**به ‏نام خدا**

تارا قشلاقی- 99443009

لینک

لینک پیاده‏سازی finetuning و نتایج (بخش *Test model*) در [لینک](https://colab.research.google.com/drive/13te2BE_Xe7Uu6rDyeWYKJJwGHySprNo-?usp=sharing) موجود است.

در فاین تیون کردن از مدل parsber v2 استفاده شده است که در حین مقدار دهی کردن tokenizer، config و model از

model\_name\_or\_path = 'HooshvareLab/bert-fa-base-uncased'

استفاده کرده‏ایم.

در ابتدا داده های موجود در فایل .txt را استخراج می‏کنیم که با توجه به پیکره‏ی PEYMA بعد از هر خط خالی، جمله‏ی جدیدی شروع خواهد شد و تگ و کلمه نیز با | از یکدیگر جدا شده اند پس با استخراج جملات و کلمه ها و تگ ها، عملیات پیش‏پردازش را انجام می‏دهیم تا داده آماده‏ی ورودی دادن به مدل برت شود.

به این صورت که بعد از فراخوانی tokenizer مربوطه روی هر کلمه که خود ممکن است شامل چند توکن شود، تگ آن کلمه را به تعداد توکن هایی که حاصل از tokenize شدن هستند تکرار می کنیم و در لیست مربوط به تگ ها متناظر با جمله، اضافه می شود سپس توکن ها را به عدد تبدیل می کنیم (قابل ورودی دادن به مدل)

همچنین لیبل ها را نیز به عدد تبدیل می کنیم (از قبل به هذ تگ عدد خاصی نسبت می دهیم)

سپس برای عملیات padding، ابتدا maxlen‌را به دست می‏آوریم ولی عددی حدود ۴۰۰ به دست آمد که طبق این باید maxlen را برابر با ۵۱۲ قرار داد و همین باعث می شد به دلیل حافظه، کد ارور بدهد. بنابراین از عدد رایج ۳۲ در بیشتر کدها استفاده کردیم.

(دلیل عدد ۴۰۰ برای تعداد توکن های یک عنصر به این دلیل بود که در فایل داده جمله های آن از هم جدا نشده بود بنابراین از آن چشم‏پوشی شد)

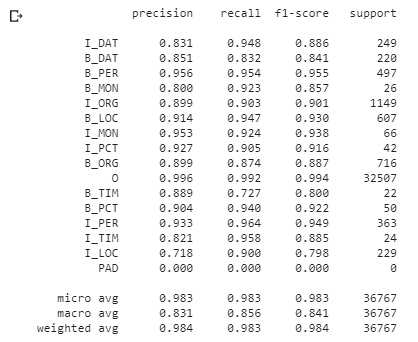
سپس بعد از انجام دادن attention mask و چشم پوشی از مقادیر صفر حاصل از padding برای آن، و عملیات تقسیم داده، و تبدیل داده ها به tensor، حال ورودی مدل آماده است.

مدل pretrained ، در دو ایپاک train شد و نتایج حاصل از loss validation همانطور که در لینک کلب هم آمده به صورت زیر است:



هم چنین مدل به دست آمده را بر روی داده های تست نیز بررسی کردیم که در لینک کلب در بخش Test model آمده است و نتایج به صورت زیر می‏باشد:

در حالتی که بین اینکه تگ اول یا در ادامه entity آمده تفاوت قائل شدیم؛



در حالتی که تگ را معیار قرار دادیم نه اینکه اول آمده یا در ادامه‏ی entity:

