به نام خدا

تارا قشلاقی- 99443009

در این تمرین ، ۶ tokenizer بررسی شده است:

- parsivar
- polyglot
- farsiyar
- spacy
- nltk
- parsipardaz

برای هر کدام از tokenizer ها، نتیجه‌ی توکن‌ها آورده شده است که با علامت = مشابه روش دستی از هم جدا شدند و برای کلمات مرکب اجزا با نیم فاصله. (برای هر ۲۰۰ جمله)

به منظور مقایسه بین tokenizer ها سه مقدار محاسبه شده است که در جدول اکسل آورده شده است. ستون هایی که برای نتیجه‌ی توکنایز شدن برای هر جمله آورده شده است:

#token_reconized_re

تعداد توکن هایی که از بین توکن های روش دستی تشخیص داده است.

is_same_re

آیا تمامی توکن‌های جمله را تشخیص داده است؟ (1 / 0)

#compound_words_true_re

از بین کلمات مرکب جمله چند مورد با توکنایزر درست تشخیص داده شده است

سپس در ادامه با استفاده از این مقادیر برای هر توکنایزر سه عدد محاسبه کرده‌ایم:

precision_tokens

دقت تشخیص توکن های درست

precision_compounds

دقت تشخیص کلمات مرکب درست

precision_sentence

تعداد جملاتی که تمامی توکن ها را تشخیص داده است

نتایج برای هر tokenizer به صورت زیر می‌باشد(برای ۱۵۶ جمله‌ی اول) :

parsivar:

precision_tokens
75.99750623
precision_compounds
1.234567901
precision_sentence
0.070512821

polyglot:

precision_tokens
73.69077307
precision_compounds
0
precision_sentence
0.057692308

farsiyar:

precision_tokens
73.87780549
precision_compounds
0
precision_sentence
0.070512821

spacy:

precision_tokens
73.31670823

precision_compounds
0
precision_sentence
0.064102564

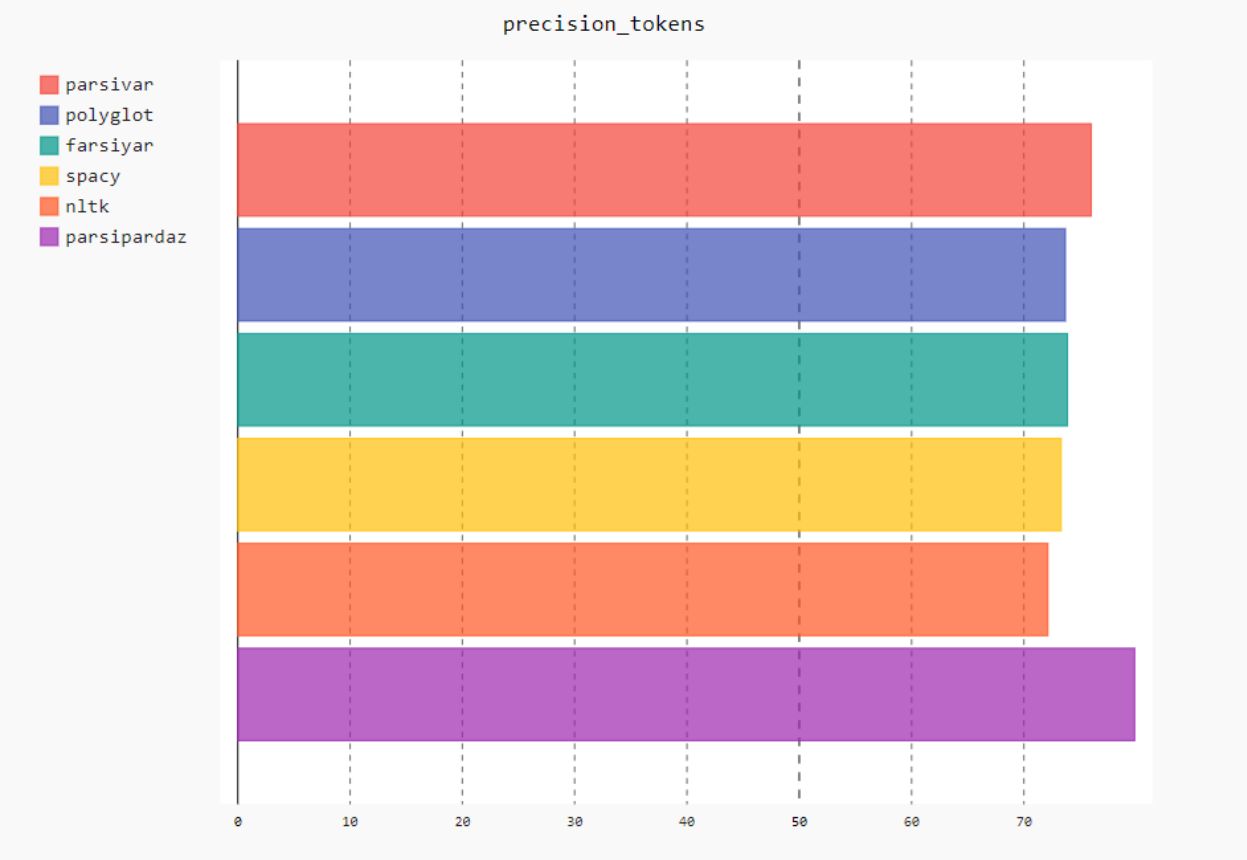
nltk

precision_tokens
72.13216958
precision_compounds
0
precision_sentence
0.051282051

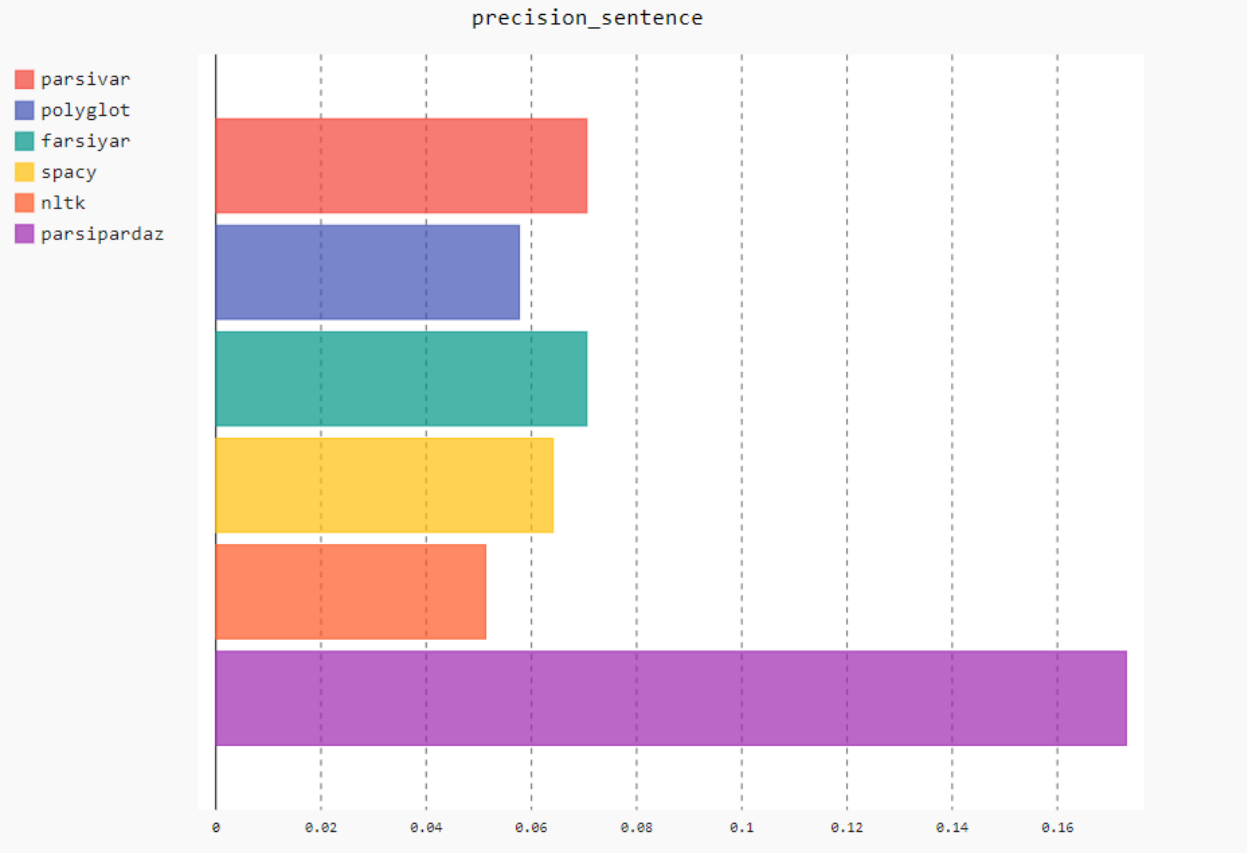
parsipardaz:

precision_tokens
79.86284289
precision_compounds
2
precision_sentence
0.173076923

برای مقایسه در یک نمودار آورده شده‌اند:



parsipardas >> parsivar >> farsivar =~ polyglot >> spacy >>nltk



parsipardas >> parsivar =~ farsivar >> spacy >> polyglot >> nltk

نتیجه : با مقایسه ی هر ۶ تا به نظر می‌رسد parsipardaz عملکرد بهتری دارد.

*کدهای استفاده شده برای برخی از توکنایزرها که api آنها در دسترس بود در *لینک* موجود است.

از توکنایزر دیگری به اسم toktok نیز استفاده شد ولی چون عملکرد پایینی داشت، آورده نشد ولی به گفته ی مستند آمده برای آن سرعت بالایی دارد.

هم چنین برای افزایش دقت در محاسبات کدی پیاده‌سازی شد که فایل آن آورده شده است.

هم چنین تعدادی از جملات ۱۵۷ به بعد نیز به صورت دستی tokenize شدند.