کاوش داده‌گان انبوه

تمرین «دو»

دکتر سامان هراتی‌زاده

دانشگاه تهران - دانشکده سامانه‌های هوشمند

نیم‌سال دوم ۱۴۰۳-۱۴۰۲

دستیاران آموزشی

محمد راشدی

آرمین رحیمی دره‌باغ

ددلاین: ساعت ۵۹:۲۳ | ۱۴۰۳/۰۴/۲۰

راهنمای تحویل

قبل از پاسخ دادن به پرسش‌ها، موارد زیر را با دقت مطالعه نمایید:

- **کیفیت گزارش شما در فرآیند تصحیح از اهمیت ویژه‌ای برخوردار است**؛ بنابراین، لطفا تمامی نکات و فرض‌هایی را که در پیاده‌سازی‌ها و محاسبات خود در نظر می‌گیرید در گزارش ذکر کنید.
- در گزارش خود برای شکل‌ها زیرنویس و برای جدول‌ها بالانویس در نظر بگیرید.
- **تحلیل نتایج الزامی می‌باشد، حتی اگر در صورت سوال اشاره‌ای به آن نشده باشد.**
- **کدهای ارسالی می‌بایست قابلیت اجرای دوباره داشته باشند، با این حال، دستیاران آموزشی ملزم به اجرای کدهای شما نیستند**؛ بنابراین، هرگونه نتیجه و یا تحلیلی که در صورت پرسش از شما خواسته شده را به طور واضح و کامل در گزارش بیاورید. در صورت عدم رعایت این مورد، بدیهی است که از نمره تمرین کسر می‌شود.
- **در صورت استفاده از Jupyter لازم است تا تمامی کد اجرا شود و خروجی هر سلول حتما در این فایل ارسالی شما ذخیره شده باشد در غیر اینصورت ورودی‌ها و خروجی‌ها متناظر می‌بایست در گزارش آورده شوند.** بنابراین برای مثال اگر خروجی سلولی یک نمودار است که در گزارش آورده‌اید، این نمودار باید هم در گزارش هم در نوت‌بوک کدها وجود داشته باشد.
- **با این که بحث در مورد تمرین‌ها منعی ندارد اما راه‌حل شما می‌بایست توسط شما (و فقط شما) باشد. همچنین، تمامی مطالب جانبی در گزارش باید رفرنس داده شود. یادآوری می‌شود که عدم صداقت علمی[1] عواقب شدیدی را به همراه دارد.**
- **استفاده از کدهای آماده برای تمرین‌ها به هیچ وجه مجاز نیست.**
- **در صورت مشاهده‌ی تقلب امتیاز تمامی افراد شرکت‌کننده در آن، به میزان بارم سوال نمره منفی لحاظ می‌شود.**
- لطفا گزارش، کدها و سایر ضمایم را به در یک پوشه با نام زیر قرار داده و آن را فشرده سازید، سپس در سامانه‌ی **Elearn** بارگذاری نمایید:

HW[Number]_[Lastname]_[StudentNumber].zip

---

[1] Academic dishonesty

کاوش داده‌گان انبوه

تمرین «دو»

دکتر سامان هراتی‌زاده

دانشگاه تهران - دانشکده سامانه‌های هوشمند

نیم‌سال دوم ۱۴۰۳-۱۴۰۲

دستیاران آموزشی

محمد راشدی

آرمین رحیمی دره‌باغ

ددلاین: ساعت ۵۹:۲۳ | ۱۴۰۳/۰۴/۲۰

- در صورت ارائه ندادن تمرین، نمره تمرین صفر در نظر گرفته خواهد شد.

در صورت وجود سوال، ابهام و یا درخواست راهنمایی با دستیاران آموزشی مرتبط با هر پرسش از طریق ایمیل‌های آورده شده در سربرگ در ارتباط باشید.

کاوش داده‌گان انبوه

تمرین «دو»

دکتر سامان هراتی‌زاده

دانشگاه تهران - دانشکده سامانه‌های هوشمند

نیم‌سال دوم ۱۴۰۳-۱۴۰۲

دستیاران آموزشی

محمد راشدی

آرمین رحیمی دره‌باغ

ددلاین: ساعت ۵۹:۲۳ | ۱۴۰۳/۰۴/۲۰

فهرست

کاوش داده‌گان انبوه

تمرین «دو»

دکتر سامان هراتی‌زاده
دانشگاه تهران - دانشکده سامانه‌های هوشمند
نیم‌سال دوم ۱۴۰۳-۱۴۰۲

دستیاران آموزشی
محمد راشدی
آرمین رحیمی دره‌باغ

ددلاین: ساعت ۵۹:۲۳ | ۱۴۰۳/۰۴/۲۰

۱ – فیلتر بلوم[۱]

هدف این تمرین استفاده از مفهوم فیلتر بلوم برای پیاده‌سازی یک سیستم احراز هویت می‌باشد. استفاده از فیلتر بلوم به ما کمک می‌کند که در مصرف حافظه و زمان پردازش صرفه‌جویی کنیم. همچنین می‌خواهیم که پیاده‌سازی سیستم به صورت client-server باشد;یعنی کلاینت استریم داده‌ها را بخواند و آن‌ها را برای احراز هویت به سرور بفرستد.

داده‌های ما در این تمرین لیستی ۲۰۰۰۰تایی از کدهای ملی است که در فایل unique_ids.csv و در پوشه q1 قرار داده شده است. توجه کنید که کد شما باید ورود داده‌ها را به صورت جریان شبیه‌سازی کنید و کدهای ملی یکی یکی خوانده شوند.

(الف) به صورت تصادفی هزار کد ملی مجاز را از لیست کدهای ملی مجاز انتخاب کنید. همچنین ۱۰۰۰ کد ملی غیرمجاز نیز تولید کنید. سپس تمام ۲۰۰۰ کد ملی را با تعدادی تابع هش(دلخواه) هش کنید. سپس به صورت client-server فیلتر بلوم را پیاده‌سازی کنید. در پایان تعداد ورودی‌های مجاز و غیرمجاز را چاپ کنید.

(ب) افزایش یا کاهش تعداد توابع هش و اندازه بردار بیتی چه اثری بر تعداد ورودی‌های غیرمجاز می‌گذارد؟ با استفاده از مقادیر مختلف برای تعداد توابع هش و اندازه بردار بیتی، درستی گذاره‌های خود را بررسی کنید.

(پ) اگر بخواهیم در روش فیلتر بلوم احتمال رخداد false positiveها برابر 1% باشد، چه کارهایی می‌توانیم انجام دهیم؟

---

[۱] Bloom Filter

۲ – شمارش کلمات متمایز با الگوریتم Flajolet-Martin

در این تمرین شما باید با استفاده از الگوریتم Flajolet-Martin تعداد تقریبی کلمات متمایز در یک استریم داده را بشمارید. دیتاست استفاده شده برای این سوال یک فایل متنی حاوی تمام آثار شکسپیر می‌باشد که در پوشه q2 قرار داده شده است. برحسب نیاز پردازش‌های لازم را بر روی متن انجام دهید(حذف کلمات نگارشی و ...). توجه کنید که پاسخ شما باید یک استریم را شبیه‌سازی کند به این معنی که شما در هر لحظه باید تنها یک کلمه را پردازش کنید. طول هش شما باید ۲۴ بیت باشد و از ۳۵ تابع هش استفاده کنید.

(الف) الگوریتم Flajolet-Martin را بر روی داده اجرا کنید. پس از اجرای الگوریتم شما باید ۳۵ تخمین از تعداد کلمات متمایز داشته باشید. حال این تخمین‌ها به روش‌های مختلف مانند میانگین، میانه و یا ترکیب آن‌ها، گروه‌بندی کنید. کدام روش بهترین تخمین را برای ما به ارمغان خواهد آورد؟

(ب) طول رشته بیتی و تعداد توابع هش چه تاثیری بر عملکرد الگوریتم دارند؟ نتیجه را در گزارش خود ذکر کنید.

۳ - مجموعه آیتم‌های پرتکرار[۱] و قوانین انجمنی[۲]

خرده‌فروشان از قوانین انجمنی برای تحلیل سبد خرید (MBA[3]) و درک الگوی خرید مشتریان استفاده می‌کنند. این اطلاعات می‌توانند برای اهداف مختلفی مانند فروش محصولات جدید به مشتریان ثابت، افزایش فروش محصولات، تبلیغات فروش، طرح‌های پاداش وفاداری، طراحی فروشگاه، برنامه‌های تخفیف و بسیاری موارد دیگر استفاده شوند.

قوانین انجمنی، قوانینی به شکل $A \rightarrow B$ هستند که بیان می‌کنند اگر تمام اعضای مجموعه‌ی A در یک سبد خرید باشند، آنگاه به احتمال زیاد B نیز در آن سبد خرید وجود خواهد داشت. پس از یافتن مجموعه‌های آیتم پرتکرار یک دیتاست، می‌توانیم از آن‌ها قوانین انجمنی سودمند را استخراج کنیم، اما هر قانونی سودمند نیست. به همین دلیل برای سنجش اهمیت و سودمندی یک قانون انجمنی، معیارهای زیر تعریف شده‌اند:

۱. **Confidence (با نماد conf(A → B))**: confidence به عنوان احتمال وقوع B در سبد خرید تعریف می‌شود اگر سبد خرید قبلاً شامل A باشد:

$$conf(A \rightarrow B) = Pr(B|A)$$

که در آن Pr(B|A) احتمال شرطی یافتن مجموعه آیتم B با فرض وجود مجموعه آیتم A است.

۲. **Lift (با نماد lift(A → B))**: lift بیان می‌کند که احتمال رخداد توامان A و B نسبت به حالتی که A و B مستقل باشند، چقدر بیشتر است.

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

---

۱ Frequent Itemsets
۲ Association Rules
3 Market Basket Analysis

که در آن $S(B) = \frac{Support(B)}{N}$ و N = تعداد کل تراکنش‌ها (سبدها) می‌باشد.

۳. **Conviction (با نماد conv(A → B))**: conviction نسبت احتمال رخداد A و B اگر نسبت به هم مستقل باشند را با فرکانس واقعی رخداد A بدون B مقایسه می‌کند.

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

(الف) یکی از اشکالات استفاده از confidence نادیده گرفتن Pr(B) می‌باشد. چرا این یک اشکال است؟ توضیح دهید چرا lift و conviction از این اشکال رنج نمی‌برند.

(ب) یک معیار متقارن است اگر measure(A → B) = measure(B → A). کدام یک از معیارهای ارائه شده در اینجا متقارن هستند؟ برای هر معیار، اگر متقارن است، اثبات کنید و یا با یک مثال نقض که نشان دهید که معیار متقارن نیست.

(پ) قوانین *دلالت کامل*[۱] قوانینی هستند که در آن‌ها وجود یک مقدم[۲]، به طور ۱۰۰٪ وجود یک تالی[۳] را پیش‌بینی می‌کند. برای مثال، در قانون $A \rightarrow B$ اگر A رخ دهد، B نیز حتما رخ خواهد داد. یک معیار *((مطلوب))* است اگر برای همه دلالت‌های کامل به حداکثر مقدار قابل دستیابی خود برسد. این کار تشخیص بهترین قوانین را آسان می‌کند. کدام یک از معیارهای فوق این ویژگی را دارند؟ شما می‌توانید موارد 0/0 را نادیده بگیرید اما سایر حالات بی‌نهایت را باید در نظر بگیرید.

---

۱ Perfect Implications

۲ Antecedent

۳ Consequent

کاوش داده‌گان انبوه

تمرین «دو»

دکتر سامان هراتی‌زاده
دانشگاه تهران - دانشکده سامانه‌های هوشمند
نیم‌سال دوم ۱۴۰۳-۱۴۰۲

دستیاران آموزشی
محمد راشدی
آرمین رحیمی دره‌باغ

ددلاین: ساعت ۵۹:۲۳ | ۱۴۰۳/۰۴/۲۰

کاربرد در توصیه محصول

عمل فروش محصولات یا خدمات اضافی به مشتریان قبلی به عنوان فروش مکمل[۱] شناخته می‌شود. پیشنهاد محصول یکی از مثال‌های فروش مکمل است که اغلب توسط خرده‌فروشان آنلاین استفاده می‌شود. یک روش ساده برای ارائه پیشنهاد محصول این است که پیشنهادات بر اساس محصولاتی که مشتریان قبلاً به صورت آنلاین مرور کرده‌اند، ارائه شوند.

فرض کنید می‌خواهیم محصولات جدیدی را به مشتریان بر اساس محصولاتی که قبلاً به صورت آنلاین مرور کرده‌اند، پیشنهاد کنیم. یک برنامه با استفاده از الگوریتم *A-priori* بنویسید تا محصولاتی را که اغلب با هم مرور می‌شوند پیدا کنید. Support را برابر ۱۰۰ قرار دهید (یعنی جفت محصولات باید حداقل ۱۰۰ بار با هم رخ دهند تا به عنوان پرتکرار در نظر گرفته شوند) و مجموعه آیتم‌های پرتکرار با اندازه ۲ و ۳ را پیدا کنید.

دیتاست ما فایل browsing.txt است که در پوشه‌ی q3 قرار دارد. هر خط از این فایل آیتم‌هایی که یک کاربر در یک session مرور کرده است را نشان می‌دهد. در هر خط، هر آیتم با یک رشته‌ی ۸ کاراکتری مشخص شده‌است و آیتم‌ها با یک فاصله از هم جدا شده‌اند.

توجه: برای بخش‌های (ت) و (ث)، قوانین باید با ترتیبی خاص در گزارش ذکر شوند اما کد شما نیازی به مرتب‌سازی خروجی ندارد. همچنین می‌توانید از دو تست صحت زیر برای اطمینان از پاسخ‌تان استفاده کنید:

۱. بعد از اولین pass ۶۴۷ آیتم پرتکرار وجود دارد (|L1| = 647)
۲. پنج جفت برتر که باید در بخش (d) تولید کنید همگی confidence بیش از 0.985 دارند.

دستورالعمل‌های دقیق‌تر در ادامه آمده است. لطفاً حداقل پنج رقم اعشار برای confidence درنظر بگیرید.

---

[۱] Cross-Selling

(ت) جفت آیتم‌های (X, Y) را شناسایی کنید به طوری که support $\{X, Y\}$ حداقل ۱۰۰ باشد. برای تمام این جفت‌ها، confidence قوانین انجمنی مربوطه را محاسبه کنید: $X \Rightarrow Y$، $Y \Rightarrow X$. قوانین را به ترتیب نزولی confidence مرتب کنید و ۵ قانون برتر را در نوشتار ذکر کنید. در صورت تساوی مقادیر confience، آن‌ها را به صورت الفبایی براساس سمت چپ قوانین مرتب کنید. (نیازی به استفاده از Spark برای بخش‌های (d) و (e) سوال نیست)

(ث) مجموعه آیتم‌های سه‌تایی (X, Y, Z) را شناسایی کنید به طوری که support $\{X, Y, Z\}$ حداقل ۱۰۰ باشد. برای همه این سه‌تایی‌ها، confidence قوانین انجمنی مربوطه را محاسبه کنید: $(X, Y) \Rightarrow Z$، $(X, Z) \Rightarrow Y$ و $(Y, Z) \Rightarrow X$. قوانین را به ترتیب نزولی confidence مرتب کنید و ۵ قانون برتر را در نوشتار ذکر کنید. در صورت تساوی مقادیر confidence، آن‌ها را به صورت الفبایی براساس سمت چپ قوانین مرتب کنید.

۴ – پیاده‌سازی و مقایسه الگوریتم‌های یافتن مجموعه آیتم‌های پرتکرار

هدف ما در این قسمت پیاده‌سازی تعدادی از الگوریتم‌های شناخته‌شده برای یافتن مجموعه آیتم‌های پرتکرار و قونین انجمنی می‌باشد. در سوال قبل شما الگوریتم A-priori را پیاده‌سازی کردید، حال در این قسمت می‌خواهیم چند الگوریتم که A-priori را بهبود می‌دهند، پیاده‌سازی کنیم. برای هر الگوریتم، بهبود و مزایا و معایب آن نسبت به الگوریتم A-priori را بیان کنید.

از همان دیتاست بخش قبل استفاده کنید. در ادامه ابتدا الگوریتم‌های ذکر شده را پیاده‌سازی کرده و با استفاده از آن‌ها مجموعه آیتم‌های پرتکرار را به دست آورید. سپس قوانین انجمنی که confidence آن‌ها بیش از ۵۰٪ را نیز به دست آورده و چاپ کنید. آیا تمام الگوریتم‌ها پاسخی مشابه به ما می‌دهند؟ در پایان الگوریتم‌ها را از نظر زمان و مصرف حافظه با هم مقایسه کنید.

(الف) PCY

(ب) Toivonen

(پ) Eclat

(ت) FP Growth

(ث) OPUS : مقاله این الگوریتم در پوشه‌ی q4 قرار داده شده است.