بسمه تعالی

دانشگاه صنعتی امیرکبیر

دانشکده مهندسی کامپیوتر و فناوری اطلاعات

پروژه دوم مبانی و کاربردهای هوش مصنوعی

توضیحات:

- مهلت تحویل پروژه تا روز جمعه ۴ بهمن در نظر گرفته شده است.
- پروژه باید به صورت انفرادی انجام شود.
- درصورت مشاهده هرگونه تقلب، نمره صفر برای کل تکلیف منظور خواهد شد.
- تمیزی و خوانایی گزارش از اهمیت بالایی برخوردار است.
- لطفا گزارش تمرین خود و فایل برنامه را با نام «P2_StudentNumber.zip» در سایت درس در مهلت معین بارگزاری نمایید.
- در ازای هر روز تاخیر ۱۰ درصد از نمره شما کسر خواهد شد.
- در صورت داشتن اشکال می‌توانید از طریق ایمیل «miladbohlouli@gmail.com» با تدریس یار درس در ارتباط باشید.

در این پروژه قصد داریم دسته‌بندی متون را با استفاده از دسته‌بند بیز و مدل‌های Unigram و Bigram انجام دهیم. بدین منظور باید از تمامی کلمات موجود در مجموعه داده به عنوان ویژگی استفاده شود. عملیات زیر را به ترتیب بر روی مجموعه داده انجام دهید.

۱- مجموعه داده را از لینک زیر[1] دریافت نمایید.

۲- مجموعه داده‌ها به دو قسمت آموزش و تست تقسیم شده‌اند. شما باید با استفاده از مجموعه داده آموزش، مدل‌های زبانی را استخراج نمایید. سپس از مجموعه تست برای ارزیابی مدل‌ها استفاده کنید. برای استخراج مدل‌ها عملیات زیر را انجام دهید.

۳- مجموعه داده‌ها در قالب زیر ذخیره شده‌اند.

ادب و هنر@@@@@@@@@@ جشنواره بین المللی موسیقی امروز آغاز می شود بخش مسابقه این جشنواره شامل موسیقی مقامی کرمان و موسیقی جوان تهران است گروه هنری پانزدهمین جشنواره بین المللی موسیقی فجر بعدازظهر امروز با هنرنمایی بیش از گروه موسیقی داخلی و خارجی و تکنوازان برگزیده شش کشور جهان در محل تالار وحدت آغاز به کار می کند این جشنواره باهدف معرفی فرهنگ موسیقی در ابعاد مختلف تشکیل گروههای موسیقی حمایت از جوانان و گسترش فرهنگ شنیداری جامعه در بخش موسیقی با شعار صلح وگفتگو در تالار فرهنگی و هنری تهران برگزار می شود بخش مسابقه جشنواره پانزدهم موسیقی فجر شامل دو بخش موسیقی مقامی کرمان و موسیقی جوان تهران است که در این بخش برگزیدگان این دو جشنواره برنامه اجرا می کنند اجرای برنامه های ارکستر سمفونیک گروه ارکستر ملی و گروه کر به همراه گروه های موسیقی کشورهای ارمنستان گرجستان تاجیکستان فرانسه اتریش و آلمان از دیگر برنامه های جشنواره پانزدهم موسیقی فجر است در جشنواره امسال موسیقی فجر از استادان علی تجویدی و جلیل شهناز از پیشکسوتان موسیقی ایرانی تجلیل می شود همچنین گروه موسیقی بانوان در طول روزهای برپایی جشنواره موسیقی فجردر تالار فرهنگ برای بانوان علاقه مند به موسیقی سنتی برنامه اجرا می کنند شرکت کنندگان در پانزدهمین جشنواره موسیقی فجر از تا بهمن ماه در تالارهای وحدت حرکت فرهنگ فرهنگسرای بهمن بنیاد آفرینش های هنری نیاوران سالن میراث فرهنگی سالن رودکی و مجموعه فرهنگی آزادی اجراخواهند داشت

شکل ۱ نمونه‌ای از مجموعه داده

همانطور که مشاهده می‌شود، کلاس هر داکیومنت قبل از آن مشخص شده است. عناوین را جدا کرده به صورت مناسب ذخیره نمایید.

[1] https://www.dropbox.com/s/q4l1gaihr29iitm/HAM-Train-Test.zip?dl=1

۴- مدل‌های زبانی را متناسب با کلاس داکیومنت استخراج نمایید. به عنوان مثال در حالت Unigram، برای هر کلمه متناسب با هر کلاس داکیومنت، احتمالی خواهد بود که با استفاده از تعداد دفعات تکرار این کلمه در این کلاس محاسبه شده است. در حالت Bigram این احتمالات برای زوج کلمات خواهد بود.

۵- از روش هموارسازی backoff استفاده نمایید. برای پارامترهای این هموارسازی۳ حالت مختلف را امتحان نمایید و بهترین حالت را گزارش نمایید.

۶- الگوریتم بیز ساده را پیاده‌سازی نمایید و بر روی مجموعه داده تست امتحان کنید. (یکی از مشکلاتی که ممکن است در این الگوریتم رخ دهد، صفر شدن حاصل‌ضرب احتمال‌ها می‌باشد که علت آن ضرب اعداد کوچک‌تر از ۱ در یکدیگر می باشد، به منظور حل این مشکل می‌توانید از لگاریتم احتمالات استفاده نمایید که در شکل زیر نمایش داده شده است.)

$$\log(P(\text{class}_i \mid \mathbf{data})) \propto \log(P(\text{class}_i)) + \sum_j \log(P(\text{data}_j|\text{class}_i))$$

شکل ۲ لگاریتم احتمالات در الگوریتم بیز ساده

۷- در انتها، برای هر کدام از داکیومنت‌های تست، کلاسی را پیش‌بینی نمایید و معیارهای ارزیابی زیر را با توجه به کلاس درست آن‌ها محاسبه نمایید.

- Precision
- Recall
- F-measure

روش به‌دست‌اوردن precision و recall در حالت چندکلاسی در مثال زیر نشان داده شده‌است:

همانطور که در جدول زیر نشان داده شده است، برای هر کدام از کلاس‌ها تعداد نمونه‌هایی که به درستی کلاس‌بندی شده‌اند نشان داده‌شده است. به عنوان مثال برای class1 ، TP1 نشان دهنده‌ی داده‌هایی است که به درستی با کلاس ۱ برچسب‌گزاری شده‌اند. این در حالی است که FP12 نشان دهنده‌ی داده‌هایی است که برچسب اصلی آن‌ها کلاس ۱ می‌باشد و مدل به اشتباه آن را کلاس۲ برچسب‌گزاری نموده است.

		Predicted		
Actual		Class1	Class2	Class3
	Class1	TP1	FP12	FP13
	Class2	FP21	TP2	FP23
	Class3	FP31	FP32	TP3

با توجه به معلومات بالا، precision و recall برای کلاس ۱ به صورت زیر تعریف می‌شوند:

$$precision_{class1} = \frac{TP1}{TP1 + Fp21 + FP31}$$

$$recall_{class1} = \frac{TP1}{TP1 + Fp12 + FP13}$$