پردازش زبان‌های طبیعی – تمرین سوم

گزارش کار – مهرشاد فلاح اسطلخ‌زیر

401521462

سوال اول:

ابتدا از Tokenizer ’bert_base_cased‘ استفاده می‌کنم و کلمات را Tokenize می‌کنم. بعد از مدل Pretained برای دسته‌بندی عبارات استفاده می‌کنم. (AutoModelForSequenceClassification) و ’bert_base_cased‘. در مرحله بعد آرگومان‌های ترین را وارد می‌کنم مثل تعداد گام و بَچ سایز و ... ابتدا بر حسب متریک accuracy داده را آموزش می‌دهم. به دقت 94 در سه گام می‌رسد مدل. در مرحله بعد بر روی دیتاست تست پیش‌بینی انجام می‌دهم و با متریک f1 و Recall مقایسه می‌کنم نتایج را که در هر دو بالای نود درصد است که نشان می‌دهد مدل overfit (بیش‌برازش) نشده است.

سوال دوم:

برای حل مسئله word analogy ابتدا فاصله کسینوسی را محاسبه باید کرد که صرفا با استفاده از کتابخانه‌های numpy مثل dot برای ضرب نقطه‌ای و np.linalg.norm برای محاسبه نرم قابل حل است. در مرحله بعد با تابع complete_analogy سعی می‌کنیم فاصله بین a-b و ? - c که یک کلمه از word_to_vec_map است را محاسبه کنم و بیشترین امتیاز را پیدا کرده و به عنوان خروجی بدهم. مشاهده می‌شود به جز مورد اول که کلمه درست berlin بود بقیه کلمات درست پیش‌بینی شده. برای بخش امتیازی هم وظیفه جلوگیری از سوگیری (مثلا جنسیتی) کلمات را انجام می‌دهم. برای این کار بردار بایاس را پیدا کرده و بردار اولیه را بر روی آن پروجکت می‌کنم و بردار اولیه را از بردار پروجکت شده کم می‌کنم تا بردار خنثی‌شده بدست بیاید. تابع بعدی که equalize است وظیفه این را دارد که یک جفت کلمه بگیرد و طبق فرمول‌های درون سوال آن‌ها را debias کند و سوگیری را از بین ببرد. در خروجی من هم قابل مشاهده است که مقادیر debias شده اما عدد نهایی کمی متفاوت است که شاید از مدل باشد.