تمرین چهارم درس پردازش زبان‌های طبیعی

گردآورندگان:

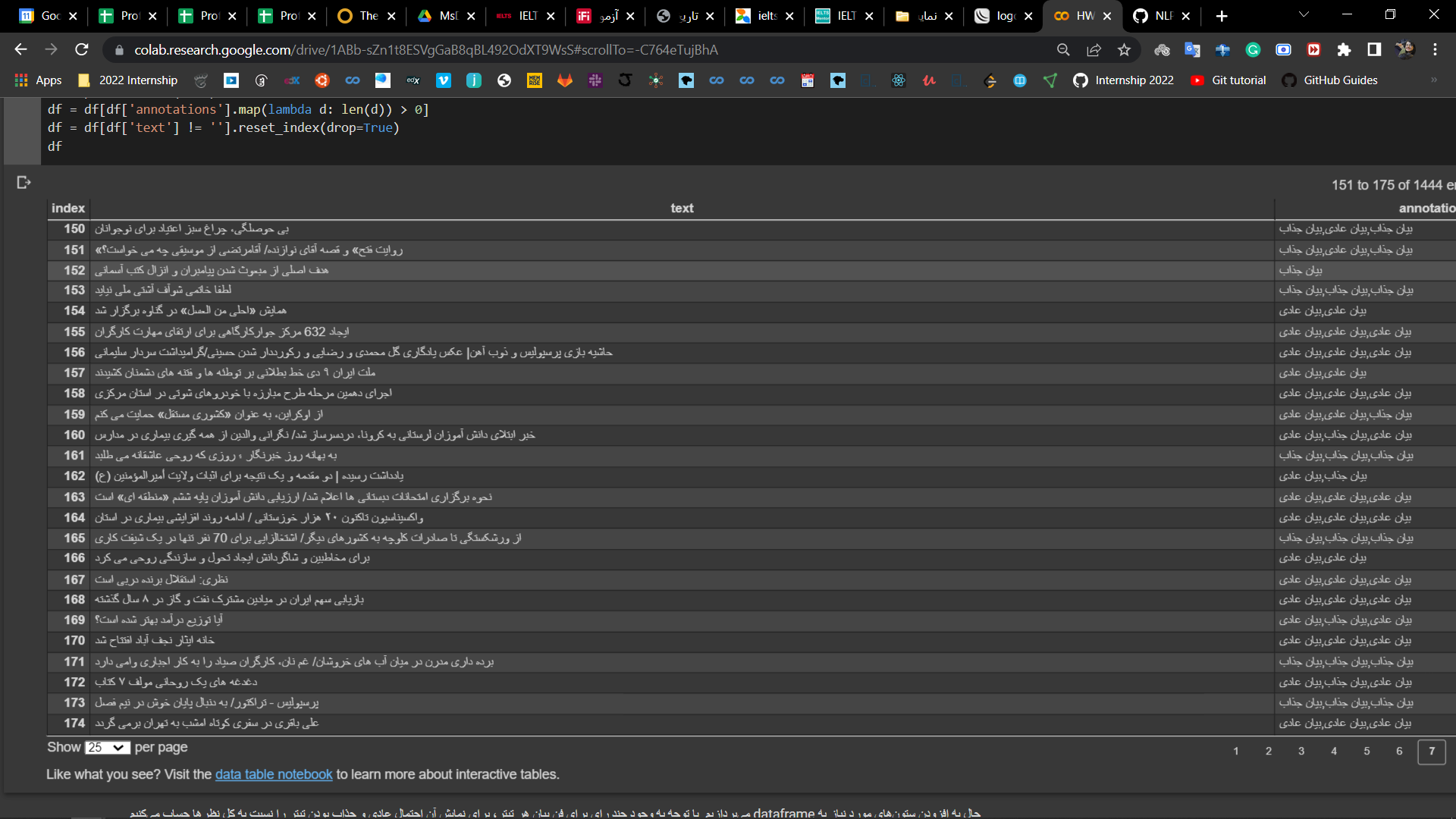
ساحل مس‌فروش

سروش تابش

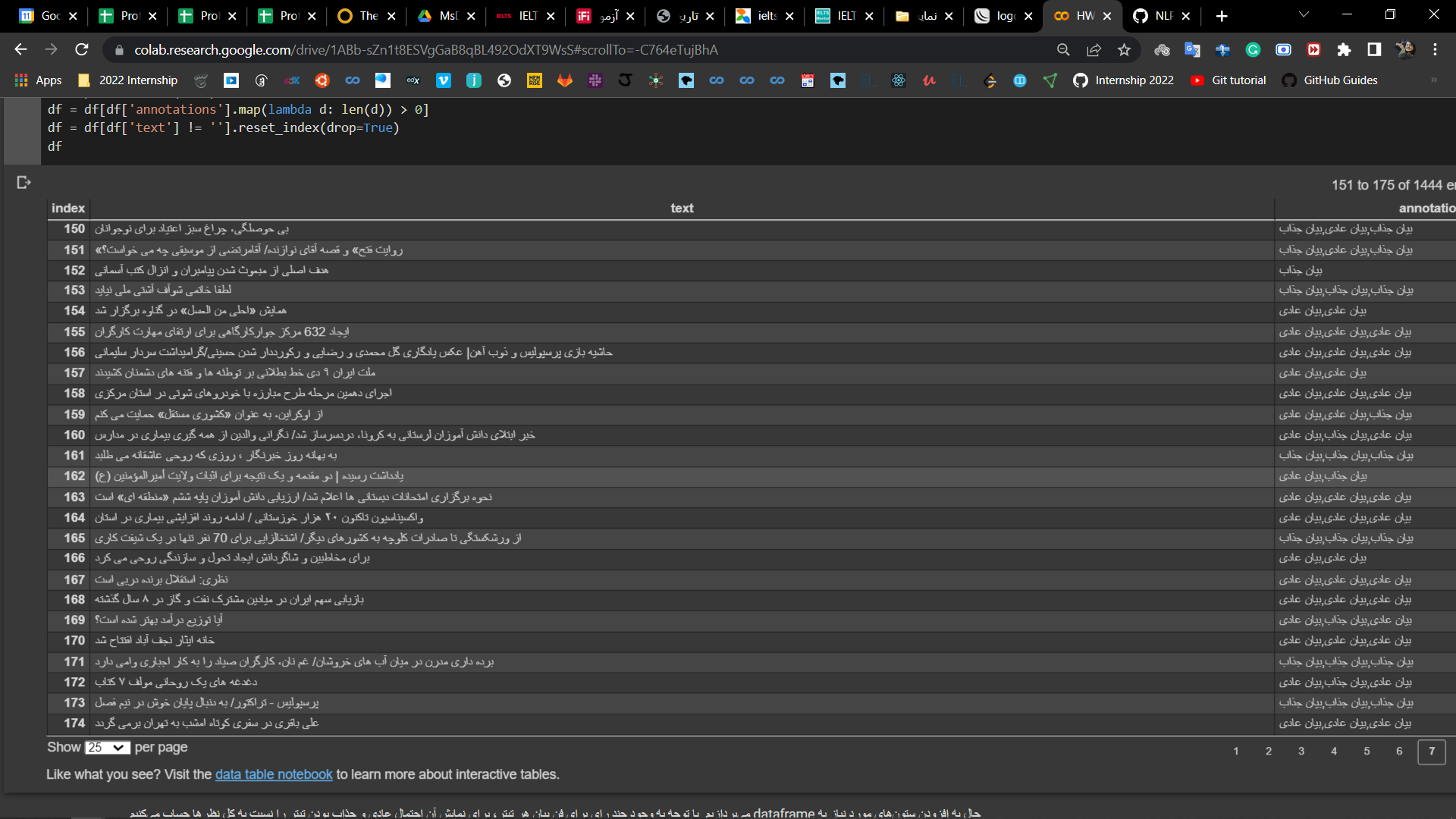
درنا دهقانی

ما در این تمرین ترک "فن بیان تیتر" را انتخاب کردیم. در این تمرین قصد داریم به کمک دو روش، متون را بر اساس جذاب بودن یا نبودن تیتر آنها طبقه‌بندی کنیم. در روش اول از یک ترنسفورمر از پیش آموزش دیده و در روش دوم از CNN استفاده می‌کنیم.

لازم به ذکر است که دیتای داده شده از چندین نظر مناسب نیست. یک دلیل آن کم بودن دیتاست. دلیل دیگر، عدم انتخاب فن بیان مناسب برای تیترهاست. مثلاً در مورد زیر بیان تیتر کاملا عادیست اما لیبل جذاب دارد.

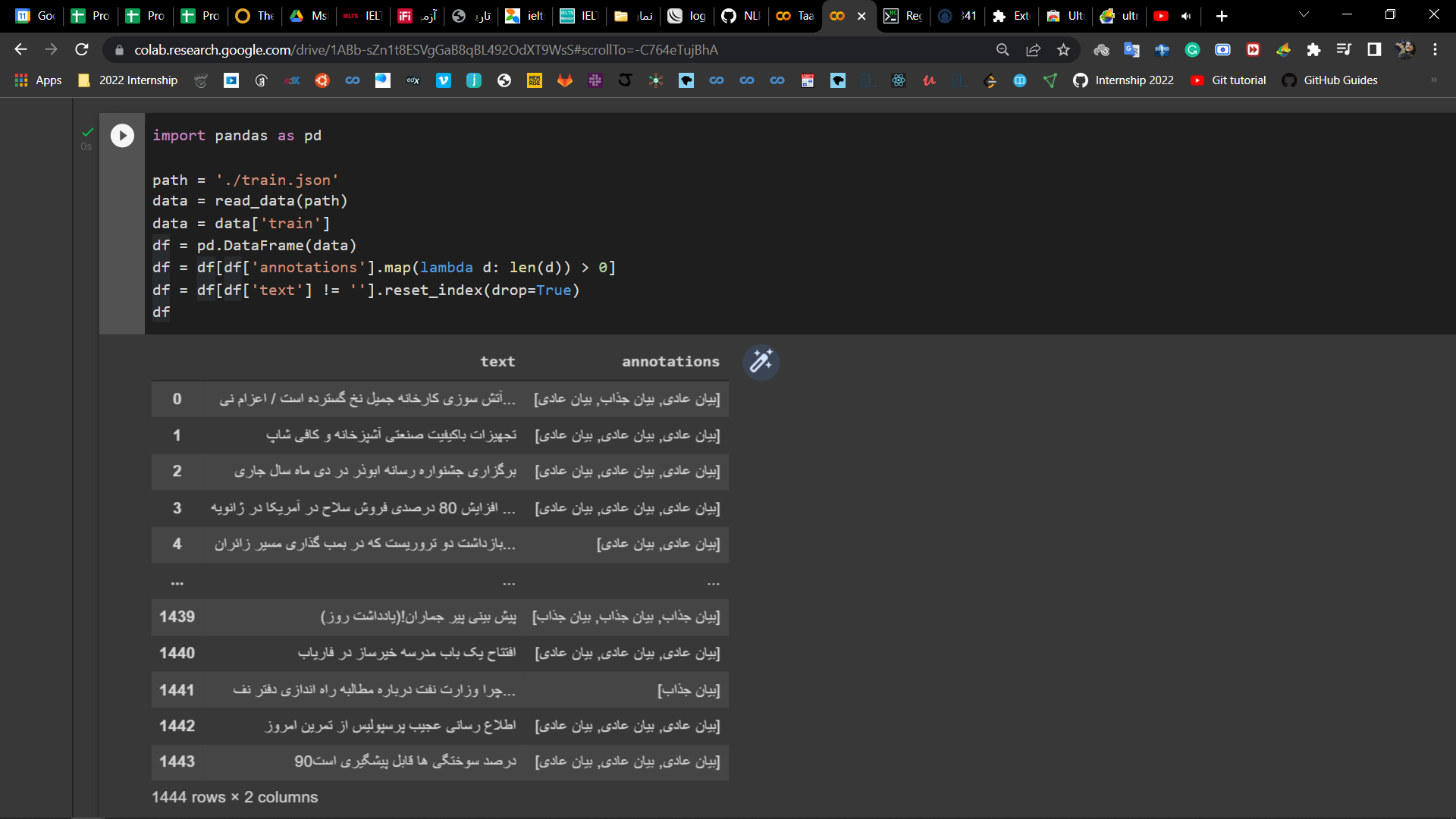


یا در این مورد، احتمال هر دو لیبل برای تیتر وجود دارد و حتی با vote گرفتن نمی‌توان در مورد آن تصمیم قطعی گرفت.

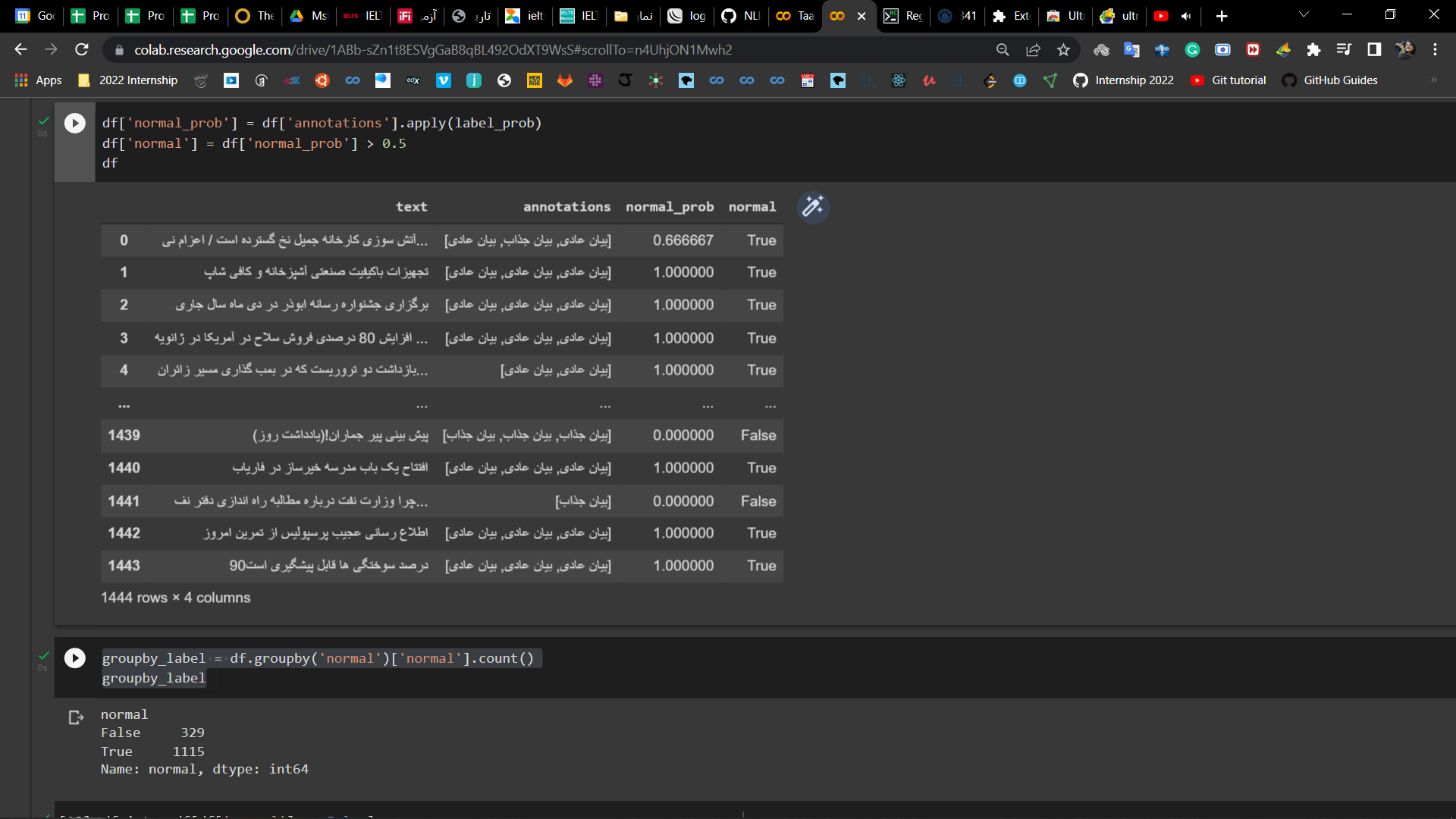


ابتدا دیتای داده شده را بر اساس یادگیری، ارزیابی و تست تقسیم می‌کنیم. با توجه به اینکه این ترک بر اساس تیتر عمل می‌کند و متن خبر تاثیری ندارد، تنها این بخش از دیتا و بیان تیتر آن را به کمک توابع read\_data، extract\_title و rewrite\_data بازنویسی می‌کنیم.

برای مشاهده‌ی ساختار داده‌ها، داده‌ی یادگیری را از فایل json می‌خوانیم و به dataframe تبدیل می‌کنیم. هم‌چنین اخباری که تیتر یا فن بیان نامشخص دارند را حذف می‌کنیم.



حال به افزودن ستون‌های مورد نیاز به dataframe می‌پردازیم. با توجه به وجود چند رای برای فن بیان هر تیتر، برای نمایش آن به کمک تابع label\_prob احتمال عادی و جذاب بودن تیتر را نسبت به کل نظرها حساب می‌کنیم.



اگر این احتمال از 0.5 کمتر بود، بیان را عادی و در غیر اینصورت جذاب فرض می‌کنیم. همچنین در صورت برابر بودن با 0.5، با اینکه هر دو لیبل محتملند، برای ایجاد تعادل در دیتاست چنین تیتری را جذاب در نظر می‌گیریم، چرا که تعداد داده‌های جذاب کمترند.

**متعادل کردن دیتاست**

در داده‌ی train تعداد تیترها با فن بیان عادی برابر با 1115 و تعداد تیترها با فن بیان جذاب برابر با 329 است که تفاوت زیاد این دو مقدار موجب تمایل بیشتر مدل به لیبل با تکرار بیشتر می‌شود. برای جلوگیری از این مسئله، با تکرار برخی داده‌های جذاب و حذف برخی داده‌های عادی بصورت تصادفی، این مقادیر را به هم نزدیک می‌کنیم.

سپس به کمک تابع cleaning به تمیز کردن تیتر اخبار می‌پردازیم. در این بخش ابتدا آن را نرمال‌سازی کرده، سپس علائم و اعداد اضافه آن را پاک می‌کنیم. نهایتاً الگوهایی مانند emojiها و symbolها نیز از آن حذف می‌گردد.

نهایتاً تابع token\_title به tokenize کردن تیتر می‌پردازد. برای این کار از tokenizer parsBert استفاده کردیم.

از توابع نوشته شده برای اضافه کردن ستون‌های موردنیاز و باز نویسی آن استفاده می‌کنیم.