به نام خدا

تمرین اول درس پردازش زبان طبیعی آماری

«آشنایی با مدل‌های زبانی و کاربردهای آن»

استاد درس: دکتر ممتازی

دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر

زمستان ۹۹

برای ارسال تمرین به نکات زیر توجه کنید.

۱- در جدول زیر نحوه اعمال نمره منفی برای تاخیر در ارسال تمرین‌ها ذکر شده‌است.

میزان تاخیر (روز)	میزان جریمه
۱ الی ۲ روز	هر روز ۵٪
۲ الی ۶ روز	هر روز ۱۰٪

در صورتی که برای ارسال تمرین‌ها بین ۷ تا ۱۴ روز تاخیر داشته باشید، نمره شما از ۵۰٪ محاسبه می‌شود و پس از این بازه به تمرین ارسالی نمره‌ای تعلق نمی‌گیرد.

۲- هرگونه کپی‌برداری در انجام تمرین‌ها موجب کسر نمره خواهد شد.

۳- آخرین مهلت ارسال تمرین، **ساعت ۲۳:۵۵ روز چهارشنبه ۱۸ فروردین** می‌باشد.

۴- فایل های ارسالی خود شامل فایل‌های پیاده‌سازی و گزارش را فشرده کنید و با عنوان **«شماره دانشجویی_HW1»** مانند `HW1_97131022` ارسال کنید.

۵- زبان برنامه‌نویسی برای انجام تمرین‌ها، پایتون یا جاوا در نظر گرفته شده‌است.

۶- کدهای ارسالی خود را برای افزایش خوانایی و درک بهتر به صورت مناسب کامنت‌گذاری کنید.

۷- در صورت هرگونه سوال یا مشکل می‌توانید با تدریس‌یار درس از طریق ایمیل زیر در ارتباط باشید.

مریم موسویان mousavian12@gmail.com

بخش اول: تعریف مسئله و معرفی دادگان

در این تمرین قصد داریم برای مجموعه دادگان اشعار فارسی[1] یک مدل زبانی و یک دسته‌بند ارائه دهیم. در این دادگان هر سه خط یک داده را تشکیل می‌دهد. خط اول برچسب داده یعنی شاعر سراینده آن بیت و دو خط بعدی دو مصرع آن بیت می‌باشد. شاعر متناظر با هر برچسب در جدول ۱ معرفی شده است. مجموعه دادگان در سه فایل `train.txt`، `valid.txt` و `test.txt` که به ترتیب برای آموزش[2]، ارزیابی[3] و آزمون[4] می‌باشد، در اختیار شما قرار گرفته‌است. چند نمونه از مجموعه دادگان مورد استفاده در جدول ۲ نمایش داده شده‌است. مجموعه دادگان و کلیه فایل‌های مورد نیاز برای این تمرین از لینک[5] موجود در پاورقی قابل دانلود است.

جدول۱- مشاعیر و برچسب‌های متناظر

شاعر	برچسب
امیر معزی	amir
ملک اشعرای بهار	bahar
قاآنی شیرازی	ghaani
امیر خسرو	Khosro
مولانا	moulavi
سنایی غزنوی	sanaee

جدول۲- چند نمونه از مجموعه داده اشعار فارسی

شاعر	مصرع اول	مصرع دوم
sanaee	باغ دل را تو از بدی کن پاک	تا برآید نهال تو چالاک
ghaani	خلق حیرت زده مانند به مانند صور	غم مخور زآنکه به یک حال نماندست جهان
bahar	هرچند گرفتارم آزادم آزاد	هرچند تهیدستم خرسندم خرسند

[1] دادگان مورد استفاده در این تمرین با اعمال تغییراتی روی دادگان مخزن https://github.com/amnghd/Persian_poems_corpus تهیه شده‌است.

[2] Train

[3] Validation

[4] Test

[5] https://drive.google.com/drive/folders/1pgibqZ6kyqFw_qm8WlsF5lM8tXAP15Jg?usp=sharing

بخش دوم: مدل‌های زبانی

در این قسمت می‌خواهیم مدل‌های زبانی یونیگرم[6] و بایگرم[7] را بر روی دادگان آموزش آماده کنیم. مدل‌های زبانی ایجادشده باید به روش Absolute Discounting هموارسازی[8] شوند و مقدار بهینه پارامتر هموارسازی با استفاده از دادگان ارزیابی محاسبه شود.

الف) مقدار perplexity را بر روی دادگان آزمون برای مدل‌های یونیگرم و بایگرم گزارش کنید.

ب) مقدار perplexity را بر روی دادگان آزمون برای مدل‌های یونیگرم و بایگرم به صورت مجزا برای هر شاعر گزارش کنید.

بخش سوم: انتخاب ویژگی

دسته‌بندی متون می‌تواند به جای استفاده از کل کلمات پیکره با کلمات محدود و مهم‌تری انجام شود. به این منظور از الگوریتم‌های انتخاب ویژگی استفاده می‌شود. در این قسمت با استفاده از هر دو معیار information gain و $\chi - \text{square}$ ویژگی (کلمه) انتخاب کنید. ۲۰۰ کلمه اول انتخاب‌شده با استفاده از هر معیار را به همراه امتیاز نظیر آن‌ها در گزارش خود قرار دهید. با در نظر گرفتن امتیاز به‌دست آمده برای هر شاعر در مقابل هر یک از کلمات منتخب نام شاعری که باعث انتخاب آن واژه شده است را بنویسید. نتایج به‌دست‌آمده برای دو روش انتخاب ویژگی را مقایسه و تحلیل کنید.

بخش چهارم: ارزیابی دسته‌بندی

در این قسمت می‌خواهیم با استفاده از مدل‌های زبانی یونیگرم و بایگرم مجموعه اشعار فارسی را با روش بیز دسته‌بندی کنیم. حاصل این دسته‌بندی تشخیص شاعر برای هر بیت ورودی می‌باشد. نتایج آزمایش‌های زیر را به صورت کامل با هم مقایسه و تحلیل کنید. برای ارزیابی دسته‌بندی از سه معیار F1-measure، precision و recall استفاده کنید. این معیارهای ارزیابی را به صورت micro average و macro average محاسبه و گزارش کنید.

---

[6] Unigram
[7] Bigram
[8] Smoothing

الف) با استفاده از ۲۰۰ ویژگی به‌دست‌آمده در بخش قبل مدل زبانی یونیگرم ایجاد کنید. با استفاده از این مدل اشعار آزمون را با روش بیز دسته‌بندی کنید. در مورد اینکه کدام روش انتخاب ویژگی بهتر است براساس نتیجه F1-measure روی دادگان ارزیابی تصمیم‌گیری نمایید.

ب) مدل یونیگرم را بدون انتخاب ویژگی ایجاد کنید و دسته‌بندی اشعار آزمون را با استفاده از روش بیز انجام دهید.

ج) در این بخش بدون انتخاب ویژگی مدل بایگرم تهیه کنید و اشعار آزمون را دسته‌بندی کنید. در این مرحله از هموارسازی به روش Absolute Discounting به منظور حل مشکل الگوهای دیده‌نشده[9] استفاده کنید. هموارسازی را با استفاده از مقادیر ۰/۱، ۰/۴ و ۰/۷ برای δ انجام دهید و بهترین مقدار را براساس دادگان ارزیابی انتخاب کنید.

بخش پنجم: تحلیل نتایج

در این قسمت بهترین مدل دسته‌بند از میان روش‌های ارزیابی‌شده در بخش چهارم را انتخاب کنید. سپس با ارائه confusion matrix نتایج دسته‌بندی را تحلیل کنید که کدام شاعرها بیشترین شباهت و کدام شاعرها کمترین شباهت را در اشعار دارند.

براساس نتایج تحلیل خود سه زوج از شاعران را انتخاب کنید:

- یک زوج سخت که شامل دو شاعر با نوشتار شبیه به هم است.
- یک زوج آسان که شامل دو شاعر با نوشتار کاملاً متفاوت است.
- یک زوج متوسط که نه خیلی سخت و نه خیلی آسان است.

سپس برای هر یک از زوج شاعران فوق یک دسته‌بندی بیز دودویی تنها بر روی اشعار همان دو شاعر مدنظر آموزش دهید و نتیجه ارزیابی دسته‌بند دودویی را گزارش کنید. مدل و پارامتر دسته‌بندی در این مرحله مشابه بخش چهارم باشد و نیاز به انتخاب مجدد مدل بهینه نیست.

موفق باشید

[9] Unseen Patterns