HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
ace_Arab.jsonl.tsv	12/3/2024	Acehnese (ace)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
16	117	93 (79.49 %)	12K	84.34 KB	49,626

Top 10 domains

Domain	Docs	% of total
iran-eng.ir	4	25.00
blogspot.com	2	12.50
bayardserviceweb.com	1	6.25
utusanmelayu.com.my	1	6.25
office-converter.com	1	6.25
salemalanzi.sa	1	6.25
onlineyoutube.com	1	6.25
all.biz	1	6.25
blogspot.it	1	6.25
blogspot.my	1	6.25

Top 10 TLDs

Domain	Docs	% of total
com	6	37.50
ir	4	25.00
com.my	1	6.25
sa	1	6.25
biz	1	6.25
it	1	6.25
my	1	6.25
net	1	6.25

Documents size (in segments)

<= 25 segments **100%** (16 documents)
> 25 segments **0%** (0 documents)
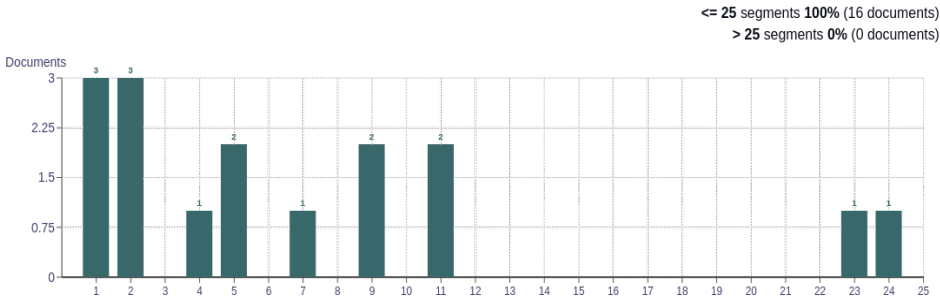
Documents by collection

wide15 (6)
wide16 (3)
cc18 (2)
wide11 (2)
3 Others (3)
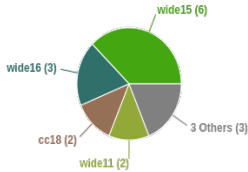
Language Distribution

Number of segments

- Arabic (ar) - 76
- Persian (fa) - 16
- French (fr) - 7
- English (en) - 6
- Malay (ms) - 4
- German (de) - 3
- Egyptian Arabic (arz) - 3
- Indonesian (id) - 2

*Acehnese (ace) identification might be inaccurate because language is not supported by Fasttext
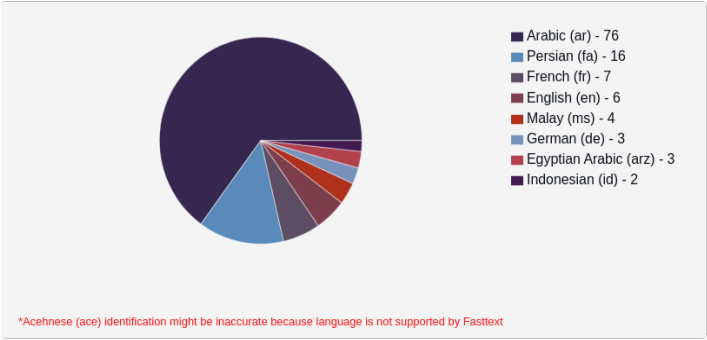
Percentage of segments in Acehnese (ace) inside documents
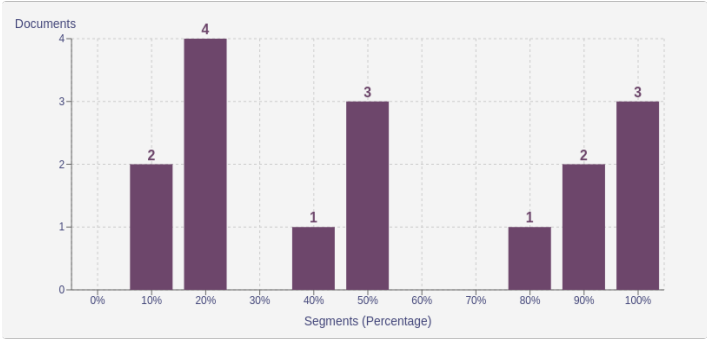
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (16 documents)

Segment length distribution by token

<= 49 tokens = **78** segments | **18** duplicates
> 50 tokens = **21** segments | **6** duplicates

- Unique segments
- Duplicated segments

Segment noise distribution

Too long	7.69 %
Too short	9.40 %
URLs	0.00 %
Bad encoding	0.00 %
Contains PII	0.85 %
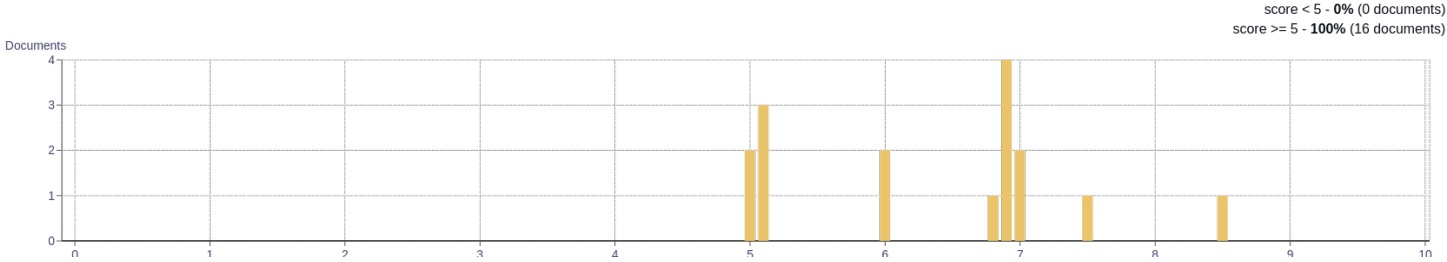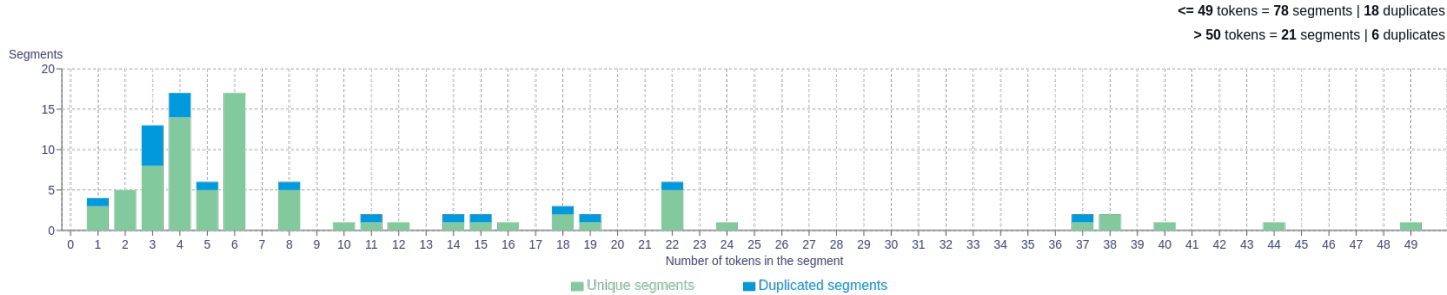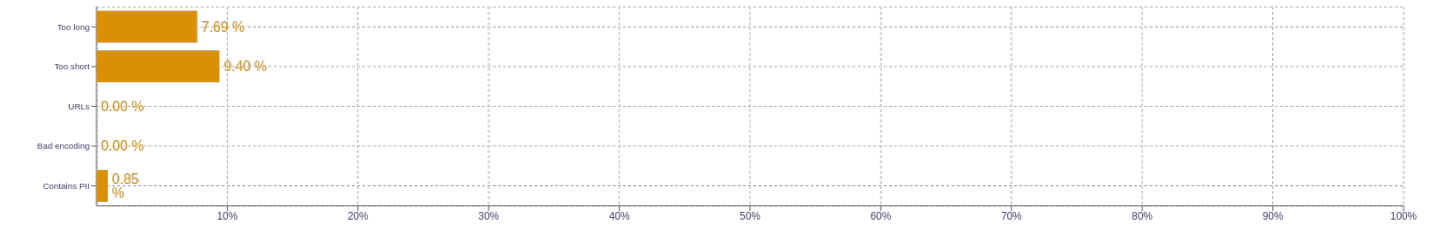
Frequent n-grams

Size	n-grams
1	کن | 1683 مدون | 1680 کجاس | 1530 خونتون | 1530 اوک | 756
2	مدون کن | 1676 کن مدون | 1673 خونتون کجاس | 1530 ل ppm | 336 اوک اوک | 755
3	کن مدون کن | 1670 مدون کن مدون | 1666 اوک اوک اوک | 754 هیوندای جینیسیس کوبیه | 56 چت ازبتا ازبتا چت | 24
4	مدون کن مدون کن | 1665 کن مدون کن مدون | 1660 اوک اوک اوک اوک | 753 سعر هیوندای جینیسیس کوبیه | 13 فورتل این اکن داقت | 8
5	کن مدون کن مدون کن | 1659 مدون کن مدون کن مدون | 1655 اوک اوک اوک اوک اوک | 752 مدونکن مدونکن مدونکن کن | 6 مدونکن مدونکن مدونکن کن | 6

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