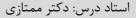
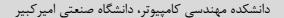


تمرین اول درس پردازش زبان طبیعی آماری

«آشنایی با مدلهای زبانی و کاربردهای آن»









برای ارسال تمرین به نکات زیر توجه کنید.

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

ميزان جريمه	ميزان تاخير (روز)	
هر روز ۵٪	۱ الی ۲ روز	
هر روز ۱۰٪	۲ الی ۶ روز	

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 نمی گیرد.

- ۲- هرگونه کپیبرداری در انجام تمرینها موجب کسر نمره خواهد شد.
- ۳- آخرین مهلت ارسال تمرین، ساعت ۵۵:۲۳ روز چهارشنبه ۱۸ فروردین میباشد.
- ۴- فایل های ارسالی خود شامل فایلهای پیاده سازی و گزارش را فشرده کنید و با عنوان «شماره دانشجویی_۱۳۳۱»
 مانند 97131022 HWI 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 کامنت گذاری کنید.
 - ۷- در صورت هرگونه سوال یا مشکل می توانید با تدریسیار درس از طریق ایمیل زیر در ارتباط باشید.

mousavian12@gmail.com

مريم موسويان

بخش اول: تعریف مسئله و معرفی دادگان

در این تمرین قصد داریم برای مجموعه دادگان اشعار فارسی ایک مدل زبانی و یک دستهبند ارائه دهیم. در این دادگان هر سه خط یک داده را تشکیل می دهد. خط اول برچسب داده یعنی شاعر سراینده آن بیت و دو خط بعدی دو مصرع آن بیت می باشد. شاعر متناظر با هر برچسب در جدول ۱ معرفی شده است. مجموعه دادگان در سه فایل txt .train.txt و txt .txt txt txt

جدول۱- مشاعیر و برچسبهای متناظر

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

جدول۲- چند نمونه از مجموعه داده اشعار فارسی

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

۱ 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 یونیگرم² و بایگرم^۷ را بر روی دادگان آموزش آماده کنیم. مدلهای زبانی ایجادشده باید به روش Absolute Discounting هموارسازی^۸ 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 مدل های یونیگرم و بایگرم به صورت مجزا برای هر شاعر گزارش کنید.

بخش سوم: انتخاب ویژگی

دستهبندی متون می تواند به جای استفاده از کل کلمات پیکره با کلمات محدود و مهم تری انجام شود. به این منظور از information gain و نتخاب ویژگی استفاده می شود. در این قسمت با استفاده از هر دو معیار χ ویژگی (کلمه) انتخاب کنید. ۲۰۰ کلمه اول انتخاب شده با استفاده از هر معیار را به همراه امتیاز نظیر آنها در گزارش خود قرار دهید. با در نظر گرفتن امتیاز به دست آمده برای هر شاعر در مقابل هر یک از کلمات منتخب نام شاعری که باعث انتخاب آن واژه شده است را بنویسید. نتایج به دست آمده 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 precision ،F1-measure و precision کنید. این معیارهای ارزیابی را به صورت macro average و شادت عرصت عدار به صورت عدورت معاسبه و گزارش کنید.

⁷ Bigram

⁶ Unigram

⁸ Smoothing

الف) با استفاده از ۲۰۰ ویژگی بهدستآمده در بخش قبل مدل زبانی یونیگرم ایجاد کنید. با استفاده از این مدل الف) با استفاده از این مدل الف) با استفاده از این مدل الف) با استفاده از این مدل ویژگی بهتر است براساس نتیجه - ۴۱ اشعار آزمون را با روش بیز دستهبندی کنید. در مورد اینکه کدام روش انتخاب ویژگی بهتر است براساس نتیجه - ۴۱ اشعار آزمون را با روش بیز دستهبندی کنید. سومیم گیری 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 و استفاده کنید. هموارسازی را با استفاده از مقادیر δ برای δ انجام دهید و بهترین مقدار را براساس دادگان ارزیابی انتخاب کنید.

بخش ينجم: تحليل نتايج

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 باشيد

⁹ Unseen Patterns