باسمه تعالی

گزارش تمرین شماره ۳ درس پردازش زبان طبیعی

استاد درس: جناب آقای دکتر باباعلی

نام دانشجو : ایمان کیانیان

شماره دانشجویی: ۶۱۰۳۰۰۲۰۳

۱- مقدمه

این تمرین، ادامه ی تمرین گذشته است. دیتاستی شامل ۲۵۹۷۹۴ کلمه آموزش و ۲۵۹۷۹۴ کلمه تست به ما داده شده است. تمام این دو دیتاست شامل tag هستند. میخواهیم با استفاده از کلمات و tag متناظر در مجموعه آموزشی یک HMM را آموزش دهیم و با استفاده از الگوریتم ویتربی، برای کلمات موجود در مجموعه تست، tag گذاری انجام دهیم. سعی بر این داریم تا tag هایی که میگذاریم به تگ اصلی در داده آموزشی شبیه باشد.

۲- پیش پردازش داده ها

ابتدا یک جدول فراوانی به ازای هر یک از کلمات در داده آموزشی ساختیم و کلماتی که فقط یکبار در train آمده بودند با UNK جایگزین کردیم. سپس در داده تست کلماتی که در واژگان داده آموزشی نبودند نیز UNK کردیم. همچنین vocab و tags را به صورت یکتا از روی داده های آموزشی بدست آوردیم. پیش پردازش خاص دیگری روی داده ها انجام نشد.

۳- آموزش مدل HMM

با استفاده از داده های آموزشی ، ماتریس A و B را محاسبه کردیم. آموزش این دو ماتریس را با روش شمارش تعداد کلمات انجام دادیم تا سرعت بالاتری را دریافت کنیم.

۴- الگوریتم ویتربی

الگوریتم ویتربی را مشابه با الگوریتمی که در توضیحات درس بود پیاده سازی کردیم. در این الگوریتم به دنبال، دنباله ی حالتی هستیم که مشاهدات تست ما را تولید میکنند. در واقع یکی از ورودی های این الگوریتم داده های تست بدون تگ هستند که وظیفه این الگوریتم پیدا کردن تگ های مناسب با توجه به ماتریس های A و B است.

۵- دقت دریافت شده

بعد از اجرای الگوریتم ویتربی بر روی داده های تست، با استفاده از تابع accuracy_score که برای عمل classification به کار میرود، مقایسه ای بین تگ های واقعی داده های تست و تگ های بدست آمده پس از اجرای الگوریتم ویتربی انجام دادیم و نتیجه دقت **۰,۹۴۰۳۰۲۷۰۱۳۷۱۰۸۶۳** را به ما داد. دقت **۹۴** درصد برای این تعداد کلمات میتواند بسیار مناسب باشد.