

## به نام خدا

امیرمحمد آزادی

NLP مرحله اول پروژه درس

موضوع انتخابی برای پروژه، دسته بندی متن بر اساس جنسیت است. در این پروژه از زبان python برای جمع آوری اطلاعات و تحلیل و آموزش مدل استفاده میشود. داده ها به زبان انگلیسی و محاوره ای است. داده ها از سطح توییتر جمع آوری میشوند. روش های مختلفی برای استخراج دیتا از توییتر وجود دارد که از جمله آنها میتوان به استفاده از کتابخانه Snsccape و یا Tweepy اشاره کرد. هر چند که مشکلاتی برای استفاده از این کتابخانه ها وجود دارد. به عنوان مثال با استفاده از کتابخانه Tweepy تنها می شود اطلاعات مربوط به یک هفته گذشته را استخراج کرد ولی میتوان از تمام امکانات API توییتر استفاده کرد. در حالیکه با استفاده از Snsccape میتوان به اطلاعات قدیمی تر دسترسی داشت ولی امکانات کمتری در اختیار ما قرار میدهد.