HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
slk_Latn.jsonl.tsv	6/12/2025	Slovak (sk)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
21,827,246	494,141,968	210,163,626 (42.53 %)	12B	69,899,657,544	70.65 GB

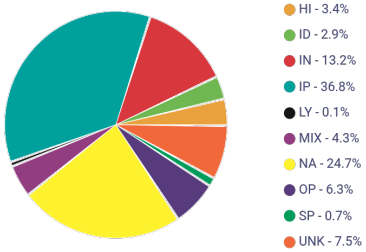
Top 10 domains

Domain	Docs	% of total
sme.sk	1.3M	6.01%
wikipedia.org	360K	1.65%
firebaseapp.com	291K	1.33%
web.app	289K	1.32%
pravda.sk	222K	1.02%
zdravie.sk	188K	0.86%
blogspot.com	161K	0.74%
24hod.sk	147K	0.67%
aktuality.sk	142K	0.65%
netky.sk	118K	0.54%

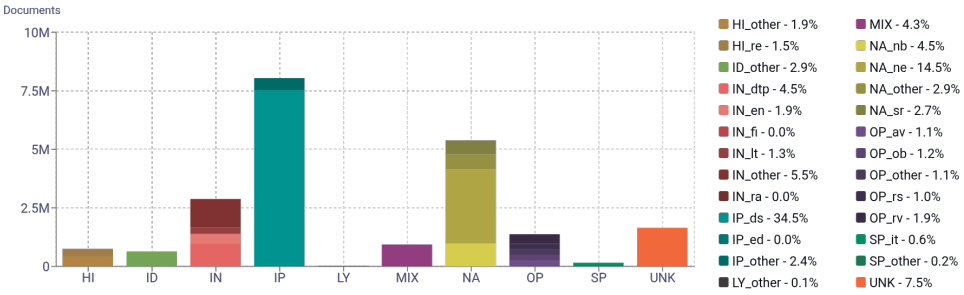
Top 10 TLDs

Domain	Docs	% of total
sk	17M	77.32%
com	2.1M	9.44%
cz	665K	3.05%
eu	628K	2.88%
org	526K	2.41%
app	289K	1.33%
net	212K	0.97%
info	122K	0.56%
ru	31K	0.14%
xyz	26K	0.12%

Register labels

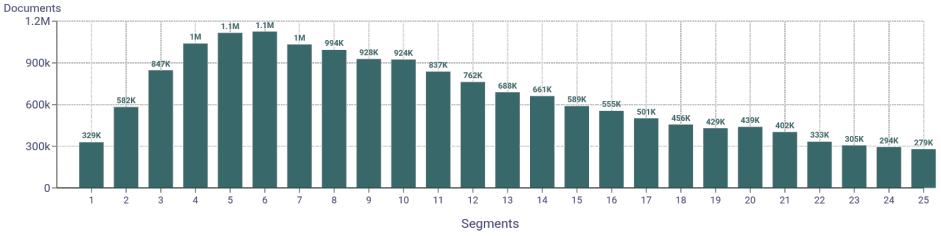


MT:5.2% | 1.1M Documents



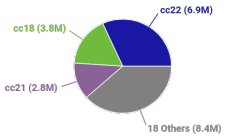
Documents size (in segments)

<= 25 segments **75.35%** (16M documents)
> 25 segments **24.65%** (5.4M documents)



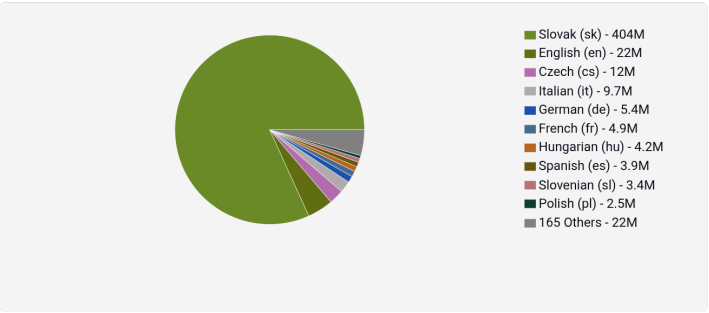
Documents by collection

CC = 71.70%
IA = 28.30%

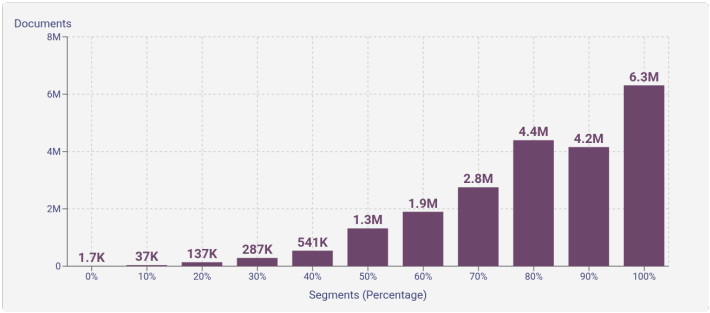


Language Distribution

Number of segments in the Slovak (sk) corpus

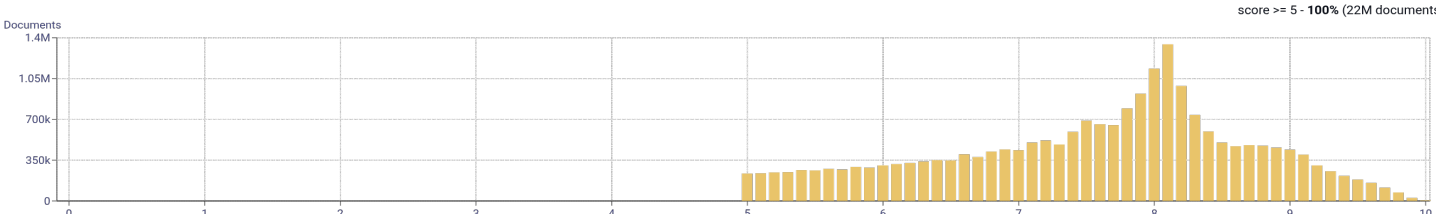


Percentage of segments in Slovak (sk) inside documents



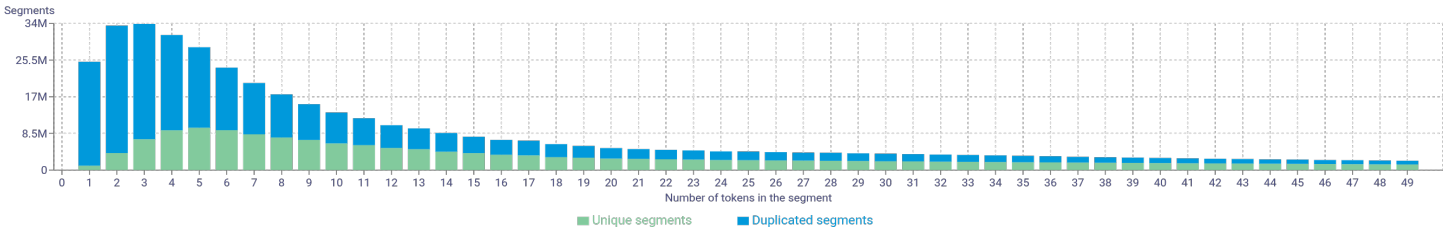
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (22M documents)

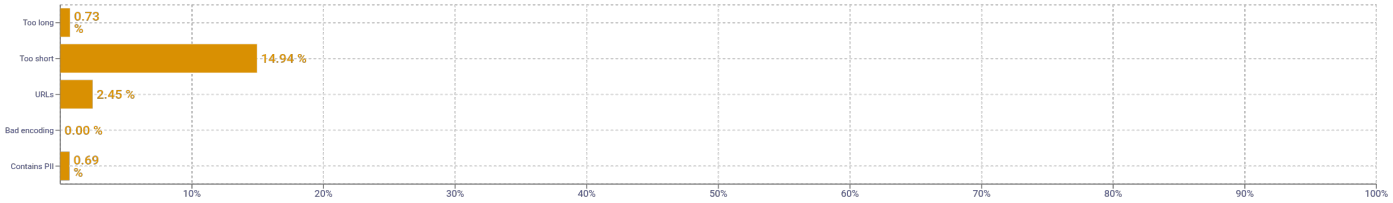


Segment length distribution by token

≤ 49 tokens = **163M** segments | **255M** duplicates
> 50 tokens = **76M** segments | **29M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	veľmi | 12389520 · roku | 12279902 · vás | 8861559 · všetky | 8764120 · u | 8655439
2	osobných údajov | 1630085 · u nás | 1542687 · slovenskej republiky | 1534554 · napriek tomu | 1427423 · upraviť zdroj | 1273910
3	znení neskorších predpisov | 566554 · ponúkame na predaj | 393576 · zmene a doplnení | 359440 · príspevok k tejto | 319844 · doplnení niektorých zákonov | 311756
4	napíše príspevok k tejto | 319633 · príspevok k tejto položke | 319629 · zmene a doplnení niektorých | 313244 · hotel počas poslednej hodiny | 193452 · prezeralo tento hotel počas | 190122
5	napíše príspevok k tejto položke | 319629 · zmene a doplnení niektorých zákonov | 303964 · Ľudí si prezeralo tento hotel | 190671 · prezeralo tento hotel počas poslednej | 190122 · najdôležitejšie správy z východu slovenska | 147665

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

Register labels

Name	Abbr.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	**HI**
Recipe	re
Informational persuasion	**IP**
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	**IN**
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	**OP**
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av