تمرین دوم پردازش زبان طبیعی

توجه:

- **کد باید فقط در زبان پایتون باشد.**
- **استفاده از کتابخانه‌های از پیش آماده مجاز است. در گزارش کتابخانه ها ذکر شوند.**
- **لطفاً علاوه بر ضمیمه کردن کد، نتایج را تحلیل و در فایل گزارش خود ضمیمه کنید.**
- **فایل گزارش به فرمت pdf و به زبان فارسی باشد.**

در این سوال با استفاده از الگوریتم ویتربی و مدل های مخفی مارکوف مسئله برچسب گذاری کلام (POS) را حل خواهید کرد. مجموعه دادگان به سه دسته آموزشی، اعتبار سنجی و آزمون تقسیم شده اند و در اختیار شما قرار گرفته اند. در هر خط دادگان کلمه و جز کلام مربوطه به آن آمده است و همچنین جملات از طریق "?.!؟؛:" از یکدیگر جدا شده اند؛ اما دادگان آزمون فاقد برچسب جز کلام است و در انتهای تمرین باید پیشبینی مدل بهینه خود را بر روی این دادگان ضمیمه کنید.

۱. در ابتدا جملات را استخراج (آموزش الگوریتم بر روی جملات است) کنید، و سپس کلمات را به همراه جز کلام شان در جملات مشخص کنید. تعداد برچسب های منحصر به فرد جزیی را گزارش دهید. تعداد کلمات منحصر به فرد در هر دادگان را گزارش دهید، همچنین تعداد کلماتی در دادگان آزمون و اعتبار سنجی هستند ولی در دادگان آموزش نیستند را هم گزارش دهید.

۲. در این بخش پارامتر های لازم را محاسبه کنید. توجه داشته باشید با توجه به اینکه ممکن است در دادگان اعتبار سنجی و آزمون کلماتی وجود داشته باشند که در دادگان آموزشی وجود ندارند، باید از الگوریتم های هموار سازی در محاسبه احتمالات تابع توزیع استفاده شود. در این تمرین از الگوریتم Add-K استفاده باید شود.

۳. مدل را برای ۷ مقدار متفاوت K، $۱, ۱۰^{-۱}, ۱۰^{-۲}, ۱۰^{-۳}, ۱۰^{-۴}, ۱۰^{-۵}, ۱۰^{-۸}$ آموزش دهید و مقدار دقت را برای داده اعتبار سنجی و آموزشی گزارش دهید.

۴. آیا الگویی بر پایه مقدار K مشاهده می کنید. همچنین مدل ها را بر اساس بیش‌برازش و کم‌برازش بررسی کنید.

۵. مدل بهینه را انتخاب کنید و بر روی داده تست پیشبین را انجام دهید، و برچسب های پیشبینی شده در یک فایلی به نام "y_test.txt" قرار دهید* به گونه ای که برچسب هر نماد فایل آزمون در خط متناظر آن قرار گیرد (فایل "y_test_example.txt" به عنوان مثال بررسی کنید).