دانشکده مهندسی برق

تحلیل داده‌های حجیم

تمرین سری دوم

استاد درس: دکتر ایمان غلامپور

طراح تمرین: حسین شریفی

بخش تئوری و تحقیق:

سوال اول)

متریک‌های زیر درنظر بگیرید:

$$S(B) = \frac{support(B)}{|Baskets|}$$

$$lift(A \rightarrow B) = \frac{conf(A \rightarrow B)}{S(B)}$$

$$conv(A \rightarrow B) = \frac{1 - S(B)}{1 - conf(A \rightarrow B)}$$

الف) توضیح دهید $lift(A \rightarrow B)$ چه رخدادی را توصیف می‌کند و مفهوم آن را شرح دهید.

ب) توضیح دهید $conv(A \rightarrow B)$ چه رخدادی را توصیف می‌کند و مفهوم آن را شرح دهید.

ج) ضعف تعریف confidence نسبت به دو متریک بالا چیست؟ با ذکر مثال شرح دهید.

سوال دوم)

فرض کنید ۱۰۰ آیتم داریم که با اعداد ۱ تا ۱۰۰ شماره‌گذاری شده‌اند و همچنین ۱۰۰ سبد داریم که آنها نیز با اعداد ۱ تا ۱۰۰ شماره‌گذاری شده‌اند. قانون قرارگیری آیتم‌ها در سبدها به این صورت است: آیتم i در سبد b قرار دارد اگر i مقسوم‌علیه b باشد.

به عنوان مثال:

- آیتم ۱ در تمام ۱۰۰ سبد وجود دارد.
- سبد شماره ۱۲ شامل آیتم های $\{۱،۲،۳،۴،۶،۱۲\}$ می‌باشد.

الف) اگر support threshold برابر با ۵ باشد، کدام آیتم‌ها Frequent هستند؟

ب) به ازای support threshold برابر با ۵، کدام جفت آیتم‌ها پرتکرار هستند؟

ج) مجموع اندازه تمام سبدها چقدر است؟

د) برای association rule های زیر، مقادیر confidence و interest را محاسبه کنید.

{۶،۹}=>۳

{۲،۴،۵}=>۳

سوال سوم)

در یک مجموعه داده، I تعداد آیتم‌ها، B تعداد سبدها و P تعداد آیتم‌های موجود در هر سبد است. با توجه به این مقادیر پاسخ دهید:

الف) فضای حافظه مورد نیاز برای ذخیره‌سازی تمام جفت آیتم‌ها با استفاده از روش triangular-matrix، با فرض ۴ بایت برای هر عنصر آرایه، چقدر خواهد بود؟

(ب) بیشترین تعداد جفت‌هایی که شمارش آن‌ها غیرصفر است، چه مقدار است؟

(ج) تحت چه شرایطی روش ذخیره سازی triples در مقایسه با حالت (الف) فضای کمتری مصرف می‌کند؟

سوال چهارم)

Element	S1	S2	S3	S4
0	1	0	1	0
1	0	1	1	0
2	1	1	0	1
3	0	0	1	1
4	1	0	0	1
5	0	1	1	0

با توجه به جدول بالا به سوالات زیر پاسخ دهید:

الف) برای هر ستون از ماتریس، minhash signature را با توجه به توابع hash زیر حساب کنید:

$h_1(x) = (3x + 1)\%6$

$h_2(x) = (4x + 3)\%6$

$h_3(x) = (5x + 2)\%6$

ب) تعیین کنید کدام یک از توابع hash بالا، جایگشت های معتبر هستند.

ج) با استفاده از minhash signature محاسبه شده، معیارJaccard تخمین زده شده را با Jaccard واقعی مقایسه کنید. (یک جدول ایجاد کنید و شباهت واقعی و تخمین زده شده را برای هر جفت ستون نمایش دهید.) برای بهبود تخمین چه راهکاری پیشنهاد می‌دهید؟

بخش عملی:

سوال اول)

در این سوال، می‌خواهیم مجموعه‌های سه کلمه‌ای از کلمات پرتکرار که در متن abstract مقالات دیتاست قبلی در کنار هم ظاهر شده‌اند را محاسبه و ارزیابی کنیم. لینک دانلود دیتاست

توجه کنید که باید از متن clean شده استفاده کنید و موقعیت سه کلمه بی‌اهمیت می‌باشد. در گزارش خود همه موارد مانند انتخاب support threshold و نرخ کاهش آن، نحوه ذخیره pair ها، توابع hash استفاده شده و... را به همراه دلایل انتخاب و طراحی آنها شرح دهید.

الف) با استفاده از الگوریتم A-priori، مجموعه‌های سه کلمه‌ای پرتکرار را پیدا و الگوریتم خود را شرح دهید و نتایج را گزارش کنید.

ب) با استفاده از الگوریتم PCY، مجموعه‌های سه کلمه‌ای پرتکرار را پیدا و الگوریتم خود را شرح دهید و نتایج را گزارش کنید.(می‌توانید از Extension های این الگوریتم هم استفاده کنید.)

ج) از معیار Jaccard طبق رابطه زیر، برای ارزیابی مدل PCY استفاده کنید. برای بهبود آن چه روشی را توصیه می‌کنید؟

$$Jaccard = \frac{TP}{TP + FP + FN} = \frac{|P \cap A|}{|P \cup A|}$$

سوال دوم)

در این قسمت قصد داریم با کمک الگوریتم LSH کدی را پیاده سازی کنیم تا مقاله‌های مشابه را بر اساس متن آن ها پیدا کنیم. الگوریتم را به نحوی پیاده سازی کنید که در ورودی، id یک مقاله داده شود و تمام مقاله‌های مشابه با آن نمایش داده شود. نحوه تنظیم پارامترهای مدل(مانند bوr) با visualization مناسب و شرح الگوریتم گزارش کنید.