بسم الله الرحمن الرحیم

تمرین دوم درس مباحث ویژه

نام استاد: حمیدرضا برادران کاشانی

طراح آموزشی: ریحانه سعیدی، مهدی دارونی

دانشکده مهندسی کامپیوتر

گروه هوش مصنوعی

توضیحات: مهلت تحویل این تمرین تا تاریخ **۱۴۰۱/۹/۵** در نظر گرفته شده است. پس از این تاریخ به مدت **۲ روز** یعنی تا تاریخ **۱۴۰۱/۹/۷** تمرینات با کسر **۳۰ درصد** از نمره تحویل گرفته می‌شود. لازم به ذکر است که به تمرینات تحویلی پس از تاریخ نمره‌ای تعلق نخواهد گرفت.

هدف آموزشی تمرین: در این تمرین قصد داریم با ساخت یک سیستم تکمیل خودکار آشنا شویم. سیستم تکمیل خودکار، سیستمی است که ممکن است هر روز ببینید، وقتی عبارتی را در گوگل سرچ می‌کنید و یا متن ایمیلی را آماده می‌کنید، اغلب پیشنهادهایی برای کمک به شما برای تکمیل عبارت‌ها نمایش داده می‌شود. شما در این تمرین یک نمونه از این سیستم‌ها را توسعه خواهید داد.

مراحل تمرین:

۱) **پیش‌پردازش:** فایل متنی را باز کنید. آن را به جملات سازنده‌اس تجزیه‌کنید. علائم نگارشی را حذف کنید، به‌طوریکه فقط اعداد و کلمات بمانند و هرگونه پیش‌پردازش که بتواند به بهبود نتایج خروجی کمک کند را اعمال کنید.

۲) **ساخت n-gramها**: از متن unigram، Bigram، Trigram و Quadrigram را با استفاده از کتابخانه NLTK استخراج کنید. تعداد تکرار هر n-gram را بیاید و ۵تا از پرتکرارترین n-gramها را گزارش کنید.

۳) **هموارسازی جملات:** در این قسمت تابعی برای هموارسازی جملات طراحی کنید. از Laplace smoothing برای unigrams و از Good-Turing smoothing برای دیگر N-gramها استفاده کنید. توضیح دهید چرا Laplace smoothing برای سایر n-gramها مناسب نیست.

۴) **پیش‌بینی کلمات:** حال می‌خواهیم با استفاده از مدل‌های ایجاد شده، کلمات جدید را با استفاده از دنباله‌ای از کلمات ایجاد شده پیش‌بینی کنیم. برای این کار تابعی طراحی کنید که مدل و دنباله کلمات را ورودی بگیرد و کلمه جدید را به عنوان خروجی برگرداند.

۵) **تولید جملات با طول معین:** یک تابع بنویسید که جملات با طول معین ایجاد کند. بدین منظور از تابع طراحی شده در قسمت قبل استفاده کنید. برای هر مدل ۵ جمله با طول ۲۰ تولید کنید.

۶) **Preplexity**: بررسی کنید چگونه می‌توان Preplexity مدل‌های ایجاد شده را برای یک جمله دلخواه محاسبه کرد و برای فایل تست ارائه شده Preplexity محاسبه و گزرش کنید.(یکبار باهموارسازی و یکبار بدون هموارسازی)

نحوه ارسال تمرین: پیاده‌سازی انجام شده را در قالب یک فایل Jupyter notebook یا Py. به همراه یک گزارشکار در قالب PDF. در کوئرا آپلود کنید. توجه داشته باشید که عدم تحویل گزارشکار با **کسر ۲۰ درصد نمره** همراه خواهد بود. همچنین می‌توانید سوالات احتمالی خود را از reyhane.saeidi2012@gmail.com بپرسید. (پ.ن: این ایمیل ادرس فقط برای پاسخ‌گویی به سوالات تمرین است و ارسال تکالیف به این آدرس نمره‌ای را به همراه نخواهد داشت.)

موفق باشید. ☺