

به نام خدا



دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)



DEPARTMENT OF COMPUTER
ENGINEERING AND IT

دانشگاه صنعتی امیرکبیر
دانشکده‌ی مهندسی کامپیوتر

تمرین امتیازی مبانی رایانش ابری
دکتر سید احمد جواد

طراح سوال:

احمدی

دی ۱۴۰۱

Recommender system

یکی از رویکردهای **collabrative filtering** انتخاب اجناس بر اساس **user-user** است که با استفاده از آن می‌توانیم سیستم‌های توصیه‌گر بسازیم. مجموعه داده مورد استفاده در این بخش در دو فایل **game.csv** و **ratings.csv** می‌باشد. در فایل اول مشخصات بازی‌ها و در فایل دوم امتیاز کاربرها با جزییات آیدی بازی، آیدی کاربر و امتیاز داده شده است.

با استفاده از رویکرد **collabrative filtering** و بر اساس روش **user-user** سیستم توصیه‌گری بسازید که با دادن آی دی از میان بازی‌هایی که به آنها نمره نداده‌اند، تعداد ۵ بازی با بیشترین شباهت را پیشنهاد کند. در حل این مساله از معیار **Cosine similarity** استفاده کنید.

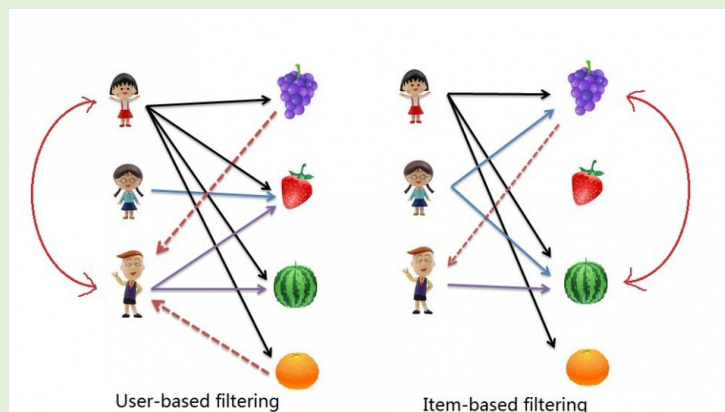
- در خروجی نام بازی‌ها را به همراه امتیاز شباهت آنها (به ترتیب نزولی امتیاز شباهت) بیاورید.
- در صورتی که دوبرنامه امتیاز یکسان داشتند، برنامه با ایندکس کمتر را انتخاب کنید.

در سال ۲۰۰۷ شرکت **Netflix** یک مسابقه برای پژوهش روی سیستم‌های توصیه‌گر و با استفاده از مجموعه داده‌ای از این شرکت ترتیب داد. در سال ۲۰۰۹ جایزه یک میلیون دلاری به یکی از پژوهشگران رسید. یکی از الگوریتم‌هایی که آن‌ها در روش خود استفاده کرده بودند همین روش ماتریس عامل بندی بود و این نشانه قدرت این روش است.

در مسائلی که ما اطلاعات خوبی از کاربر (مثل سن، جنسیت، شغل و ...) داریم اما در مورد آیتم‌ها دچار کمبود اطلاعات هستیم یا به دست آوردن ویژگی‌های آیتم‌ها دشوار است سیستم‌های توصیه‌گر فیلتر مشارکتی بسیار مناسب‌اند. در این روش براساس شباهت رفتاری و عملکردی کاربرانی که در گذشته الگوی رفتاری مشابهی با کاربر فعلی داشته‌اند، پیشنهادها ارائه می‌شود. شاید تعریف آن کمی پیچیده باشد ولی به طور ساده روش فیلتر مشارکتی بر این فرض استوار است که کاربرانی که یک سری نظرهای مشابه درباره یک آیتم دارند، درباره آیتم‌های دیگر هم نظرهای مشابه دارند. فیلتر مشارکتی خود به دو دسته تقسیم می‌شود:

مبتنی بر کاربر (User-based)،

مبتنی بر آیتم (Item-based).



در روش مبتنی بر کاربر میزان شباهت سلیقه ی کاربران مشخص می شود و با توجه به آن، مقدار علاقه مندی کاربر به یک آیتم که تا به حال ندیده است مشخص می شود. در این سیستم از تشکیل بردار کاربر، که شامل تمام آیتم هایی است که کاربر امتیاز دهی کرده است به همراه امتیاز داده شده به آن ها، شروع می کنیم. سپس میزان تشابه کاربرها در یک ماتریس $n \times n$ که n تعداد کاربران است محاسبه می شود. برای محاسبه ی میزان تشابه می توان از تشابه کسینوسی استفاده کرد. حال ماتریس پیشنهادها می تواند تشکیل شود. امتیاز داده شده به آیتم ها در میزان شباهت کاربری که آن آیتم را امتیاز دهی کرده است با کاربر فعال ضرب می شود و به عنوان امتیاز پیشبینی شده در ماتریس پیشنهادها گذاشته می شود. این کار برای تمام آیتم هایی که کاربر ندیده است انجام می شود و سپس امتیازها مرتب می شوند و پیشنهادهای با بیشترین امتیاز، به کاربر هدف توصیه می شوند. کاربرها تمایل دارند محصولات را بخرند که کاربرهای با سلیقه مشابه آنها خریده اند.

برای مثال در جدول زیر کاربر U_1 احتمالا تمایل دارد تا محصول I_2 را بخرد، زیرا کاربرهای U_1 و U_4 محصول I_1 را می پسندند و کاربر U_4 امتیاز بالایی به محصول I_2 داده است.

I_4	I_3	I_2	I_1	
۵	۵	؟	۴	U_1
	۱	۲	۴	U_2
۴	۲		۳	U_3
		۴	۴	U_4
۵	۳	۱	۲	U_5

منبع^{۲۱}

¹ <https://www.geeksforgeeks.org/user-based-collaborative-filtering/>

² <https://sokanacademy.com/blog/recommender-system-collaborative-filtering-cf>

توضیحات مهم:

- تمامی مستندات خود شامل گزارش و کدهای خود را در یک فایل فشرده با فرمت zip ذخیره کرده و با عنوان `#StudentId_HW۴.zip` بارگذاری نمایید .
- تمرین بدون گزارش و درک عمیق نسبت به اهداف تمرین، فاقد ارزش می‌باشد و **نمره‌ای به آن تعلق نمی‌یابد**.
- مطابق قوانین دانشگاه هرگونه کپی برداری **ممنوع** می‌باشد و برای شناسایی تقلب در این تمرین از Apache MOSS برای تشخیص مشابهت در کدها استفاده خواهد شد، در صورت مشاهده تشابه، نمره هر دو طرف صفر در نظر گرفته می‌شود.
- در هر مرحله، نتایج خود را تحلیل کنید.
- کدهای خود را برای خوانایی بیشتر کامنت گذاری کنید.
- در تمامی سوال‌ها تنها مجاز به استفاده از کتابخانه‌های `numpy`، `apache spark` و `pandas` می‌باشید. توجه داشته باشید کد می‌بایست توسط شما پیاده سازی شده و از استفاده از کدهای آماده اجتناب کنید.