به نام خدا

پروژه ی دوم درس داده کاوی

پیاده سازی Iceberg Cube

در این تمرین قصد داریم با استفاده از مجموعه داده ی داده شده پیاده سازی Iceberg Cube را انجام دهیم. همان طور که در کلاس ذکر کردیم ، مجموعه داده ی مورد نظر ما دارای ۲۰ نوع آیتم (از آیتم صفر تا آیتم نوزده) است که هر کدام را می توان به منزله ی یک بعد در این تمرین در نظر گرفت.

Iceberg Cube مورد نظر برای پیاده سازی باید دارای ۵ بعد باشد و از این رو آیتم های صفر تا چهار را به عنوان بعدهای مطلوب خود در نظر بگیرید. بدیهی است که مابقی آیتم ها در هر تراکنش بلا استفاده خواهند شد و نیازی به آن ها نخواهیم داشت. در نظر داشته باشید که با احتساب پنج بعد اول تعداد زیادی از تراکنش های مجموعه داده ی ما حذف خواهند شد چون شامل آیتم های صفر تا چهار نیستند.

می دانیم که هر Cube از تعدای Cuboid تشکیل شده است و اگر دارای n بعد باشیم تعداد کل Cuboid ها برابر با 2^n خواهد بود. کوچکترین Cuboid ، Apex نام دارد و به اصطلاح صفر بعدی می باشدکه حاصل تجمیع (Aggregate) تمامی ابعاد است و پس از آن Cuboid های یک بعدی تا n بعدی (در این تمرین ۵ بعدی) را خواهیم داشت که حالت n بعدی ، Base Cuboid نام دارد.

برای تفهیم بهتر مطلب یکی از Cuboid های پیاده سازی شده را در شکل زیر مشاهده می نمایید:

	A	B	C	D	E	Aggregate
▶	*	*	*	0	0	30173
	*	*	*	0	1	21454
	*	*	*	1	0	25608
*	*	*	*	1	1	13688

جدول فوق برای یک Cuboid دو بعدی ساخته شده است که در آن ویژگی های اول تا سوم مقادیر * را اختیار نموده اند. قابل ذکر است که در این شکل هر کدام از ستون ها نمایانگر یکی از پنج ویژگی (آیتم صفر تا چهار) می باشد و هر کدام از ویژگی ها می توانند مقدار صفر یا یک (به معنای عدم حضور یا حضور در تراکنش ها) را اختیار کنند.

ستون آخر که Aggregate نام دارد جمع تعداد تمامی تراکنش ها با شرایط مورد نظر است. برای مثال سطر اول جدول فوق بیان می کند که در مجموعه داده ی ما تعداد تراکنش هایی که فاقد ویژگی چهارم و پنجم باشند (آیتم سه و آیتم چهار) برابر با ۳۰۱۷۳ می باشد. بدیهی است که در این حالت مقدار ویژگی های اول تا سوم مد نظر ما نمی باشند (به اصطلاح Don’t Care هستند) و می توانند در تراکنش حضور داشته باشند یا حضور نداشته باشند.

با توجه به توضیحات داده شده و با استفاده از مجموعه داده مورد نظر باید تمامی ۳۲ عدد Cuboid به صورت Iceberg Cube پیاده سازی شوند. یک پیشنهاد ساده برای پیاده سازی این مسئله این است که در ابتدا تمامی ۳۲ عدد Cuboid به صورت ساده ساخته شوند و سپس در ادامه با قرار دادن یک حد آستانه (Min_Sup) کلی بر روی ۳۲ عدد Cuboid ، تمامی سلول هایی (Cell) که مقدار Measure برای آن ها (در این مسئله مقدار تجمیع شده یا Aggregate) از حد آستانه کمتر است ، حذف شوند.

نکات مربوط به پیاده سازی و گزارش:

- محدودیتی برای زبان پیاده سازی وجود ندارد و دانشجویان می توانند با زبان دلخواه پیاده سازی مربوطه را انجام دهند.
- تمامی ۳۲ عددCuboid باید به صورت کامل (به صورت Iceberg) پیاده سازی شوند
- در صورت وجود شباهت میان پیاده سازی های انجام شده نمره ی تمرین میان افراد خاطی تقسیم خواهد شد
- خروجی برنامه برای تمامی ۳۲ عدد Cuboid باید در گزارش به صورت کامل آورده شود و نتایج پیاده سازی و نتایج موجود در گزارش باید با یکدیگر مطابقت داشته باشند
- از قرار دادن کد به صورت صرف در گزارش خودداری کنید
- گزارش باید مراحل پیاده سازی و نحوه ی عملکرد کلی شما را شامل شود بگونه ای که در صورت عدم حضور دانشجو نحوه ی پیاده سازی کاملا درک گردد
- گزارش شما باید به فرمت PDF باشد
- تمرین شما باید شامل دو پوشه باشد به نام های Report و Source Code باشد که در پوشه ی اول گزارش مربوطه و در پوشه ی دوم پیاده سازی انجام شده قرار داده می شود. فایل نهایی، فایل Zip شده ی این دو پوشه است که باید با فرمت زیر نام گذاری شود:

DataMining_Exc#1_Your Name_Your Student ID

پاینده باشید