به نام خدا

تمرین سری اول پردازش زبان طبیعی

پاییز ۱۴۰۴- دانشگاه گیلان

مدرس: یاسمن برشبان

سوال۱: معیار ارزیابی (۱۵ نمره)

برای یک دسته بندی دوکلاسه نتایج زیر به دست آمده است.

الف) TN,TP, FN,FP را مشخص کنید.

ب) معیارهای precision, recall, F1 و accuracy را محاسبه کنید.

	Predicted (positive)	**Predicted (Negative)**
Actual (Positive)	**۴۵**	**۱۵**
Actual (Negative)	**۱۰**	**۳۰**

سوال۲: نایوبیز: (۲۰ نمره)

با توجه به جدول زیر مشخص کنید سند شماره ۶ مربوط به چه کلاسی است. از روش نایوبیز با استفاده از هموارسازی add-1 استفاده کنید.

$$\hat{P}(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|}$$

Documents	**Words**	**Class**
1	Technology, Stemming, Lemmatization, Learning	AI
2	Recognition, Learning, Data	AI
3	Neural, Network, Lemmatization	AI
4	Router, Switch, Network	Network
5	Network, Bandwidth, Router	Network
6	Recognition, Learning, Router	?

سوال ۳: پیش پردازش + نایو بیز (۳۰ نمره- پیاده سازی)

در این سؤال می‌خواهیم بر روی دیتاست Sentipers تحلیل احساسات انجام دهیم. جملات در این دیتاست امتیازی بین −۲ تا +۲ دارند .برای سهولت، احساسات را به سه دسته زیر تبدیل کنید:

- Positive برچسب‌های +۱ و +۲
- Neutral برچسب ۰
- Negative برچسب‌های −۱ و −۲

از نسخه یکپارچه دیتاست استفاده کنید. لینک داده

الف) پیش‌پردازش متن (۵ نمره)

با استفاده از ابزارهای Hazm یا Parsivar مراحل زیر را روی ستون متنی اعمال کنید:

۱.نرمال‌سازی

۲. Lemmatization یا Stemming

۳. حذف کلمات توقف(Stopwords)

در پایان، با توجه به پیش پردازش های انجام شده چهار نسخه مختلف از داده بسازید. ۱-نسخه خام(Raw Text) ، ۲-نسخه فقط نرمال‌سازی‌شده، ۳- نسخه نرمال + stopword و ۴-نسخه full-preprocessed کامل‌ترین نسخه

ب) تقسیم‌بندی داده و آموزش مدل Naïve Bayes (۲۰ نمره)

برای هر یک از چهار نسخه داده بالا مراحل زیر را انجام دهید

۱. ابتدا داده را Shuffle کنید.
۲. سپس با استفاده از Stratified Split به ۷۰ درصد آموزش و ۳۰ درصد تست تقسیم کنید:
۳. یک طبقه‌بند Multinomial Naive Bayes آموزش دهید.
۴. روی مجموعه آزمون موارد زیر را گزارش کنید:
   - Precision برای هر سه کلاس
   - Recall برای هر سه کلاس
   - F1-score برای هر سه کلاس
   - F1 Macro و Accuracy کل
   - Confusion Matrix

ج) مقایسه داده خام و داده پیش‌پردازش‌شده (۵ نمره)

با توجه به نتایج به دست آمده از مرحله قبل به سوالات زیر پاسخ دهید.

۱. کدام مراحل پیش‌پردازش بیشترین تاثیر را بر بهبود مدل داشته‌اند؟
۲. آیا حذف stopwords باعث کاهش یا افزایش دقت شده؟ چرا؟
۳. پیش‌پردازش چگونه روی کلاس‌های مختلف (مثبت/منفی/خنثی) تأثیر گذاشته؟

سؤال ۴ – پیاده‌سازی سیستم بازیابی اطلاعات (IR) فارسی با TF-IDF بر روی PQuAD (۳۵ نمره)

در این تمرین قصد داریم یک سیستم ساده‌ی بازیابی پاسخ (Answer Retrieval) ایجاد کنیم که با دریافت یک سؤال، بتواند پاراگراف مناسب را از میان اسناد موجود در دیتاست PQuad پیدا کرده و رتبه‌بندی کند.

الف) آماده‌سازی داده و پیش‌پردازش (۵ نمره)

۱ .دیتاست PQuAD را بارگذاری کنید.
۲ .از ۱۰،۰۰۰ مورد اول بخش train به‌عنوان مجموعه اسناد (Documents / Index) استفاده کنید.

- برای هر سند، از ستون context متن پاراگراف را استخراج کنید.
- **.از** ۱،۰۰۰ نمونه اول بخش test به‌عنوان مجموعه پرسش **(Queries)** استفاده کنید.
- از ستون question به عنوان متن سؤال استفاده کنید.
- برای هر سؤال، contextمرتبط (در دیتاست موجود است) سندِ صحیح محسوب می‌شود.

۴ .روی هر دو مجموعه اسناد و پرسش‌ها پیش‌پردازش فارسی انجام دهید:

- نرمال‌سازی (اصلاح ی/ک، حذف علائم اضافی و ...)
- حذف علائم نگارشی
- حذفstopword های فارسی
- توکن‌سازی

ب) ساخت TF-IDF و انجام بازیابی (۱۵ نمره)

۱ .با استفاده از TfidfVectorizer، متن ۱۰،۰۰۰ پاراگراف train را به ماتریس TF-IDF تبدیل کنید.
۲ .متن ۱،۰۰۰ سؤال test را نیز با *همان vectorizer* و با *همان vocabulary* به TF-IDF تبدیل کنید

۳. برای هر سؤال شباهت آن را با تمام ۱۰،۰۰۰ سند با استفاده از cosine similarity محاسبه کنید. اسناد را بر اساس شباهت از بیشترین به کمترین رتبه‌بندی کنید.سندِ درست همان contextمرتبط با سؤال است که در دیتاست مشخص شده است. هنگام ارزیابی، رتبه این سند را در لیست خروجی پیدا کنید.

ج) ارزیابی سیستم MRR (۵ نمره)

روی ۱،۰۰۰ سؤال test ، موارد زیر را محاسبه کرده و گزارش کنید:

Precision@5

Recall@5

MRR (Mean Reciprocal Rank)

د) مقایسه عملکرد با و بدون پیش‌پردازش (۵ نمره)

تمام مراحل بخش ب و ج را یک بار دیگر انجام دهید، اما این بار بدون هیچ پیش‌پردازشی .نتایج دو حالت را مقایسه کنید:

- دقت و MRR چقدر تغییر کرده‌اند؟
- کدام مقادیر بهبود یافته یا بدتر شده‌اند؟
- چرا پیش‌پردازش در برخی پرسش‌ها اثر بیشتری گذاشته است؟

هـ) تحلیل خطا و ارائه پیشنهادهای بهبود (۵ نمره)

برای ۳ تا ۵ سؤال که سیستم در آن‌ها عملکرد ضعیفی داشته (مثلاً سند درست در رتبه‌های پایین قرار گرفته):

۱ .متن سؤال را نشان دهید.
۲ .سند درست را نمایش دهید.
۳ .سه سند اول رتبه‌بندی‌شده را مقایسه کنید.
۴ .توضیح دهید چرا TF-IDF اشتباه کرده است
۵ .حداقل سه راهکار برای بهبود سیستم پیشنهاد کنید .

موفق باشید