بسم الله الرحمن الرحیم

تمرین سوم درس مباحث ویژه

نام استاد: حمیدرضا برادران کاشانی

طراح آموزشی: ریحانه سعیدی، مهدی دارونی

دانشکده مهندسی کامپیوتر

گروه هوش مصنوعی

توضیحات: مهلت تحویل این تمرین تا تاریخ **۱۴۰۱/۱۰/۶** در نظر گرفته شده است. پس از این تاریخ به مدت **۳ روز** یعنی تا تاریخ **۱۴۰۱/۱۰/۹** تمرینات با کسر **۳۰ درصد** از نمره تحویل گرفته می‌شود. لازم به ذکر است که به تمرینات تحویلی پس از تاریخ نمره‌ای تعلق نخواهد گرفت.

هدف آموزشی تمرین: این تمرین از دو بخش تشکیل شده است. در بخش اول یک سیستم توصیه‌گر ساده را پیاده‌سازی می‌کنیم و در بخش دوم به بررسی تحلیل احساسات خواهیم پرداخت.

بخش اول: سیستم توصیه‌گر

مراحل تمرین:

۱) **آماده‌سازی داده‌ها**: دیتاست مورد استفاده، TMDB است که حاوی اطلاعات فیلم‌ها است. شما می‌توانید این دیتاست را به‌صورت دستی یا از طریق دستور wget! آن را در داخل گوگل کولب دانلود کنید.

۲) **پیش‌پردازش:** عنوان، ژانر و ... برخی از فیلم‌ها ممکن است از دو بخش تشکیل شده باشد. مثلاً science fiction از دو بخش تشکیل شده است. پس بهتر است فضاهای خالی را حذف کنید.

۳) **تبدیل متون به بردار**: برای تبدیل متن به بردار، روش های مختلفی وجود دارد. Idf*Tf از جمله روش‌های ساده‌ای است که می‌توان به کمک آنها متون را به بردار تبدیل کرد تا آماده استفاده در سیستم

توصیه‌گر شوند. نحوه استخراج TF*IDF را یکبار به‌صورت دستی و بار دیگر با استفاده از کتابخانه scikit-learn بررسی کنید. سپس این دو روش را به صورت جداگانه بر مجموعه داده‌ها اعمال کنید.

۴) **ساخت سیستم توصیه‌گر:** در این سیستم توصیه‌گر قصد داریم. با دریافت اسم یک فیلم که در دیتاست موجود است، فیلم‌های شبیه به آن را بیابیم. بدین منظور ابتدا نوع فیلم (type) مورد جست‌و جو را تبدیل به بردار TF*IDF کنید و از طریق شباهت کسینوسی، میزان شباهت آن را با تمامی فیلم‌های موجود در دیتاست پیدا کنید. سپس بر حسب میزان شباهت کسینوسی فیلم‌ها را مرتب کنید و ۵ فیلم با بیشترین امتیاز را برگردانید. این سیستم را بر روی فیلم‌های Scream 3، Mortal Kombat و Runaway Bride اعمال کنید و نتیجه را گزارش کنید. (توجه داشته باشید که نتایج باید یکبار با بردار TF*IDF دستی و بار دیگر با استفاده از کتابخانه scikit-learn گزارش کنید.)

بخش دوم: تحلیل احساسات

۱) **آماده‌سازی داده**: دیتاست مورد استفاده، Twitter US Airline است. همانند بخش قبلی این دیتاست را آماده کنید.

۲) **پیش‌پردازش**: در این تمرین قصد داریم، یک دسته‌بند باینری طراحی کنیم. اما دیتاست ارائه شده، شامل سه دسته نظرات مثبت، نفی و خنثی است. در اولین مرحله از پیش‌پردازش داده‌هایی که خنثی هستند را حذف کنید. سپس با دستور factorize یا هر دستور دیگر برچسب داده‌های مثبت را به عدد صفر و برچسب داده‌های منفی را به عدد یک تبدیل کنید.

۳) **تبدیل متون به بردار**: بدین منظور به هر یک از کلمات منحصر به فرد یک عدد منحصر به فرد می‌دهیم و سپس آن کلمه را با عدد اختصاص داده شده جایگزین می‌کنیم. برای این هدف ابتدا تمامی داده‌های مربوط به ستون text را بازیابی کنید. سپس با استفاده از متد tensorflow.keras.preprocessing.text متون را توکن‌بندی کنید. سپس با اعمال توابع fit_on_texts() و text_to_sequence() متون را تبدیل به بردار کنید. از آنجایی که برای پردازش نیازمند داده‌هایی با بعد یکسان هستیم لذا بر روی دیتاست متد pad_sequences را هم اعمال کنید.

۴) **ساخت مدل دسته‌بند**: در این بخش با استفاده از keras، یک مدل پنج لایه که شامل لایه‌های زیر است را بسازید و مدل را برای پنج ایپاک متوالی آموزش دهید. (راهنمایی)

Layer (type)	Output Shape	Param #
embedding (Embedding)	(None, 200, 32)	423488
spatial_dropout1d (SpatialDropout1D)	(None, 200, 32)	0
lstm (LSTM)	(None, 50)	16600
dropout (Dropout)	(None, 50)	0
dense (Dense)	(None, 1)	51

۵) **رسم نمودار**: در انتها روند بهبود accuracy در پنج ایپاک آموزش را نمایش دهید.

نحوه ارسال تمرین: پیاده‌سازی انجام شده را در قالب یک فایل Jupyter notebook یا Py. به همراه یک گزارشکار در قالب PDF. در کوئرا آپلود کنید. توجه داشته باشید که عدم تحویل گزارشکار با **کسر ۲۰ درصد نمره** همراه خواهد بود. همچنین می‌توانید سوالات احتمالی خود را از reyhane.saeidi2012@gmail.com بپرسید. (پ.ن: این ایمیل ادرس فقط برای پاسخ‌گویی به سوالات تمرین است و ارسال تکالیف به این آدرس نمره‌ای را به همراه نخواهد داشت.)

موفق باشید. ☺