

**دانشکدگان فنی دانشگاه تهران**

**نام افراد گروه:**

**محمدباربد امیرمزلقانی - 810102348**

**پروژه**

**هوش مصنوعی در سیستم های نهفته**

بهمن 1402

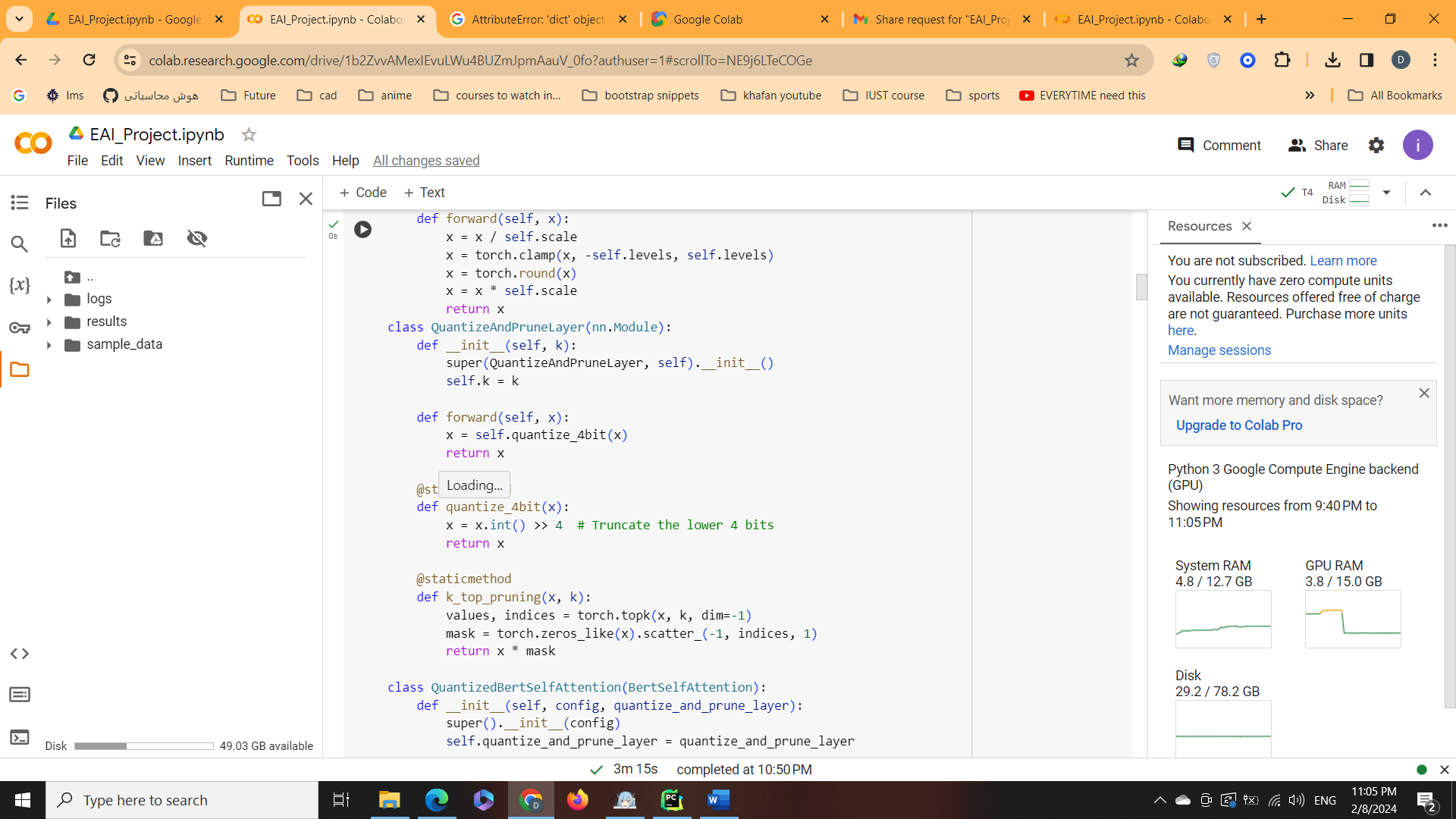
فهرست مطالب

[مرحله اول 2](#_Toc148982102)

# مرحله اول

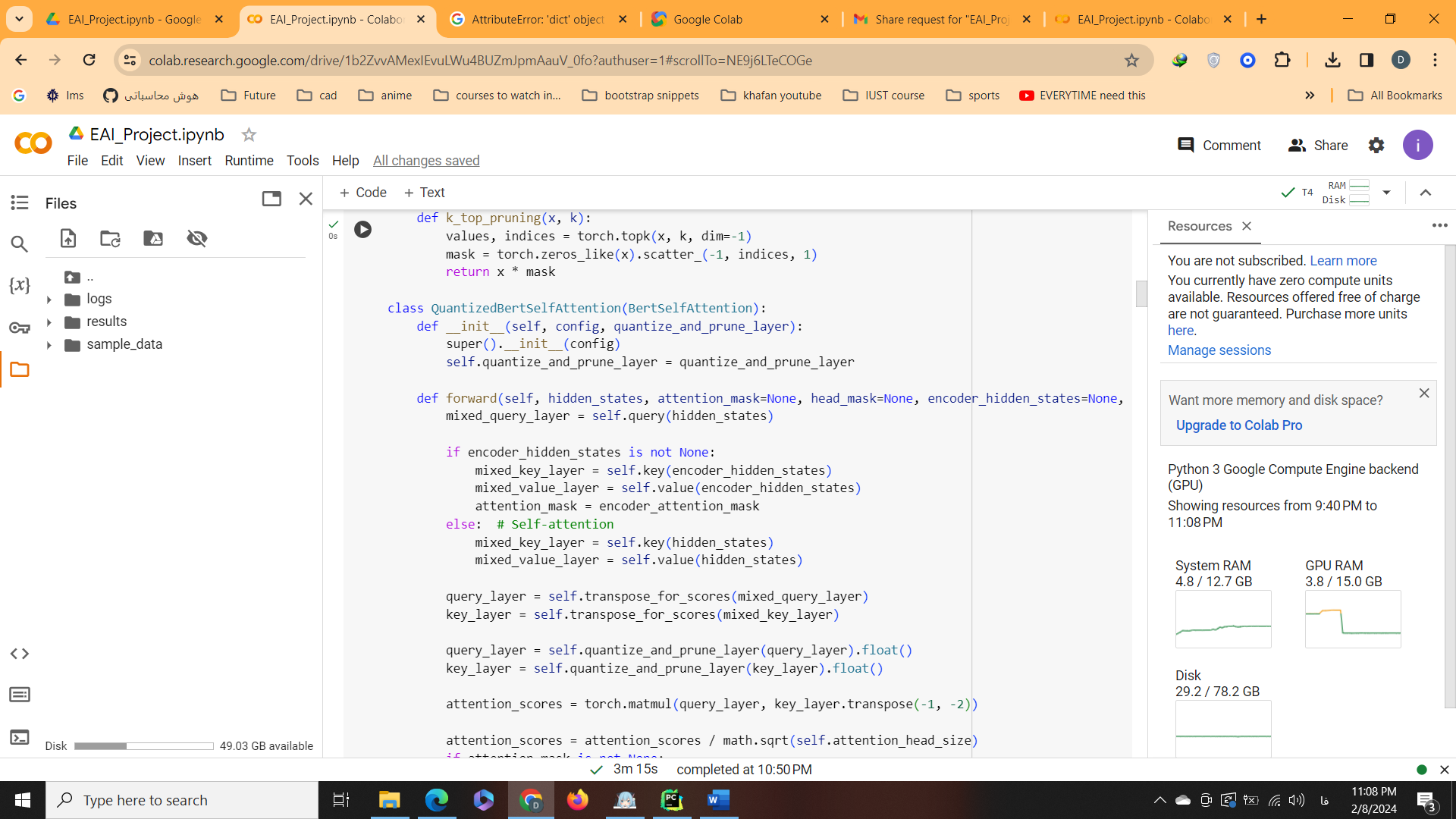
قسمت 1:

با توجه به خواسته سوال برای prune و quantize کردن ابتدا یک کلاس برای این کار تعریف میکنیم که با دریافت یک لایه یا آن را کوانتایز میکند یا به کمک الگوریتم top-k پرون میکند :

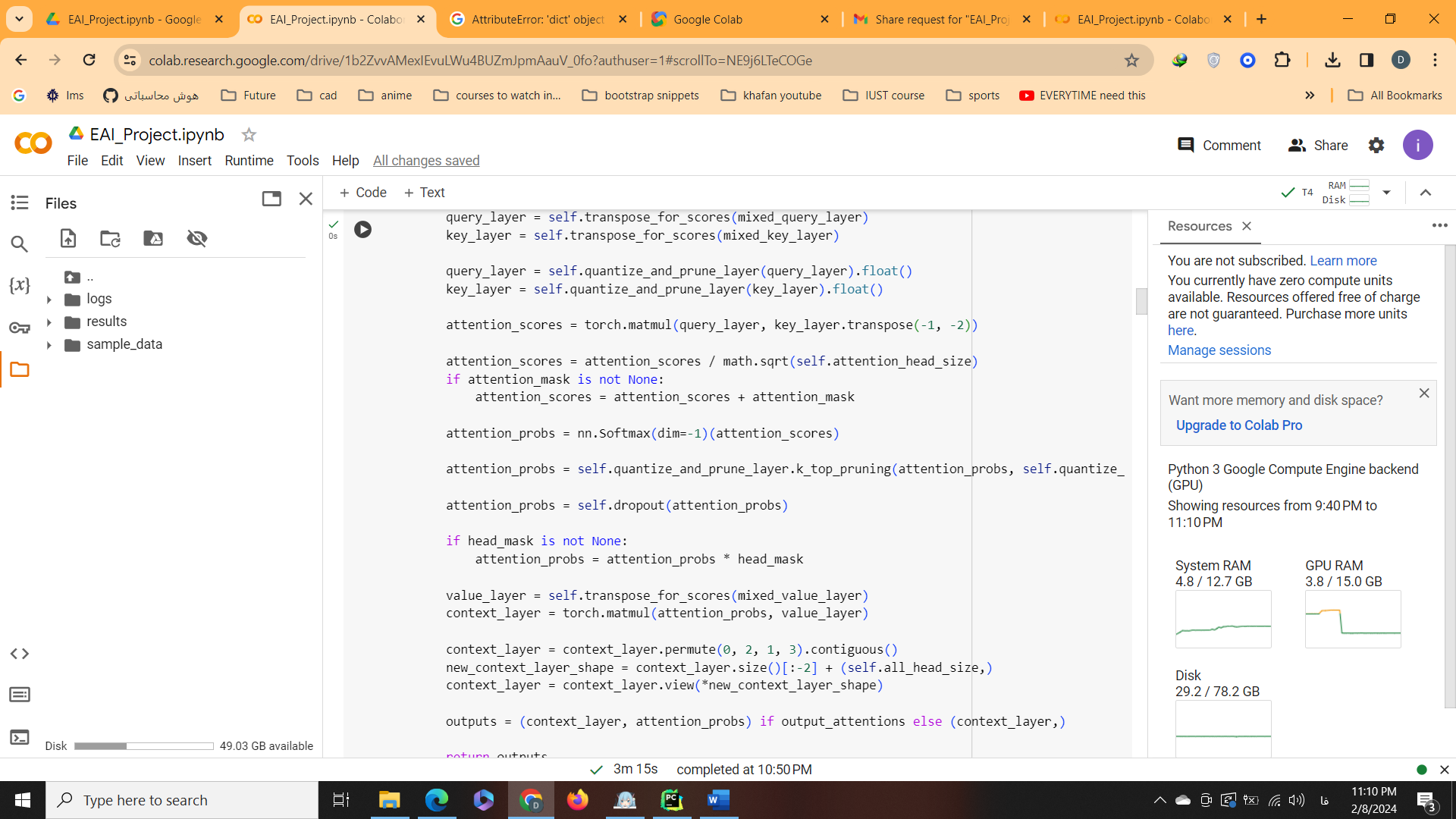


در ادامه مکانیزم self-attention را با توجه به روش هرس 1 انتخاب میکنیم و آن را چند مرحله تغییر میدهیم .

ابتدا query, key tensor را کوانتیزه میکنیم که در ادامه میبینید :

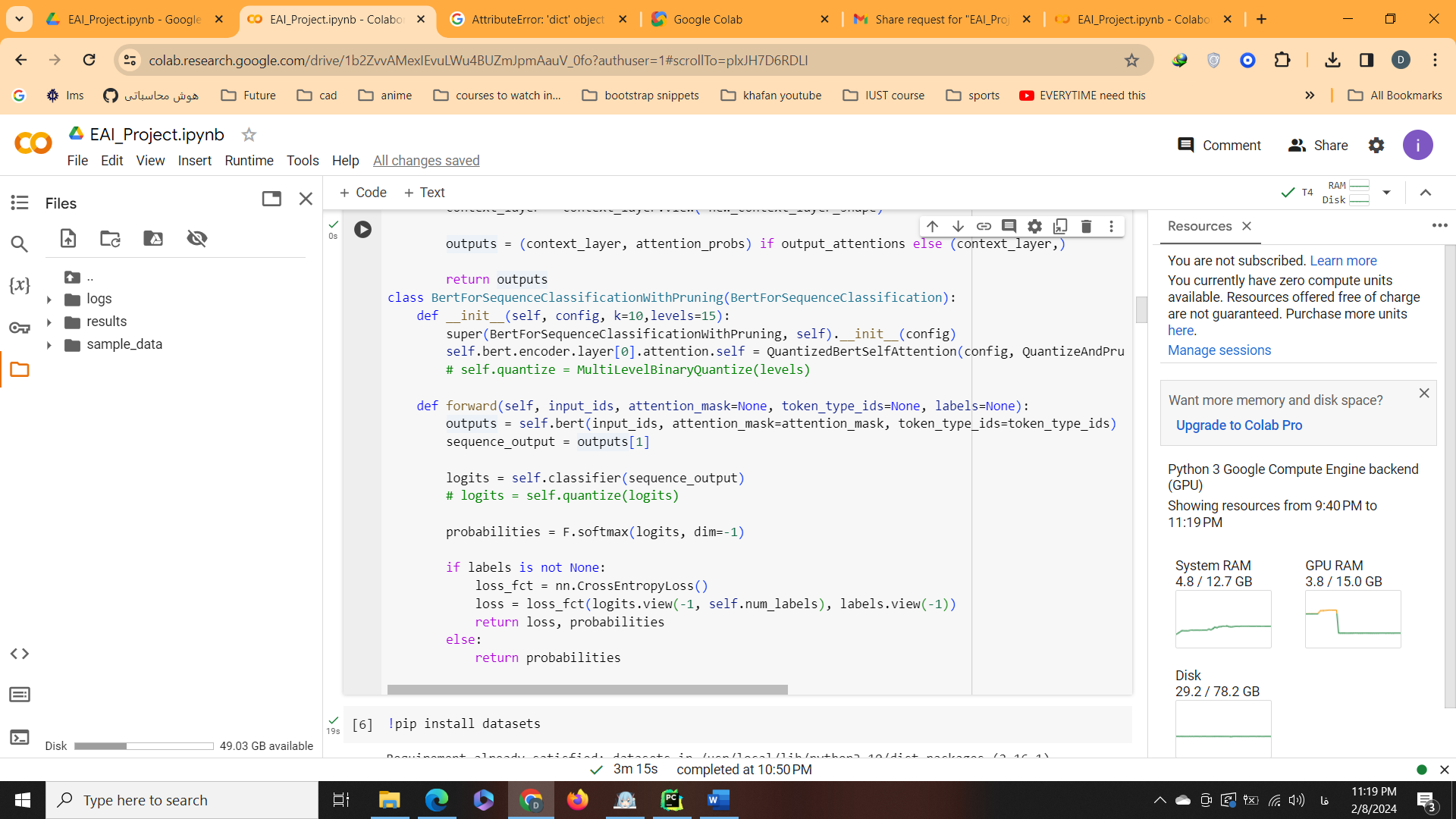


سپس عملیات top-k را طبق خواسته سوال و همچنین طبق رابطه پس از softmax اعمال میکنیم :

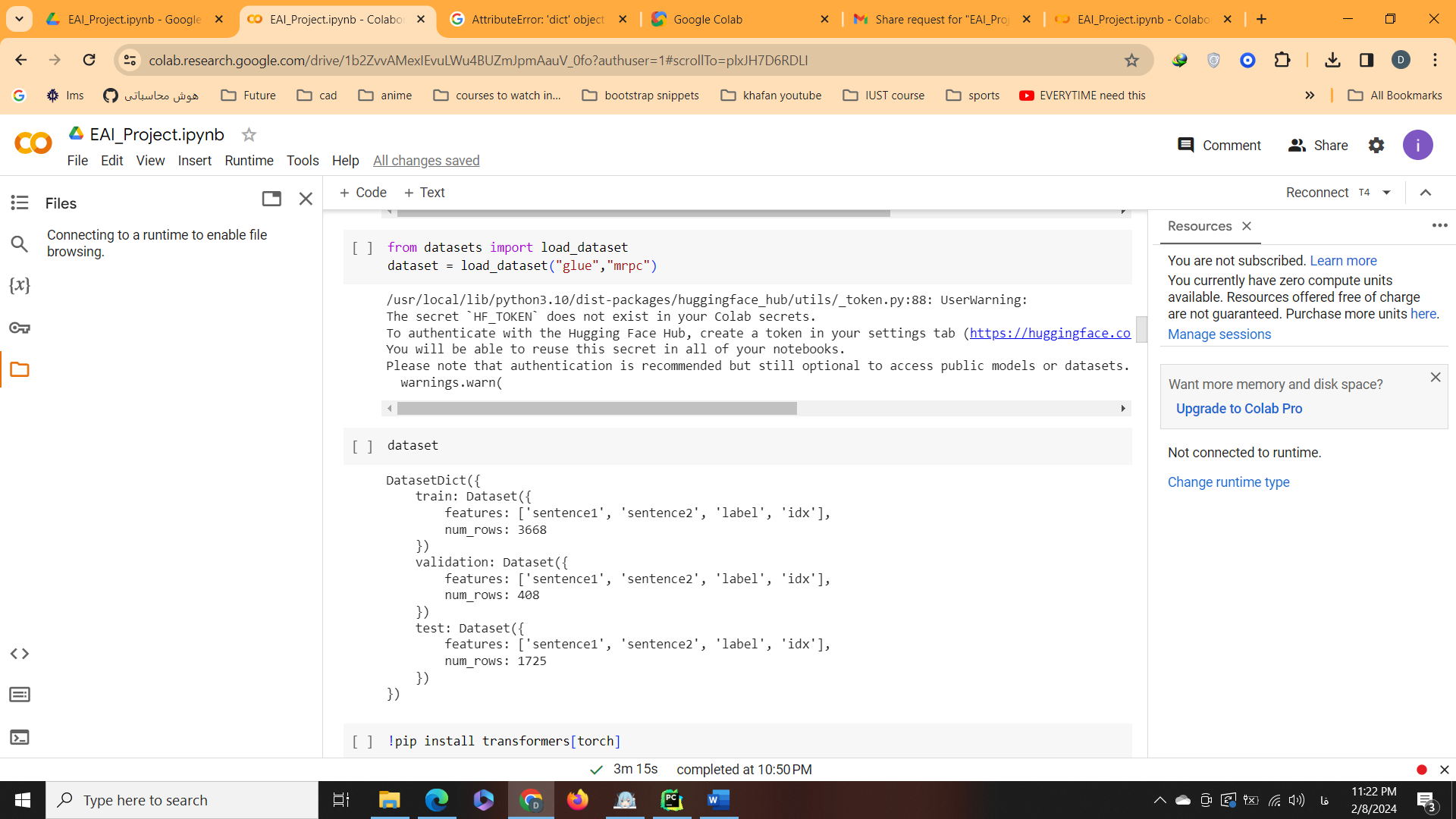


در گام بعد به سراغ تعریف کلاس مدل میرویم که بیس کار همان مدل bert ساده است اما تغییراتی هم دارد مثلا همین مکانیزم جدیدی که در بالا تعریف کردیم ،

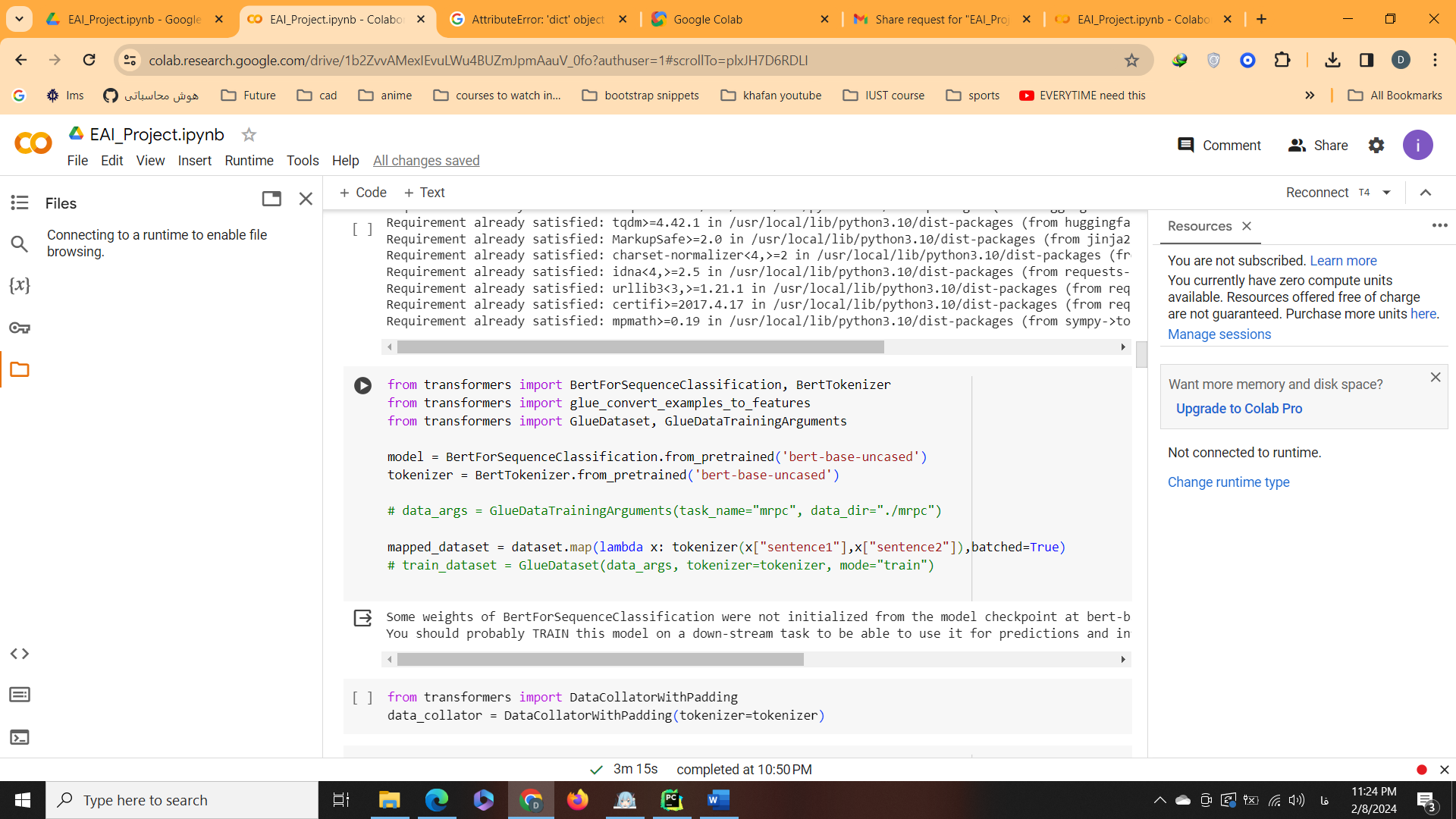
این کد یک کلاس برای دسته بندی دنباله با استفاده از مدل BERT را تعریف می کند که قابلیت هرس را نیز دارد. در متد forward، خروجی های مدل BERT برای داده های ورودی محاسبه می شود و سپس از آن برای محاسبه احتمالات استفاده می شود. اگر برچسب ها نیز وجود داشته باشد، خطا نیز محاسبه می شود.کد این قسمت را در ادامه مشاهده میکنیم :

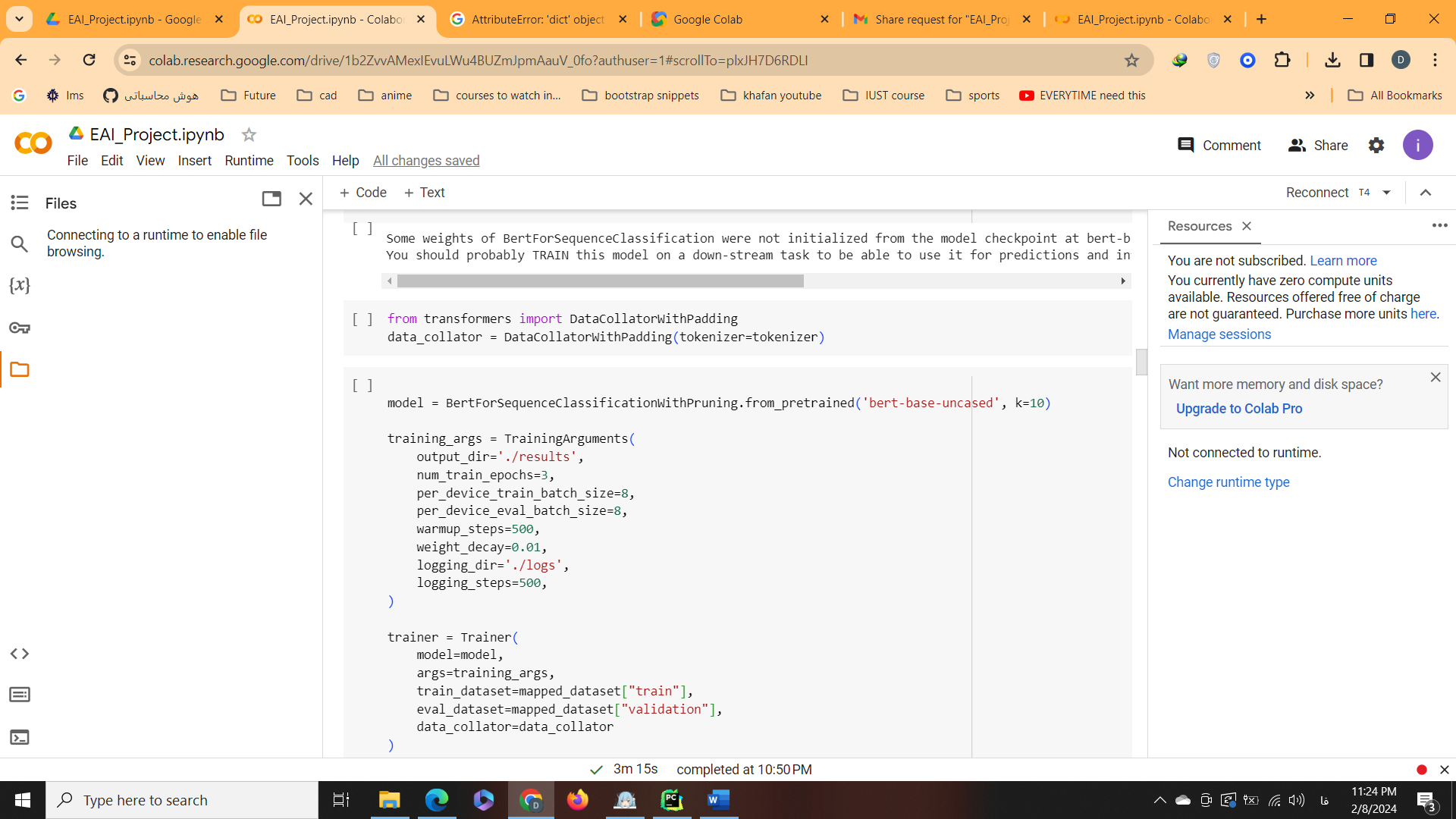


در ادامه برای ترین نیاز به دیتاست داریم که با توجه به متن پروژه یک تسک از مجموعه glue انتخاب میکنیم، در اینجا mrpc را انتخاب کردیم که در ادامه نحوه لود و ساختار آن را مشاهده میکنیم .

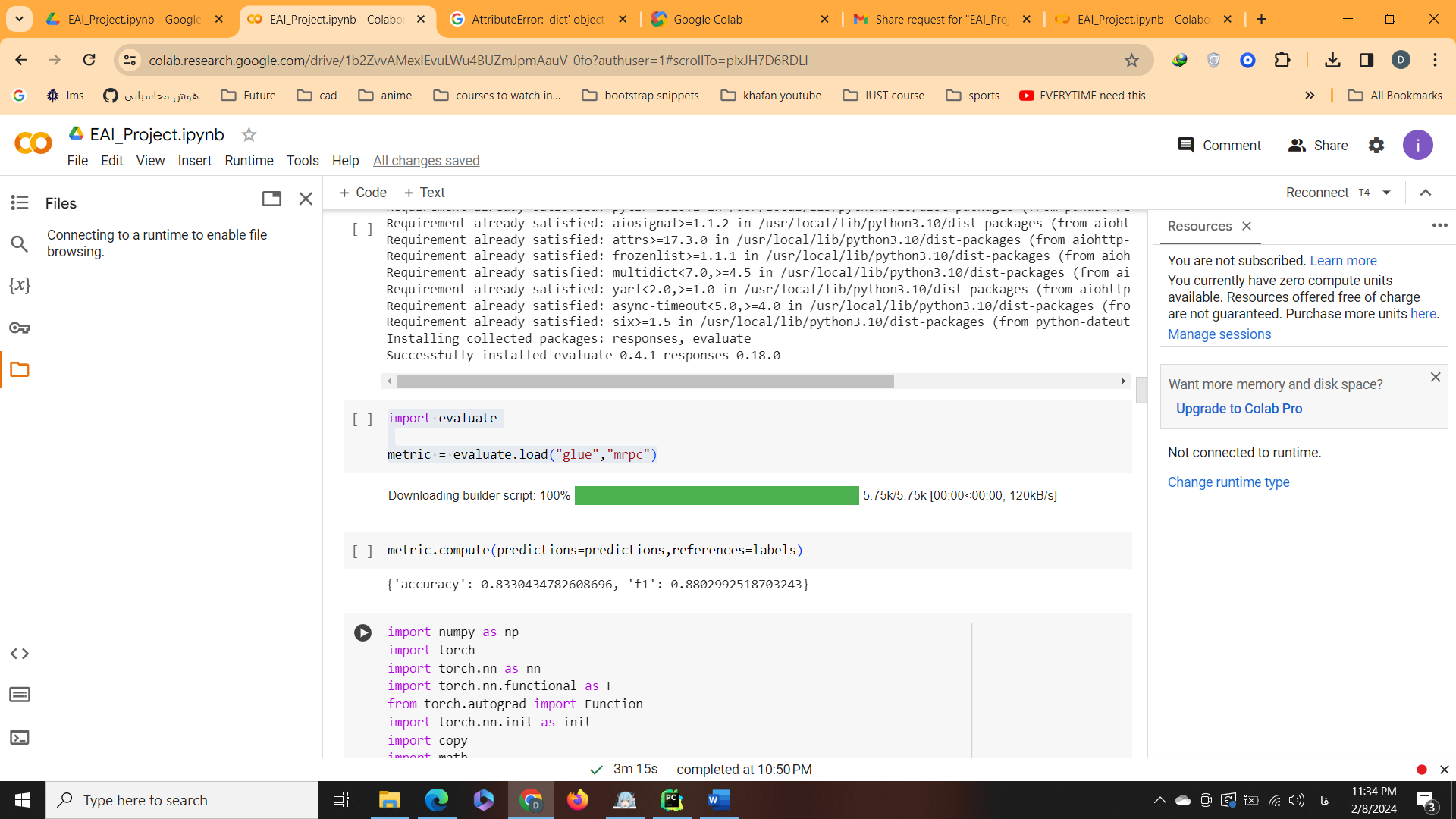


در ادامه داده ها را توکنایز میکنیم و پارامتر های ترین را کانفیگ میکنیم :

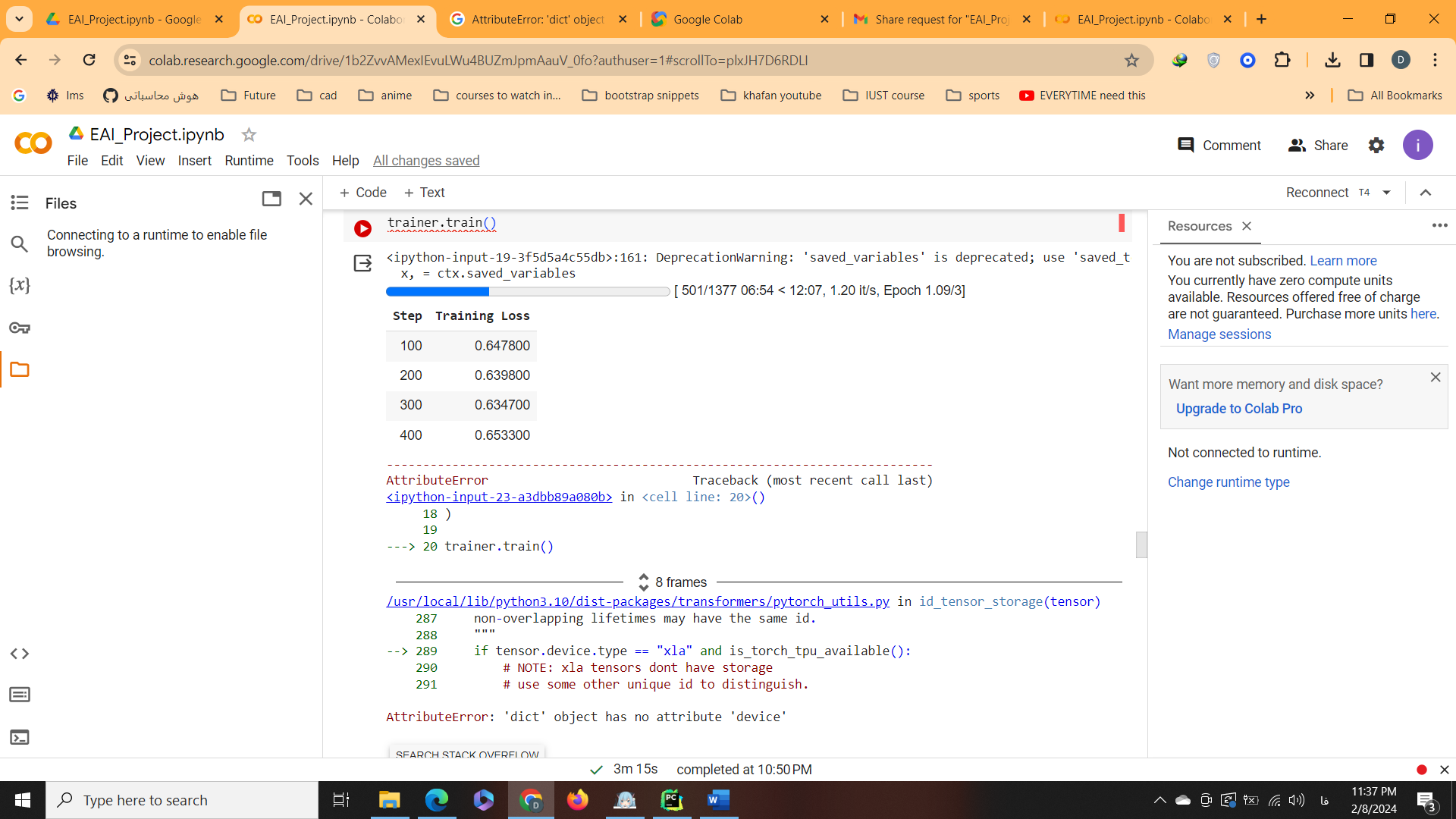


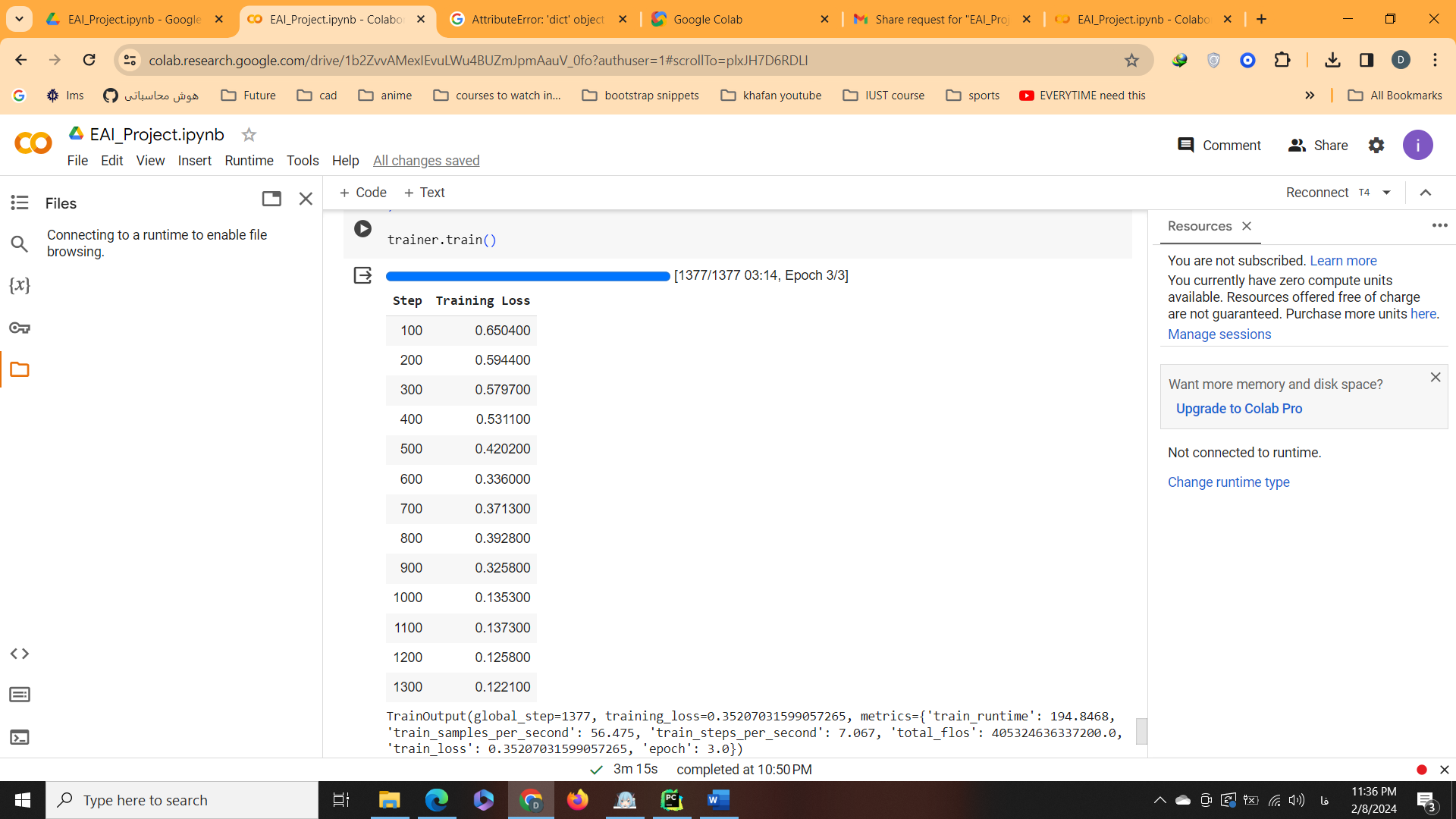


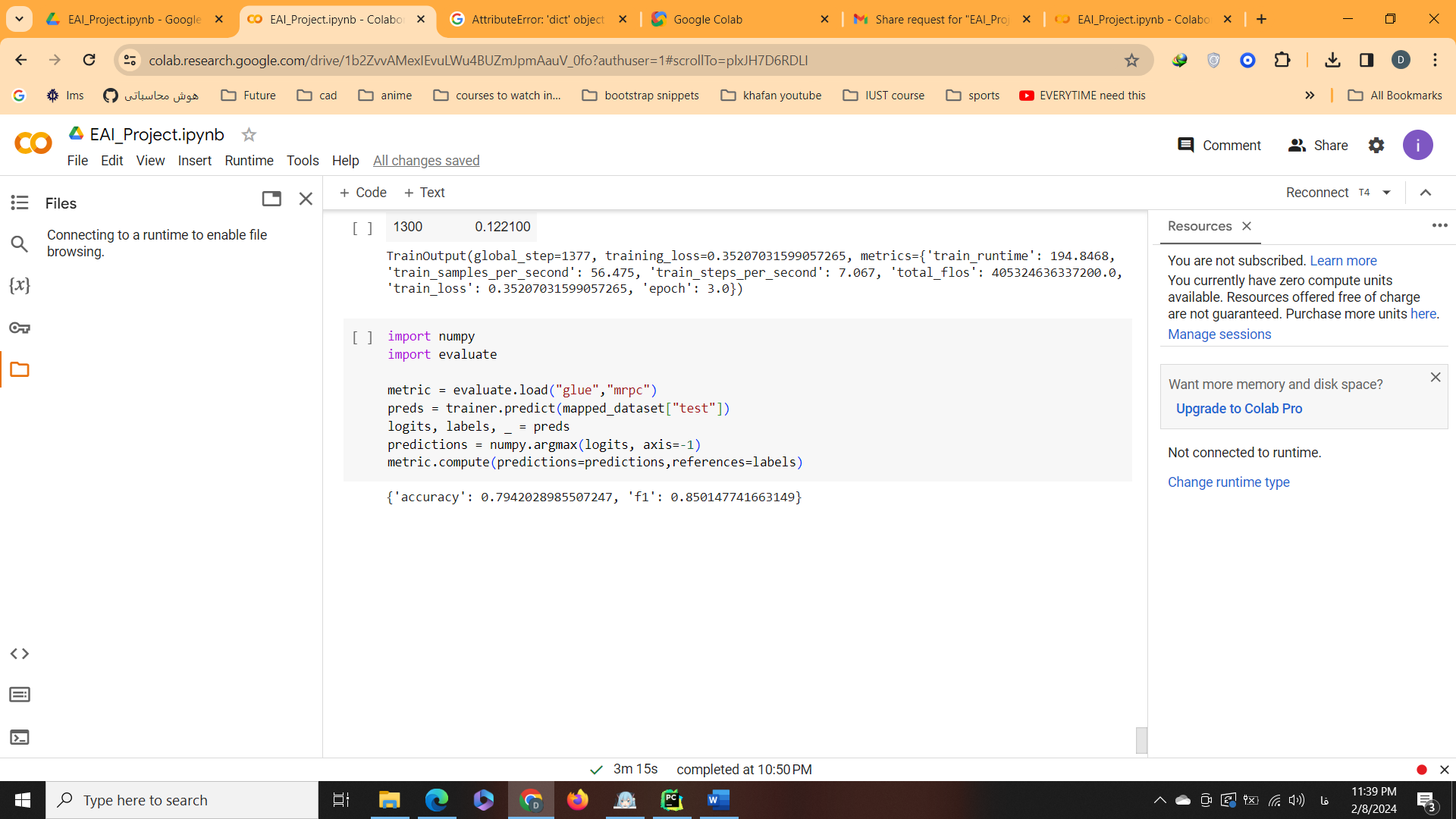
در نهایت برای سنجش مدل از کتابخانه evaluate استفاده میکنیم



و به عنوان گام نهایی از تابع prepare که در اختیارمان گذاشتید استفاده میکنیم وقتی از rebnet استفاده کردیم ارور عجیبی خورد که در گام 500 به این برخورد میکنیم با راهنمایی کمک مدرس به سراغ lsq رفتیم و نتیجه هم گرفتیم :







و مدل را به ازای پارامتر های مختلف تست کردیم مثلا چند بیتی بودن و k-top 5 10 و 15

