بسمه تعالی

دانشگاه صنعتی امیرکبیر

دانشکده مهندسی کامپیوتر و فناوری اطلاعات

پروژه سوم مبانی و کاربردهای هوش مصنوعی

توضیحات:

- مهلت پروژه تا تاریخ ۲۰ مرداد ۹۹ درنظر گرفته شده است.
- پروژه باید به صورت انفرادی انجام شود.
- در صورت مشاهده‌ی هرگونه شباهت در کد یا گزارش، نمره‌ی صفر برای کل پروژه لحاظ خواهد شد.
- تمیزی و خوانایی گزارش از اهمیت بالایی برخوردار است.
- لطفا گزارش پروژه را به همراه کد در فایلی با نام "Prj3_StudentNumber.zip" در سایت درس در مهلت معین بارگزاری نمایید.
- به ازای هر روز تاخیر ده درصد از نمره‌ی شما کسر خواهد شد.
- در صورت داشتن هرگونه سوال می‌توانید از طریق ایمیل Tabatabaeifateme@gmail.com با تدریس یار درس در ارتباط باشید.

تابستان ۹۹

در این پروژه قصد داریم با استفاده از مدل‌های زبانی n-gram، الگوریتمی پیاده‌سازی کنیم که بتواند کلمه‌ای که از یک جمله حذف شده‌است را به درستی حدس بزند که برای این کار نیاز به روش‌های پردازش متن داریم. برای پیاده‌سازی مطابق مراحل زیر عمل کنید.

۱. مجموعه داده را از این لینک دانلود نمایید. پانصد جمله‌ی اول از فایل "train_v2.txt" را به عنوان داده‌های آموزشی درنظر بگیرید. صد جمله تصادفی از فایل "test_v2.txt" را نیز (که شامل تعدادی جمله است که از هر جمله یک کلمه حذف شده است) برای ارزیابی مدل انتخاب کنید.

۲. مدل‌های زبانی را برای مجموعه داده‌ی آموزشی استخراج نمایید. به عنوان مثال، در حالت Unigram برای هر کلمه در جمله احتمالی خواهد بود که با استفاده از تعداد دفعات تکرار این کلمه محاسبه شده است. در حالت Bigram، این احتمالات برای زوج کلمات خواهد بود.

۳. از روش هموارسازی backoff برای 3-gram استفاده نمایید. برای پارامترهای این هموارسازی سه حالت مختلف را امتحان کنید و بهترین حالت را گزارش نمایید.

۴. برای تشخیص مناسب ترین کلمه برای جای خالی، باید براساس مدل n-gram استفاده شده، کلمه‌ای که بیشترین احتمال را دارد، انتخاب کنید.

۵. در انتها، برای مجموعه داده‌ی تست درنظر گرفته شده، معیار دقت (Accuracy) را گزارش کنید.

موفق باشید