به نام خدا

تمرین سوم درس جستجو و بازیابی اطلاعات در وب، «سامانه‌های توصیه‌گر»

استاد درس: دکتر ممتازی

پاییز ۱۴۰۲– دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر

نکاتی در مورد این تمرین که نیاز به توجه و دقت دوستان دارد:

۱- برای ارسال پاسخ تمرین‌های این درس، **مجموعا ۱۰ روز** زمان تاخیر مجاز در نظر گرفته شده‌است و در صورت تجاوز مجموع زمان تاخیرها از مقدار در نظر گرفته‌شده، پاسخ ارسال‌شده مورد بررسی قرار نخواهد گرفت.

۲- برای طرفین مشارکت‌کننده در هرگونه کپی‌کردن، بدون اغماض، نمره **منفی ۱۰۰** در نظر گرفته می‌شود.

۳- آخرین مهلت ارسال تمرین، ساعت **۲۳:۵۵ روز سه‌شنبه ۱۰ بهمن ۱۴۰۲** می‌باشد. این زمان با توجه به شرایط، جمع‌بندی‌ها، زمان لازم برای امتحانات و سایر تمرین‌ها در نظر گرفته شده‌است و **قابل تمدید نمی‌باشد**.

۴- دوستان فایل ارسالی خود را به صورت فشرده و به صورت «شماره دانشجویی_HW3» مانند HW3_400131123 نام‌گذاری کنید. در این فایل باید مواردی نظیر کدها، فایل گزارش و سایر موارد موردنیاز در هنگام بررسی و نمره‌دهی وجود داشته باشد و تنها این فایل جهت نمره‌دهی در نظر گرفته می‌شود.

۵- زبان برنامه‌نویسی پاسخ این تمرین تنها می‌تواند **پایتون** باشد.

۶- به صورت مناسب کامنت‌های لازم را در کدهای خود قرار دهید. به صورتی که بتوان حداقل روال اجرا و موارد مورد نیاز را درک کرد.

۷- سعی کنید ابتدا تمامی سوالات و بخش‌ها را مطالعه کنید.

۸- استفاده از کتابخانه‌های آماده به جز موارد مطرح شده در تمرین مجاز **نمی‌باشد** و شما باید موارد خواسته‌شده را پیاده‌سازی نمایید.

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

n.gholinezhad@aut.ac.ir

بخش اول: معرفی مجموعه‌داده

مجموعه‌داده‌ای[1] که در اختیار شما قرار داده شده است با عنوان بازخورد ضمنی[2] در ادبیات شناخته می‌شود. این نوع مجموعه‌داده فاقد امتیاز صریح کاربر به اقلام بوده و صرفا نشان دهنده تعامل کاربر با آن آیتم است به این صورت که عدد ۱ نشانگر این است که کاربر با آیتم تعامل داشته است و این تعامل می‌تواند به هر نحوی باشد(خرید، بازدید، کلیک و ...) و عدد ۰ نشانگر عدم تعامل کاربر با آن آیتم است.

جدول۱: توضیحات مجموعه‌داده

توضیحات	ویژگی
شناسه یکتا کاربر	user_id
شناسه یکتا آیتم	item_id
امتیاز کاربر به آیتم(تبدیل شده به حالت ۰ و ۱)	rate
متن نظر	review_text

بخش دوم: سامانه توصیه‌گر مبتنی بر Matrix Factorization (۴۰ امتیاز)

هدف از این بخش پیاده سازی الگوریتم گرادیان کاهشی تصادفی[3] برای ساختن یک سیستم توصیه‌گر شخصی‌سازی شده مبتنی بر Matrix Factorization است. در این رویکرد ما به دنبال پیدا کردن دو ماتریس P و Q هستیم به نحوی که $R \cong QP^T$ در این رابطه، ماتریس R بیانگر ماتریس تعامل کاربر-اقلام می‌باشد اندازه ماتریس R برابر m*n است که m تعداد اقلام و n تعداد کاربرها می‌باشد.

تابع خطا برای این مسئله به صورت زیر تعریف می‌شود

$$E = (\sum_{(u,i) \in training} (r_{iu} - q_i . p_u^T) + \lambda (\sum_i \|q_i\|_2^2 \sum_u \|p_u\|_2^2)$$

با توجه به تابع خطای بالا الگوریتم گرادیان تصادفی کاهشی را پیاده سازی کنید.

مقدار λ را برابر ۰.۰۰۱ در نظر بگیرید. همچنین ماتریس P و Q را ۶۴ بعدی در نظر بگیرد، به عبارت دیگر ویژگی پنهان[4] کاربر و آیتم ۶۴ بعد خواهد بود..

[1] Dataset
[2] Implicit feedback
[3] Stochastic Gradient Descent
[4] Latent feature

بخش سوم: ترکیب مدل بخش دوم با بردارهای معنایی عصبی (۴۰ امتیاز)

در این بخش قصد داریم ماتریس Q به دست آمده در بخش دوم را با استفاده از اطلاعات جانبی موجود در مجموعه‌داده غنی‌تر کنیم. در مجموعه‌داده‌ای که در اختیار شما قرار داده شده است برای هر آیتم مجموعه‌ای از نظرات ثبت شده است، وظیفه شما این است که با استفاده از مدل BERT که در تمرین دوم با آن آشنا شدید برای هر آیتم، بازنمایی تمام نظرات ثبت شده برای آن آیتم را به دست آورید، سپس بین بردارهای نظرات به دست آمده برای هر آیتم میانگین بگیرید. بعد از عمل میانگین‌گیری برای هر آیتم یک بردار ۷۶۸ بعدی خواهید داشت.

در ادامه، طبق بخش الف و ب با استفاده از برداری که برای هر آیتم به دست آمده یک ماتریس Q جدید تشکیل دهید و نتایج حاصل را ارزیابی کنید.

الف) جایگزینی: با استفاده از روش کاهش بعد PCA، بردار ۷۶۸ بعدی به دست آمده برای هر آیتم را به بردارهای ۶۴ بعدی تبدیل کنید و ماتریس Q جدید را بسازید. ماتریس جدید به دست آمده از این بخش را با ماتریس Q بخش دوم جایگزین کنید. برای ماتریس P از همان ماتریس نهایی به دست آمده از بخش دوم استفاده کنید.

ب) الحاق: بردار ۷۶۸ بعدی به دست آمده برای هر آیتم از مدل BERT و بردار متناظر آن آیتم از ماتریس Q بخش دوم را با یکدیگر الحاق کرده و بردار جدید را که دارای ابعاد ۸۳۲ است، با استفاده از PCA به یک بردار ۶۴ بعدی تبدیل کنید و ماتریس Q جدید را بسازید. همانند حالت الف ماتریس P بخش دوم را برای این بخش نیز استفاده کنید.

بخش چهارم: ارزیابی نتایج (۲۰ امتیاز)

برای هر دو روش مطرح شده در بالا، با در نظر گرفتن مجموعه‌داده بخش test کیفیت عملکرد سامانه توصیه‌گر را با استفاده از معیارهای ارزیابی Recall، NDCG و rank correlation محاسبه کنید. معیار Recall و NDCG را برای ۲۰ آیتم برتر محاسبه کنید.

*راهنمایی در خصوص نحوه ارزیابی: برای هر کاربر، امتیاز احتمالی کاربر به تمامی آیتم‌هایی که در زمان آموزش مشاهده نکرده است را با استفاده از مدل آموزش دیده شده به دست آورید سپس آیتم‌ها را بر حسب امتیاز پیش‌بینی شده مرتب‌سازی کنید(پس از مرتب‌سازی ۲۰ آیتم بالای لیست برای هر کاربر همان ۲۰ آیتم برتر هستند)

بخش آخر: برخی نکات در مورد گزارش و تمرین

- دادگان مطرح شده در این تمرین و تمامی بخش‌ها همراه با صورت تمرین در سایت درس قرارداده شده است.
- در این تمرین شما مجاز به استفاده کتابخانه‌های زیر و موارد مشابه و هم‌کاربرد با آن‌ها می‌باشد:

numpy, scipy, pandas, genism, pickle, tensorflow, pytorch, keras

- در این تمرین سعی شده است علاوه بر آشنایی شما با کاربرد مباحث ارائه‌شده در کلاس و لمس بهتر آن‌ها، خلاقیت و حل چالش شما نیز ارزیابی شود. لذا در صورتی که در این تمرین چالشی وجود دارد که شما راه‌حلی برای آن ارائه دادید و استفاده کردید، آن را در گزارش بیان کنید. اما اگر مشکلی بزرگ وجود دارد که نیاز به بررسی مجدد دارد، آن را از طریق ایمیل با تدریس‌یاران درس مطرح کنید.
- در صورتی که هر گونه پیش‌پردازش بر روی دادگان انجام دادید آن را در گزارش خود بیان کنید.
- این تمرین **۱ نمره** از بارم کلی تمرین‌های شما را با توجه به پوشش مباحث و حجم تمرین دارد. امتیاز این تمرین از ۱۰۰ محاسبه می‌شود که بارم هر بخش مشخص شده است.
- در تمامی بخش‌ها، میزان نتایج در ارزیابی شما تاثیر چندانی ندارند (مگر اینکه بسیار دور باشد). بلکه میزان تسلط، دیدگاه و پیاده‌سازی، تحلیل‌ها و خلاقیت شماست که در نمره شما تاثیر مستقیم دارد و بر اساس این موارد مورد ارزیابی قرار می‌گیرید.

موفق باشید

قلی نژاد