به نام خدا دانیال ریاضتی – 98243029

گزارش پروژه پایانی دوم درس هوش مصنوعی:

در این پروژه از ما خواسته شده تا با گرفتن یک لیست از عبارات انگلیسی، عبارت معادل آن در فارسی در ستون مقابل آن نمایش داده شود که این عبارت فارسی مطابق با تلفظ درست آن عبارت میبایست تولید شود.

ابتدا یک مرحله از این پروژه برای گردآوری دیتاست مناسب که بعد از آن برای استخراج تلفظ صحیح استفاده میشود، است که یک فایل **csv** که حاوی **10000** لغت پایه انگلیسی و حروف اضافه از سایت دانشگاه **MIT** که در لینک زیر است را به تلفظ صحیح آن تبدیل کرده ام.

https://www.mit.edu/~ecprice/wordlist.10000

در تبدیل به تلفظ از سایت https://easypronunciation.com/ و در مواردی از لایبرری **eng_to_ipa** که کد آن نیز در گیتهاب مشاهده نمودم استفاده کرده ام.(سیستمی رول بیسد و با **query** زدن به دیتابیس دانشگاه **Carnegie-Mellon University**)

سپس با دریافت یک کلمه، به ازای هر عبارت ابتدا محتویات آن را از هم **split** کردم و با استفاده از لایبرری **nltk.tag** و **nltk.tokenize** که می آید محتویات عبارات را از هم جدا سازی کرده و یک تگ به ازای هر کلمه در آن میدهد تا نوع کلمه(اسم یا فعل یا صفت یا مصدر یا...) را مشخص کند. همچنین اگر کلمه ای حاوی **prefix** بود، بخش اضلی را جدا کرده و جدا جدا فارسی سازی میشود تا دقت تلفظات حفظ شود.

از یک لایبرری دیگر نیز به اسم **nltk.stem.snowball** هم استفاده کردم تا اگر کلمه به فرمت خاص مثلا گذشته یا **ing** فرم بود بیاید آن را به ریشه اصلی کلمه برگرداند تا فارسی سازی آن ممکن شود.

اگر هم کلمه یک اسم خاص بود هم از فارسی سازی آن معمولی و با مپ کردن حروف به یک دیکشنری که تلفظ ها و گویش فارسی آن ها را در آن ست کرده ام صورت میگیرد.(نحوه تعیین گویش ها به صورتی است که برای حروف صدادار موقعیت حرف صدا دار و حروف بعدی آن و این که ایا ابتدا یا انتها یا میانه کلمه است بسیار اهمیت دارد که برای حالات خاص صدا گذاری حروف صدا دار با توجه به اولویت های گویشی که همه آن ها و حالات خاص گویش را در سایت https://www.englishlanguageclub.co.uk/ مشاهده کرده ام عمل کرده ام) همچنین اگر ابتدای کلمات با فونتیک های خاص مثلا ə شروع شود با یکسریی رول ها چک میکنم که در ابتدا فارسی آن حرف ا یا آ قرار دهد.

در انتها نیز پس از فارسی سازی شدن تک بخش های کوچک ان ها را با استفاده از لایبرری **Hazm** بخش ها را کنار هم قرار داده ام و علایم نگارشی متناظر را کم یا افزایش داده ام.

در پایان نیز آن ها را در لیست **csv** نهایی خود متناظرا قرار داده ام.