Department of
Computer Engineering

دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

تمرین اول پردازش داده‌های حجیم

استاد درس: دکتر چهرقانی

مهلت ارسال تمرین: ۱۴۰۱/۱/۱۵

بهار ۱۴۰۱

بخش اول – سوالات تشریحی

Association Rules

۱ جدولی که در زیر مشاهده می‌کنید، سبدهای خرید مشتریان یک فروشگاه ارائه‌دهنده‌ی کنسول‌های بازی است. با توجه به این جدول به سوالات زیر پاسخ دهید.

شماره تراکنش	دیسک‌های بازی خریداری شده
۱	MAFIA: Trilogy, FIFA 22, The Last of Us Part II
۲	FIFA 22, MAFIA: Trilogy, The Last of Us Part II, Far Cry 6
۳	FIFA 22, Horizon Forbidden West
۴	Dying Light 2, FIFA 22, MAFIA: Trilogy
۵	Horizon Forbidden West, The Last of Us Part II
۶	FIFA 22, Horizon Forbidden West
۷	Horizon Forbidden West, Far Cry 6
۸	FIFA 22, Horizon Forbidden West, The Last of Us Part II
۹	Dying Light 2, FIFA 22, Horizon Forbidden West
۱۰	FIFA 22, MAFIA: Trilogy
۱۱	GTA V, Ghost of Tsushima, The Last of Us Part II, Far Cry 6
۱۲	GTA V, Ghost of Tsushima, The Last of Us Part II, Far Cry 6
۱۳	GTA V, Gran Turismo 7
۱۴	FIFA 22, GTA V, Ghost of Tsushima
۱۵	FIFA 22, Gran Turismo 7, The Last of Us Part II
۱۶	GTA V, Gran Turismo 7, Far Cry 6
۱۷	FIFA 22, Gran Turismo 7, The Last of Us Part II
۱۸	FIFA 22, GTA V, Gran Turismo 7
۱۹	FIFA 22, GTA V, Gran Turismo 7
۲۰	Horizon Forbidden West, MAFIA: Trilogy, GTA V, Far Cry 6

الف) بیشترین اندازه‌ی ItemSet ای که میتوانیم استخراج کنیم، چقدر است؟

ب) عبارتی برای بیشینه‌ی تعداد ItemSet هایی که با اندازه‌ی ۳ که می‌توانیم از این اطلاعات استخراج کنیم بنویسید.

ج) کدام ItemSet با اندازه‌ی حداقل ۲، بیشترین support را دارد؟

د) آیا جفت association rule هایی به فرم $A \rightarrow B$ و $B \rightarrow A$ که confidence یکسانی داشته باشند، در این مجموعه‌ها وجود دارد؟ در صورتی که وجود دارد، یکی از آن‌ها را ذکر کنید.

۲ فرض کنید شش item داریم: A, B, C, D, E, F. پس از یکبار اجرای الگوریتم Tivonen به این نتیجه رسیده‌ایم که maximal frequent itemset ها برابر با مجموعه‌های ABC, AD, BD, CD, E هستند. مرز منفی[۱] این مجموعه‌ها را به دست آورید.

Map Reduce

۳ رابطه‌ی R(A,B) شامل چهار تاپل[۲] مقابل است: {(1,10), (2,10), (3,11), (4,10)}. رابطه‌ی S(B,C) نیز از ۵ تاپل مقابل تشکیل شده است: {(10,20), (11,21), (12,22), (10,23), (11,24)}. اگر با استفاده از الگوریتم MapReduce ی که در اسلایدهای درس با آن آشنا شدید، join دو رابطه‌ی R و S را به دست بیاوریم:

الف) توسط همه‌ی mapper ها چند جفت کلید-مقدار ایجاد می‌شود؟ آن‌ها را بنویسید.

ب) حداقل اندازه‌ی reducer برای اجرای الگوریتم روی این داده‌ها چند است؟

ج) تعداد تاپل‌های خروجی چند است؟ آن‌ها را بنویسید.

Locality-sensitive hashing

۴ ماتریس ورودی زیر را در نظر بگیرید. برای این ماتریس ابتدا جایگشت متفاوت در نظر بگیرید، سپس از طریق آن ماتریس M (Signature matrix) را به دست آورید. در نهایت شباهت ماتریس ورودی و ماتریس M را برای ۳ جفت ستون دلخواه بسنجید (از معیار شباهت Jaccard استفاده کنید).

7	1	4
2	3	3
5	5	1
1	8	2
8	7	5
4	6	7
6	2	6
3	4	8

0	1	1	0	1
0	0	1	1	0
1	0	0	1	1
1	0	0	1	1
1	1	0	1	1
1	1	1	0	0
0	1	1	0	0
0	1	1	0	1

[۱] Negative border

[۲] Tuple

بخش دوم – سوالات پیاده‌سازی

Map Reduce

۱ در این بخش می‌خواهیم با پیاده‌سازی الگوی برنامه‌نویسی Map-Reduce آشنا شویم. توصیه می‌شود که برای این بخش از تمرین از PySpark استفاده کنید، چرا که به راحتی می‌توانید آن را در پلتفرم گوگل کولب نصب کرده و با استفاده از آن کدهای خود را اجرا کنید.

مجموعه‌داده‌ی مورد استفاده در این سوال در فایل dataset1.txt قرار گرفته است. هر سطر از این مجموعه‌داده شامل آیدی یک کاربر و با یک tab آیدی کاربرانی که با آن دوست هستند، آمده است. می‌خواهیم برای هر کاربر، از میان کسانی که در حال حاضر با آن فرد دوست نیستند، کاربرانی که با آن‌ها بیشترین دوست مشترک را دارد، به عنوان افرادی که ممکن است بشناسند پیشنهاد دهیم. در پیاده‌سازی و گزارش خود موارد زیر را در نظر بگیرید:

۱. کدها حتما با استفاده از الگوی برنامه‌نویسی Map-Reduce نوشته شوند.
۲. در خروجی برای کاربران با آیدی‌های ۹۸، ۱۳۵، ۱۱۷، ۹۱۱، ۸۸۰۴ تعداد ۱۰ دوست پیشنهادی را چاپ کنید.
۳. در گزارش علاوه بر قرار دادن خروجی الگوریتم برای کاربران ذکر شده، کد نوشته شده را توضیح دهید. این توضیح کافی است شامل بخش مربوط به توابع map و reduce باشد.

کاربرد Association Rules

۲ در این تمرین می‌خواهیم با کاربرد الگوریتم A-priory برای توصیه اقلام آشنا شویم. مجموعه‌داده‌ی مورد استفاده در این بخش در فایل games_library.txt قرار گرفته است. هر سطر از این مجموعه‌داده شامل لیستی از بازی‌های محبوب یک گیمر است که با استفاده از space از یکدیگر جدا شده‌اند. می‌خواهیم یک توصیه‌گر بنویسیم که بازی‌هایی را که توسط گیمرهای زیادی به صورت توام محبوب بوده‌اند، پیشنهاد دهد.

۱. جفت بازی‌های (X,Y) را که دارای support حداقل ۱۰۰ هستند، بیابید. برای همه‌ی این جفت‌ها، مقدار confidence مربوط به قوانین $X \to Y,\ Y \to X$ را محاسبه نمایید. در انتها به ترتیب نزولی آنها را مرتب کرده و ۵ قانون اول را گزارش دهید.
۲. مجموعه‌های سه‌تایی بازی‌های (X,Y, Z) را که دارای support حداقل ۱۰۰ هستند، بیابید. برای همه‌ی این سه‌تایی‌ها، مقدار confidence مربوط به قوانین $(X,Y) \to Z,\ (X,Z) \to Y\ ,(Y,Z) \to X$ را محاسبه نمایید. در انتها به ترتیب نزولی آنها را مرتب کرده و ۵ قانون اول را گزارش دهید.

توجه: در صورت برابری، از ترتیب lexicographical استفاده نمایید.

۳ در این بخش می‌خواهیم با کاربرد LSH برای یافتن تقریبی نزدیک‌ترین همسایه‌ها، آشنا شویم. در ابتدا مفاهیم مرتبط با LSH برای حل این مسئله را مرور خواهیم کرد. فرض کنید مجموعه‌داده‌ی A شامل n نقطه است که در metric space با معیار فاصله‌ی $d(.,.)$ قرار گرفته‌اند. ثابت c را به عنوان عددی بزرگتر از یک در نظر میگیریم. با داشتن این فرض‌ها، مسئله‌ی نزدیک‌ترین همسایه‌ها با تقریب (c, λ) را به صورت زیر می‌توان تعریف کرد:

۱. یک نقطه دلخواه z به ما داده شده است.
۲. فرض می‌کنیم نقطه‌ای به نام x در مجموعه‌ی داده وجود دارد، طوری که $d(x,z) \leq \lambda$ است.
۳. الگوریتم تقریبی نقطه‌ای به نام x' بر می‌گرداند، طوری که $d(x',z) \leq c\lambda$ است.
۴. در این صورت، پارامتر c نشان‌دهنده‌ی فاکتور تقریب بیشینه برای این مسئله است.

در این مسئله فرض می‌کنیم، LSH family H از توابع هشی تشکیل شده است که برای معیار فاصله‌ی $d(.,.)$ $(\lambda, c\lambda, p_1, p_2)$ $sensitive$ هستند. تابع G را نیز به صورت زیر تعریف می‌کنیم:

$$G = H^k = \{g = (h_1, \dots, h_k) | h_i \in H, \forall\, 1 \leq i \leq k\}, \quad k \log_{\left(\frac{1}{p_2}\right)}(n)$$

فرایند زیر را برای حل مسئله در نظر می‌گیریم:

۱. انتخاب $L = n^{\rho}$ عضو تصادفی $g_1, \dots, g_L$ از G، طوری که $\rho = \frac{\log\left(\frac{1}{p_1}\right)}{\log\left(\frac{1}{p_2}\right)}$
۲. هش کردن همه‌ی نقاط داده به همراه نقطه‌ی پرس‌وجو با استفاده از $g_i (1 \leq i \leq L)$.
۳. بازگرداندن حداکثر 3L از نقاط داده به صورت کاملا تصادفی، از مجموعه‌ی L تا باکتی که نقطه‌ی پرس‌وجو به آن‌ها هش شده است.
۴. از میان نقاطی که در گام سوم انتخاب شدند، آن نقطه‌ای را که به نقطه‌ی پرس‌وجو از همه نزدیک‌تر است یک تقریب (c, λ) از نزدیک‌ترین همسایه است.

تا به اینجای کار با نحوه‌ی تعریف و حل مسئله‌ی یافتن نزدیک‌ترین همسایه به صورت تقریبی از طریق LSH، برای فهم کدی که در فایل lsh.py در اختیارتان قرار گرفته است، آشنا شدیم.

مجموعه‌داده‌ی مورد استفاده در این سوال در فایل patches.csv قرار گرفته است. هر سطر از این مجموعه داده یک تصویر ۲۰×۲۰ است، که توسط یک بردار ۴۰۰ بعدی بازنمایی شده است. در این سوال می‌خواهیم میزان کارایی تقریب با استفاده از LSH را با روش جستجوی خطی مقایسه نماییم. از معیار فاصله‌ی L_1 برای تعیین شباهت میان تصاویر استفاده می‌شود.

توضیحات کد: در کد اولیه ارائه شده در این تمرین، مواردی که میبایست توسط شما تکمیل گردند، توسط Todo مشخص شده‌اند. شما میبایست از توابع راه اندازی و جستجو استفاده کرده و تابع جستجوی خطی خود را پیاده‌سازی نمایید. می‌توانید از پارامترهای

پیش فرض برای این تمرین که برابرند با L=10، k=24 استفاده نمایید؛ هر چند دست شما برای استفاده از هر مقدار دیگری برای این تمرین باز است مادامی که دلایل خود را برای انتخاب آن‌ها ذکر نمایید.

الف) به صورت خلاصه، توضیح دهید که عملکرد و نحوه پیاده‌سازی تابع lsh_search در فایل lsh.py به چه صورتی است.

ب) برای هر یک از اندیس‌های $\{100, 199, 300, 399, 500, 599, 700, 799, 900, 999\}$، سه مورد نزدیک‌ترین همسایه را با استفاده از هر دو روش LSH و جستجوی خطی بدست آورید. میانگین زمان جستجوی خود را برای هر یک از این دو مورد ذکر کرده و با هم مقایسه نمایید.

ج) با فرض اینکه $\{z_j \mid 1 \leq j \leq 10\}$ مجموعه تصاویر مورد نظر ما که در آن z_j تصویری است از ستون $100j$ باشد و $\{x_{ij}^*\}_{i=1}^3$ سه نزدیک‌ترین همسایه درست z_j باشند که از روش جستجوی خطی بدست آمده اند، میزان خطای زیر را گزارش دهید.

$$error = \frac{1}{10}\sum_{j=1}^{10}\frac{\sum_{i=1}^{3} d(x_{ij}, z_j)}{\sum_{i=1}^{3} d(x_{ij}^*, z_j)}$$

د) نمودار مقدار خطا را یکبار به صورت تابعی از L (L=10, 12, …, 20) و با ثابت نگاه داشتن مقدار k برابر با k=24) و یکبار به صورت تابعی از k (k=16, 18, 20, 22, 24) و با ثابت نگاه داشتن مقدار L برابر با L=10) رسم نموده، مقادیر را گزارش نمایید. به طور خلاصه نمودارها را تحلیل نمایید.

ه) با استفاده از هر یک از دو روش مورد مقایسه در این سوال، ۱۰ همسایه‌ی نزدیک برای تصویر موجود در ستون صدم را یافته و به همراه خود تصویر رسم نمایید. در انتها عملکرد دو روش را مقایسه نمایید.

نکات مربوط به تحویل تمرین

- مجموعه‌های داده و فایل‌های مرتبط با تمرین را می‌توانید از طریق لینک زیر یا سامانه درس دانلود کنید:
- https://drive.google.com/file/d/1-OCBGBtKoY_PadKHcXDyWxHQ2BS8Nulo/view?usp=sharing

- **کد:** دقت داشته باشید که استفاده از کتابخانه‌های آماده برای بخش‌های خواسته‌شده، در پیاده‌سازی مجاز نیست.
- **گزارش**: ملاک اصلی انجام تمرین گزارش آن است و ارسال کد بدون گزارش فاقد ارزش است. برای این تمرین یک فایل گزارش در قالب pdf تهیه کنید و در آن برای هر سوال، تصاویر ورودی، تصاویر خروجی و توضیحات مربوط به آن را ذکر کنید. سعی کنید توضیحات کامل و جامعی تهیه کنید.
- **تذکر ۱:** مطابق قوانین دانشگاه هر نوع کپی برداری و اشتراک کار دانشجویان غیر مجاز بوده و شدیدا برخورد خواهد شد.
- **تذکر ۲:** برای سهولت در انجام تمرینات، توصیه می‌شود که پلتفرم کولب گوگل استفاده نمایید.
- **تذکر ۳:** در نظر داشته باشید کد های شما باید قابلیت اجرا در هنگام ارائه را داشته باشند. همچنین بر روی کدهای خود مسلط باشید.
- **کانال درس:** اطلاعیه‌های مربوط به درس کانال زیر قرار می‌گیرند:
- https://t.me/+cLCmyX2sIPVjN2I0
- **راهنمایی:** در صورت نیاز میتوانید سوالات خود را در خصوص تمرینات، از طریق ایمیل زیر بپرسید.
- E-mail: bigdata.aut.1401@gmail.com

- **ارسال:** پاسخ سوالات تشریحی، فایل های کد و گزارش خود را در یک فایل فشرده قرار داده و با نام با فرمت HW1_StudentID ارسال نمایید.