

General overview

Corpus	Analytics date	Language
ssw_Latn.jsonl.tsv	9/22/2024	Swati (ss)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
2,036	62,126	39,431 (63.47 %)	1.3M	8.42 MB	8,760,975

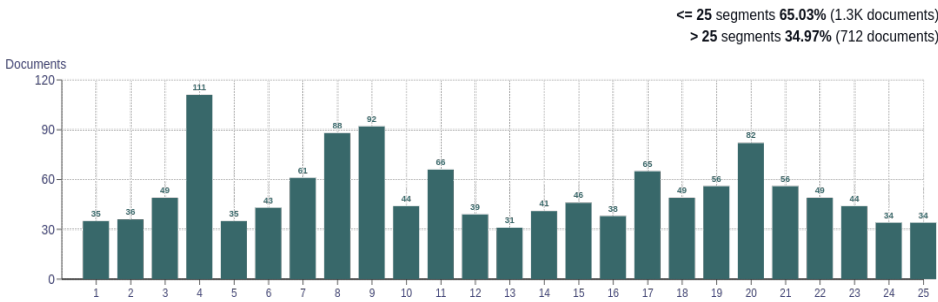
Top 10 domains

Domain	Docs	% of total
jw.org	599	29.42
biblesa.co.za	446	21.91
wikipedia.org	400	19.65
southafrica.co.za	242	11.89
myconstitution.co.za	27	1.33
sciencegraph.net	20	0.98
gotquestions.org	15	0.74
nalibali.org	15	0.74
digitallibrary.io	13	0.64
shuters.co.za	12	0.59

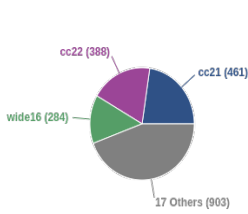
Top 10 TLDs

Domain	Docs	% of total
org	1.1K	53.14
co.za	776	38.11
com	54	2.65
net	26	1.28
gov.za	22	1.08
org.za	20	0.98
io	14	0.69
ac.za	11	0.54
co.uk	7	0.34
co.sz	7	0.34

Documents size (in segments)

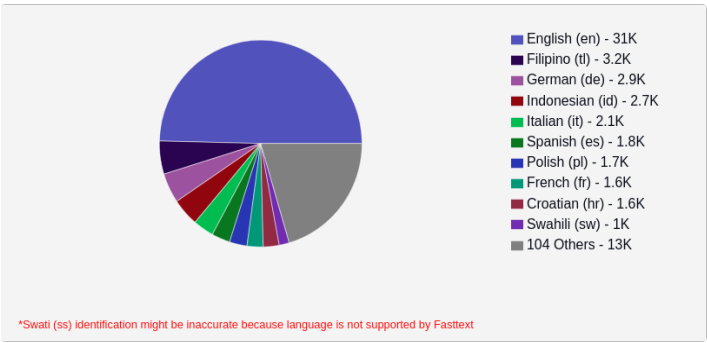


Documents by collection

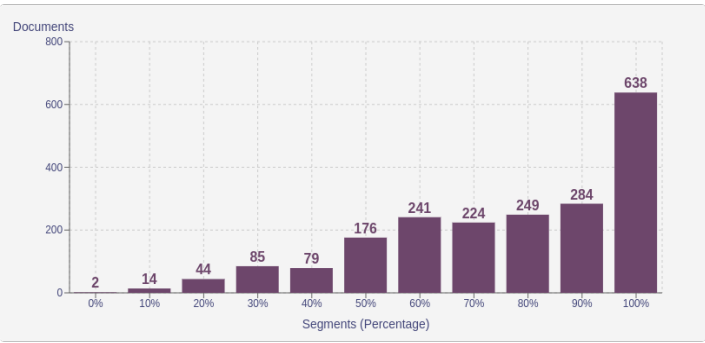


Language Distribution

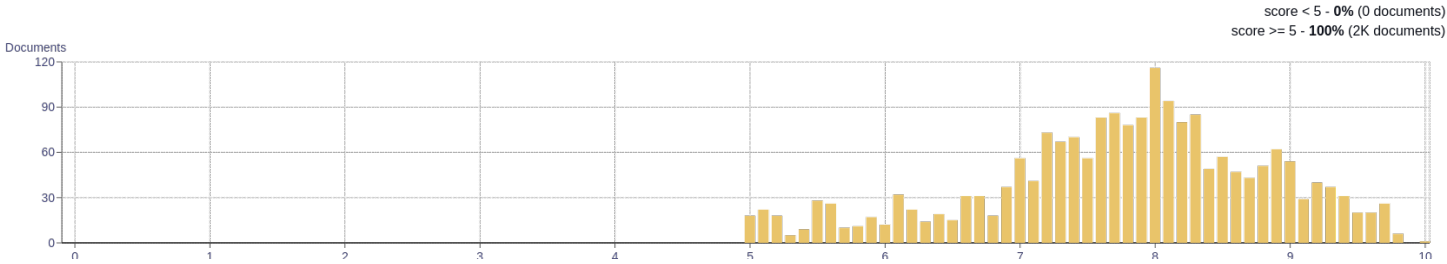
Number of segments



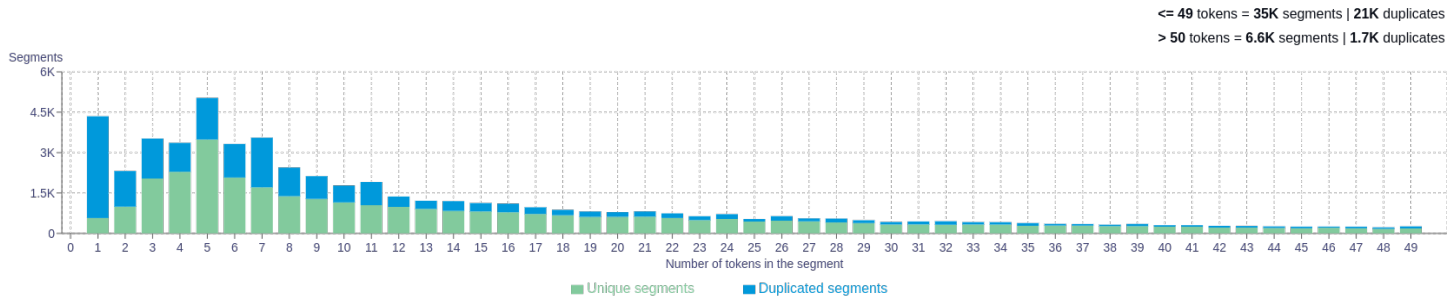
Percentage of segments in Swati (ss) inside documents



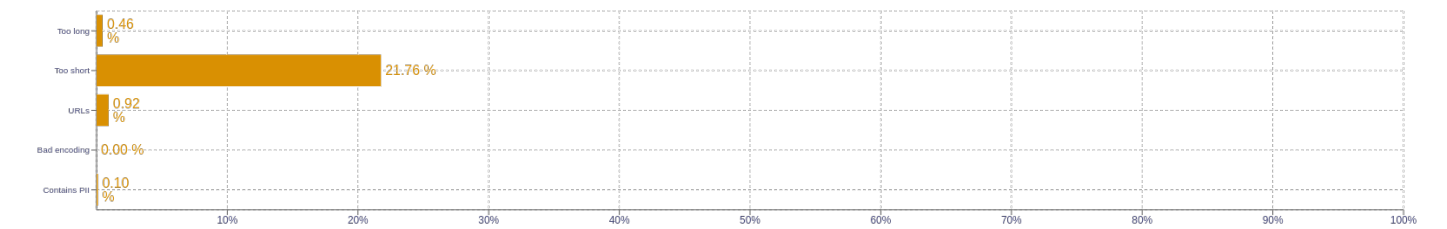
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>stfu 8185</div> <div>noobs 8185</div> <div>bantfu 3879</div> <div>nkulunkulu 3525</div> <div>ke 2986</div>
2	<div>stfu noobs 8185</div> <div>edit source 1029</div> <div>wonkhe umuntfu 520</div> <div>bonkhe bantfu 445</div> <div>uma ngabe 433</div>
3	<div>emazinga kutemfundvo nekucecesha 270</div> <div>cweluco cweluco cweluco 236</div> <div>wekucinisekisa emazinga kutemfundvo 234</div> <div>wonkhe umuntfu unelilungelo 193</div> <div>new world translation 121</div>
4	<div>cweluco cweluco cweluco cweluco 235</div> <div>wekucinisekisa emazinga kutemfundvo nekucecesha 228</div> <div>translated by phindile malotana 109</div> <div>world translation of the 103</div> <div>ungatsandza yini kufundza lesihloko 103</div>
5	<div>cweluco cweluco cweluco cweluco cweluco 234</div> <div>world translation of the holy 103</div> <div>ungatsandza yini kufundza lesihloko ngalolulwimi 103</div> <div>translation of the holy scriptures 103</div> <div>new world translation of the 103</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>