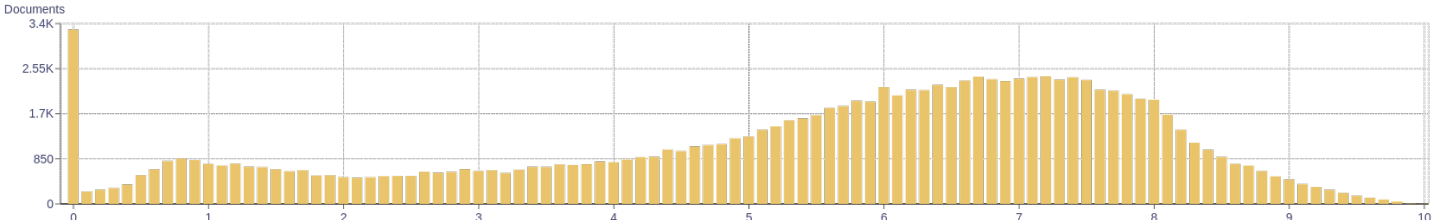


Percentage of segments in Welsh (cy) inside documents



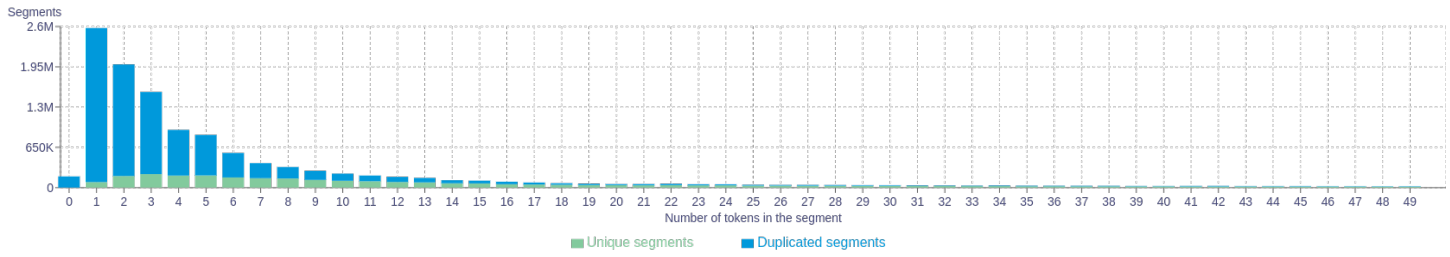
Distribution of documents by document score

score <= 5 - **33.35%** (37K documents)
score > 5 - **66.65%** (74K documents)

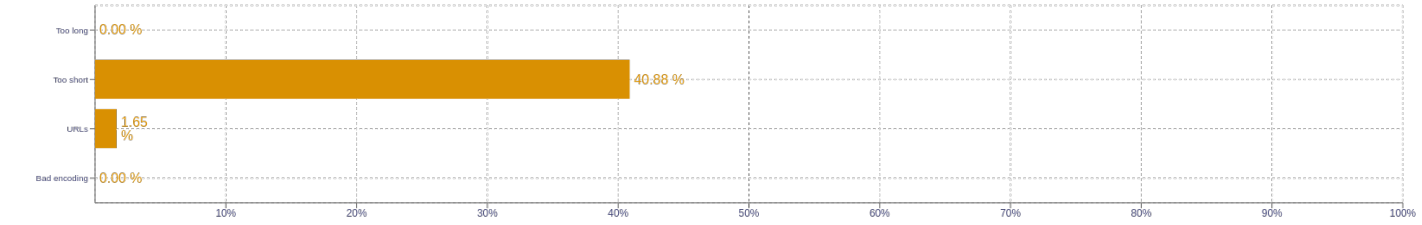


Segment length distribution by token

<= 49 tokens = **3M** segments | **9.1M** duplicates
> 50 tokens = **665K** segments | **138K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	r | 1130142 · n | 691744 · cymru | 447316 · newydd | 194136 · gemau | 142327
2	llywodraeth cymru | 30872 · polisi preifatrwydd | 27898 · support script | 23697 · senedd chevron | 20876 · gemau ar-lein | 20452
3	cod y dudalen | 18584 · telerau ac amodau | 18499 · share to facebook | 14947 · share to twitter | 14944 · copy to clipboard | 14331
4	browser does not support | 23791 · golygu cod y dudalen | 18512 · share to twitter share | 14943 · share to facebook share | 14334 · copy to clipboard share | 14330
5	browser does not support script | 23697 · clipboard share to facebook share | 14328 · facebook share to twitter share | 14327 · twitter share to linkedin video | 9548 · newidiwyd y dudalen hon ddiwethaf | 6644

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