به نام خدا

دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

درس پردازش داده‌های حجیم

استاد حقیرچهرقانی

تمرین اول

علیرضا مازوچی

۴۰۰۱۳۱۰۷۵

بخش اول: سوالات تشریحی

سوال ۱

الف) با بررسی لیست تراکنش‌ها به نام بازی‌های زیر بر می‌خوریم:

- Horizon Forbidden West
- GTA V
- Gran Turismo 7
- Ghost of Tsushima
- Dying Light 2
- MAFIA: Trilogy
- FIFA 22
- The Last of Us Part II
- Far Cry 6

تعداد این‌ها ۹ تاست و یک مجموعه می‌تواند حداکثر شامل تمام اعضا باشد؛ پس حداکثر می‌توان یک مجموعه‌ی ۹تایی تشکیل داد؛ اگر support برابر با صفر باشد.

ب) همانطور که در قسمت قبل گفته شد، ۹ آیتم وجود دارد. هر مجموعه سه‌تایی باید انتخاب ۳ عضو از این‌ها باشد. پس بیشینه تعداد مجموعه‌های سه تایی $\binom{9}{3} = 84$ است که برای support برابر با صفر رخ خواهد داد.

ج) بدیهی است که یک مجموعه دارای Support کمتر مساوی‌ای از زیرمجموعه‌های خود است. پس برای پیدا کردن مجموعه با اندازه حداقل ۲ که بیشترین Support را دارد لازم و کافی است که تمام مجموعه‌های با اندازه دقیقا ۲ را بررسی کنیم. به نظر می‌رسد مجموعه‌ی {FIFA 22, The Last of Us Part II} با ۵ بار تکرار بیشترین فراوانی را در میان نامزدها دارد.

د) فرمول Confidence به شرح زیر است:

$$Conf(I \to j) = \frac{Support(I \cup j)}{Support\ (I)}$$

صورت این کسر برای هر دو قانون $A \to B$ و $B \to A$ یکسان است. پس برای برابری Confidence کافی است تا Support آیتم A با Support آیتم B برابر باشد. به عنوان مثال GTA V و The Last of Us Part II هر دو ۸ بار تکرار شده‌اند.

سوال ۲

یک مجموعه جز مرز منفی است اگر و فقط اگر خودش غیرفراوان باشد ولی تمام زیرمجموعه‌های بلافاصله‌اش فراوان باشد. با این تعریف می‌توان اعضای زیر را جز مرز منفی قرار داد:

$$\{ABD, BCD, ACD, AE, BE, CE, DE, F\}$$

سوال ۳

الف) ۹ جفت کلید-مقدار زیر تشکیل می‌شود:

$$(10, (R, 1)), (10, (R, 2)), (11, (R, 3)), (10, (R, 4))$$

$$(10, (S, 20)), (11, (S, 21)), (12, (S, 22)), (10, (S, 23)), (11, (S, 24))$$

ب) جفت کلید-مقدار با توجه به مقادیر کلید در مرحله grouped by به سه گروه تقسیم می‌شوند:

$$(10, (R, 1)), (10, (R, 2)), (10, (R, 4)), (10, (S, 20)), (10, (S, 23))$$

$$(11, (R, 3)), (11, (S, 21)), (11, (S, 24))$$

$$(12, (S, 22))$$

بنابراین حداقل به سه Reducer برای مرحله بعد نیاز خواهیم داشت.

ج) خروجی هر یک از Reducer ها عبارت است از:

$$(10, (R, 1)), (10, (R, 2)), (10, (R, 4)), (10, (S, 20)), (10, (S, 23)) \rightarrow (1, 10, 20), (1, 10, 23), (2, 10, 20), (2, 10, 23), (4, 10, 20), (4, 10, 23)$$

$$(11, (R, 3)), (11, (S, 21)), (11, (S, 24)) \rightarrow (3, 11, 21), (3, 11, 24)$$

$(12, (S, 22)) \rightarrow -$

لذا هشت خروجی زیر را خواهیم داشت:

$(1, 10, 20), (1, 10, 23), (2, 10, 20), (2, 10, 23), (4, 10, 20), (4, 10, 23), (3, 11, 21)$

$, (3, 11, 24)$

سوال ۴

در صورت سوال یک ماتریس جایگشت داده شده است و من هم از همان استفاده می‌کنم (اگر قرار به انتخاب اختیاری جایگشت باشد) ماتریس M بدین ترتیب حاصل می‌شود:

۱	۳	۲	۱	۱
۵	۱	۱	۳	۱
۱	۴	۳	۱	۱

برای سنجش شباهت، از سه جفت ستون اول (از سمت چپ) استفاده می‌کنم:

شباهت واقعی ستون اول و دوم: $\frac{۲}{۷}$

شباهت هش ستون اول و دوم: ۰

شباهت واقعی ستون اول و سوم: $\frac{۱}{۸}$

شباهت هش ستون اول و سوم: ۰

شباهت واقعی ستون دوم و سوم: $\frac{۴}{۶}$

شباهت هش ستون دوم و سوم: $\frac{۱}{۳}$