کاوش داده‌گان انبوه

دکتر سامان هراتی زاده
دانشگاه تهران - دانشکده علوم و فنون
ترم اول سال تحصیلی – زمستان ۱۳۹۹

دستیاران آموزشی
مهدی صادقی (mahdisadeghi@ut.ac.ir)
شیوا پارسا راد (shiva.parsarad@ut.ac.ir)
ایمان برازنده (barazandeh.iman@ut.ac.ir)

مجموعه تمرین پنجم :

شناسایی مجموعه اقلام پرتکرار و گاهی نیز کم تکرار!

مهلت تحویل این تمرین سه شنبه ۳۰ دی ساعت ۲۳:۵۹ می‌باشد.

هدف از این تمرین آشنایی با یکی از تکنیک‌های بنیادی داده کاوی یعنی کاوش مجموعه اقلام پرتکرار است. کاوش مجموعه اقلام پرتکرار یکی از تکنیک‌های اکتشافی بدون ناظر است که برای شناسایی همبستگی‌های با اهمیت بین اقلام درون داده‌گان انبوه استفاده می‌شود. این تکنیک این روزها کاربردهای جذابی در بیوانفورماتیک، امنیت و حریم خصوصی، ساختن سبد سهام[1] و زمینه‌های بسیار دیگر دارد.

۱-(طراحی و پیاده سازی الگوریتم)

داده مورد استفاده در این تمرین شامل سبدهای خرید یک خرده فروشی است. هر سطر از این دیتاست نشانگر یک سبد خرید و در هر سطر تعدادی شناسه محصول وجود دارد که نشانگر محصولات خریداری شده است. در این تمرین شما فقط با product_id و order_id کار خواهید داشت.

الف) با استفاده از الگوریتم Apriori و support=0.03 کالاهای پر تکرار در سبدهای خرید را پیدا کنید.

ب) با استفاده از الگوریتم PCY و support=0.04 کالاهای پر تکرار در سبدهای خرید را پیدا کنید.

[1] Stock Portfolio

۲-(پژوهش)

الف) یکی از بهبودهایی که روی تعریف کلاسیک مجموعه اقلام پرتکرار برای بعضی کاربردها ضروی است، اضافه شدن وزن به اقلام درون مجموعه‌ها است، به عبارت دیگر مجموعه اقلام پرتکرار وزن‌دار[۲] .

چرا وزن؟ بدیهی است که در بسیاری از کاربردهای دنیای واقعی، اقلام درون مجموعه‌ها اهمیت و ارزش یکسانی ندارند. بعنوان مثال، سهام شرکت‌های مختلف درون یه یک سبد سهام را در نظر بگیرید که هر کدام نرخ سود متفاوتی در یک روز معاملاتی دارند. با بررسی یک تحقیق کاربردی معتبر، الگوریتم شناسایی مجموعه اقلام پرتکرار وزن‌دار و همچنین کاربردی که از آن استفاده کرده را با جزئیات کافی به ما نیز معرفی کنید.

ب) شاید برایتان جالب باشد که یک زمینه کاربردی و تحقیقاتی به نام **کاوش مجموعه اقلام کم تکرار**[۳]!!!! نیز وجود دارد. کمی در این مورد تحقیق کرده و الگوریتم‌ها و کاربردهای آن را گزارش کنید.

[2] Weighted Frequent Itemset
[3] Infrequent Itemset Mining

نکات تکمیلی

- از زبان برنامه سازی پایتون کنید.
- کدها با نامگذاری مناسب در پوشه SOURCES قرار گیرد.
- گزارش کامل تمرین را با ساختار مناسب، بدون ابهام و به زبان ساده نوشته و با فرمت PDF در پوشه دیگری قرار دهید. برای نگارش گزارش‌ها نیز امتیاز مثبت و منفی در نظر گرفته شده است، پس برای نگارش گزارش خود وقت مناسبی اختصاص دهید.
- پاسخ تمرین خود را به ایمیل barazandeh.iman@ut.ac.ir ارسال نمایید. هر دو پوشه را در یک فایل ZIP با نام MMDS_STID_FullName_HW# مثلا MMDS_830123451_ImanBarazandeh_HW1 قرار داده و در ایمیلی با همین عنوان نیز ارسال کنید.
- در صورت مشخص شدن هر نوع تقلب انفرادی یا مشترک نمره تمرین صفر در نظر گرفته می‌شود.
- به ازای هر روز تاخیر ۲۵٪ از امتیاز تمرین از دست خواهید داد.
- هر گونه ابهام را با دستیاران آموزشی مطرح کنید.