به نام خدا

دانشکده مهندسی کامپیوتر و فناوری اطلاعات

تمرین سوم برنامه‌نویسی
مبانی و کاربردهای هوش مصنوعی

توضیحات:

- تمامی فایل‌های تمرین (فایل pdf گزارش و فایل‌های کد) را در یک فایل zip ذخیره کرده و با نام AIP3_stdudentNumber در courses بارگذاری نمایید.
- مهلت تمرین تا ۲۰ بهمن ۹۹ می‌باشد.
- تمرین‌های باید تک نفری انجام شوند و با هرگونه مشابهت در کدها برخورد خواهد شد.
- در صورت داشتن هر گونه سوال به ah.rasoulian@gmail.com ایمیل دهید.

نیم‌سال اول ۰۰-۹۹

این شعر از کیست؟

شرح مسئله:

در این تمرین قصد داریم برنامه‌ای طراحی کنیم که با دریافت یک مصرع شعر، نام شاعر سراینده‌ی آن را حدس بزند.

برای این مساله، مجموعه شعرهای مربوط به سه شاعر پرآوازه ایرانی (فردوسی، حافظ و مولانا) داده شده است. این مجموعه به دو مجموعه‌ی داده‌های آموزشی[1] و آزمایشی[2] تقسیم شده است.

در مجموعه‌ی آموزشی سه فایل .txt قرار دارد که حاوی اشعار هرکدام از شاعرهای نامبرده است (فرمت این فایل‌ها به این صورت است که در هر خط یک مصراع از شاعر مورد نظر آورده شده است). شما می‌بایست با استفاده از مجموعه‌ی آموزشی، یک مدل زبانی برای هرکدام از این شاعرها تولید کرده و سپس براساس این مدل‌ها تشخیص دهید که هر مصراع در پیکره آزمایشی متعلق به کدام شاعر است.

مراحل انجام کار:

برای این مساله بایستی گام‌های زیر را طی کنید.

۱- ساخت یک واژه‌نامه[3] با استفاده از مجموعه‌ی آموزشی:

برای این منظور می‌توانید کلمات موجود در مجموعه‌ی آموزشی را استخراج و شمارش نمایید. سپس کلماتی که تکرار آن‌ها کم است را حذف نمایید (براساس یک حدآستانه به عنوان مثال کلماتی که کمتر از ۲ بار در کل مجموعه داده تکرار شده‌اند) و آنچه باقی می‌ماند را به‌عنوان واژه‌نامه درنظر بگیرید.

۲- ساخت مدل زبانی برای هرکدام از شاعرها

در این مرحله برای هرکدام از شاعرها یک مدل زبانی مطابق با آنچه که در کلاس گفته شده تولید نمایید. برای این منظور از مدل‌های زبانی بایگرام و یونیگرام استفاده کرده و به‌منظور هموارسازی از روش Backoff model با استفاده از رابطه‌ی زیر کمک بگیرید.

$$\hat{P}(c_i|c_{i-1}) = \lambda_3 P(c_i|c_{i-1}) + \lambda_2 P(c_i) + \lambda_1 \epsilon$$

$$\lambda_1 + \lambda_2 + \lambda_3 = 1$$

$$0 < \epsilon < 1$$

---

[1] Train set
[2] Test set
[3] Dictionary

۳- محاسبه‌ی احتمال مصراع‌های موجود در مجموعه‌ی آزمایشی با استفاده از مدل‌های زبانی ساخته شده

به ازای هرکدام از مدل‌های زبانی که در گام دوم ساخته شد، احتمال مربوط به هرکدام از مصرع‌های موجود در مجموعه‌ی آزمایشی را محاسبه نمایید. به‌این‌ترتیب برای هرکدام از مصرع‌ها، سه مقدار احتمال محاسبه می‌شود. در نهایت شاعری که احتمال بیش‌تری با توجه به مدل زبانی متناظر با او به‌دست آمد را به عنوان شاعر آن مصرع معرفی کنید.

(فرمت فایل آزمایشی به این صورت است که در هر خط یک مصرع به همراه برچسب متناظر با شاعر آن مصرع آورده شده است که با کاراکتر "t\" از یکدیگر جدا شده‌اند در اینجا اشعار فردوسی با برچسب یک، اشعار حافظ با برچسب ۲ و اشعار مولانا با برچسب ۳ مشخص شده‌اند).

گزارش:

مساله را به ازای پارامترهای مختلف λ و ϵ حل کرده (حداقل ۲ حالت برای λ و ۲ حالت برای ϵ، مجموعا ۴ حالت) و برای هر کدام از آنها دقت تشخیص برنامه را ارزیابی و گزارش کنید. سپس با توجه به دقت‌های به دست آمده، بهترین پارامترها و علت برتری آنها را ذکر کنید.

برای محاسبه دقت، نسبت تعداد مصراع‌هایی که برنامه شما شاعر آن را درست حدس زده است به تعداد کل مصراع‌های موجود در پیکره آزمایشی را محاسبه کنید.