**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 بینتوکن های روش دستی تشخیص داده است.

**is\_same\_re**

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

**#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 |

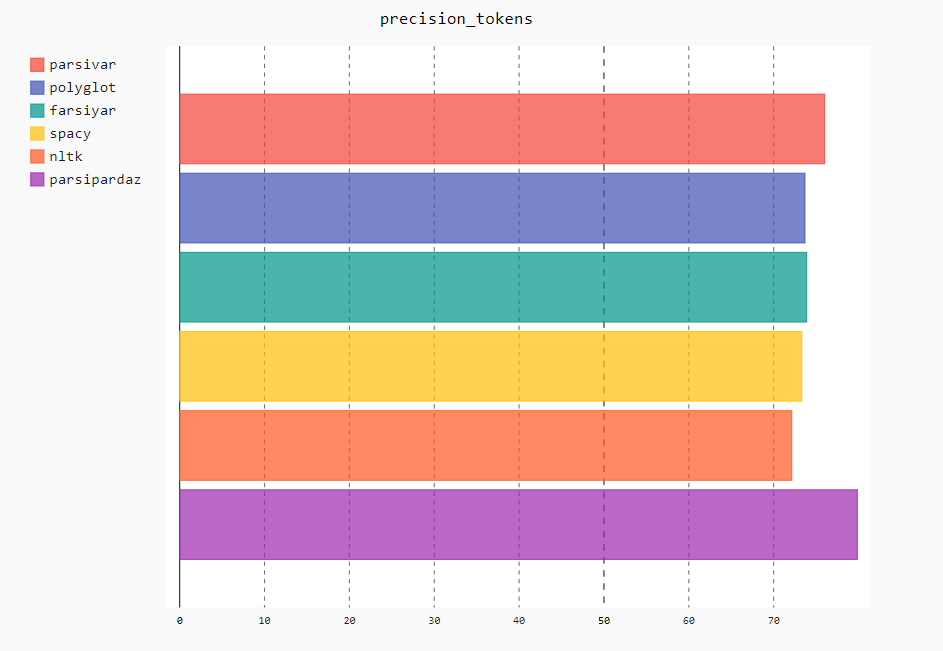
nltk

|  |
| --- |
| **precision\_tokens** |
| 72.13216958 |
| **precision\_compounds** |
| 0 |
| **precision\_sentence** |
| 0.051282051 |

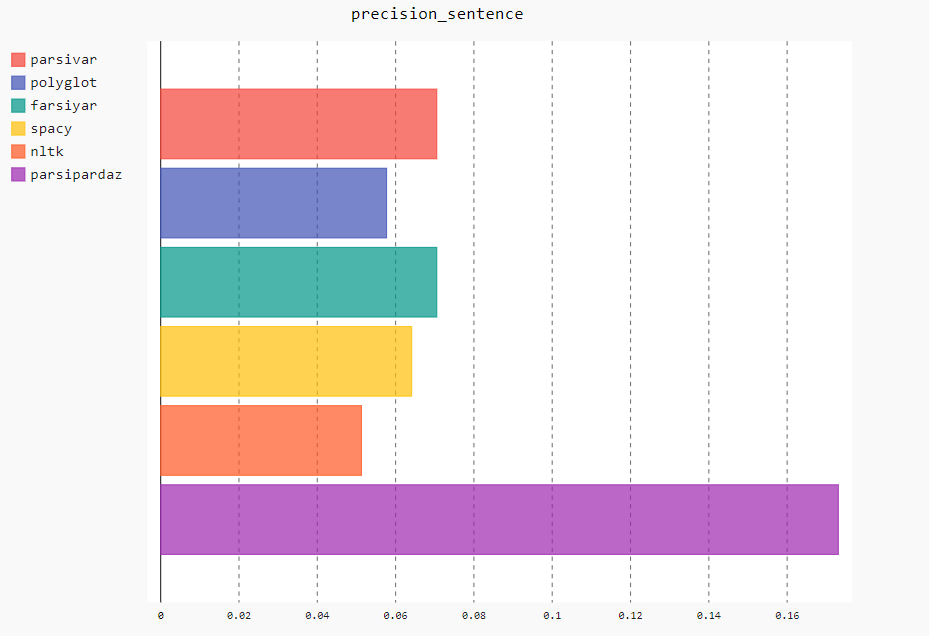
parsipardaz:

|  |
| --- |
| **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*](https://colab.research.google.com/drive/1C2xu8BSwy9ShRVPSLAsCnjbvy7D4rJRk?usp=sharing) 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.