تمرین ۲-۱-الگوریتم‌های تخمین تعداد متمایز و بسامد

۱- الف-چرا از معادله زیر جهت تعیین رتبهٔ ورودی استفاده می‌شود؟ به بیان دیگر، چرا نمایه برابر مقدار یک با کمترین ارزش به عنوان معیار رتبه انتخاب می‌شود؟ آیا در مقالات اصلی همین مورد پیش رفته است؟

$$\text{رتبه}(i) = \begin{cases} \min_{i_k \neq 0} k, \forall i > 0 \\ \ell, \qquad i = 0 \end{cases}$$

ب- دلیل استفاده از مقادیر درهم به جای ورودی در روش‌های تخمین تعداد منحصر به فرد چیست؟ می‌توان از نمایش عددی دیگری با همین کارائی استفاده کرد؟

ج- روش مقدار ضرائب در روش تخمین شمارش احتمالاتی و لگ‌لگ و ابرلگ‌لگ بررسی و گزارش شود.

۲- الگوریتم‌های تخمین تعداد اعضای متمایز را روی متن شاهنامه اعمال کنید. از روش‌های تخمین دقیق و شمارش خطی، شمارش احتمالی، الگوریتم لگ‌لگ، الگوریتم ابرلگ‌لگ، و ابرلگ‌لگ++ استفاده کنید. ممکن است قدم‌هائی برای آماده‌سازی اعم از استانداردسازی نیاز داشته باشید. موارد را لحاظ کنید.

۳- ادغام دو مجموعه در روش شمارش احتمالاتی و لگ‌لگ به چه نحوی انجام می‌پذیرد.

۴- مجموعه‌ای با 2^{16} رشته ۳۲ بیتی را تولید کنید به طوری که هر بیت در هر رشته از توزیعی یکنواخت تصادفی تبعیت کند. مقادیر حاصل را چون مقدار در هم لحاظ کنید. الف- به نظر شما با توجه به شرایط اشاره شده جهت تولید داده تصادمی وجود دارد؟ ب- نمودار هیستوگرامی از اعمال چند روش شمارش احتمالی (تک سطلی)، PCSA، لگ‌لگ، و ابرلگ‌لگ بر ۱۰۰ بار تولید چنین رشته‌هایی رسم کنید، به طوری که محور x میزان تخمین بدست آمده و محور y تعداد تکرار را به دست دهد (یاداوری هیستوگرام). در صورتی که نیاز بود محور x را به صورت لگاریتمی نمایش دهید. میانگین‌گیری حسابی و هارمونیک چقدر تاثیر داشته‌اند؟ میزان خطاها،تاثیر نرمال‌سازها و میزان خطا را بر اساس معادلات تخمین خطا بررسی کنید. ج- در نموداری دیگر تاثیر تغییر m را با مقادیر ۱۶ و ۶۴ و ۲۵۶ بر ابرلگ‌لگ رسم و تحلیل کنید. در کلاس عنوان شد نسبت $n/m>20$ مناسب است. چنین مواردی را تحلیل کنید. د- الگوریتم ابرلگ‌لگ را با الگوریتم CVM مقایسه کنید.

۵- روش طرح کمینه را با میانگین بررسی کنید و نتایج آن را با میانه بسنجید.

۶- در تحلیل مجموعه‌داده‌های بزرگ، نظیر پایگاه داده تراکنش‌های فروش اینترنتی، یکی از چالش‌های اساسی، یافتن محصولاتی یا کالاهایی با فراوانی بیشتر از آستانه‌ای مشخص (پربسامدها) بدون مصرف حافظه زیاد و با تعداد پویش کم روی داده است. روش کلاسیک A-Priori با وجود کارایی تحلیلی بالا، در مواجهه با تعداد بسیار زیاد اقلام (مثلاً صدها هزار محصول منحصربه‌فرد) و محدودیت حافظه، با مشکل مواجه می‌شود. از سوی دیگر، الگوریتم‌های جریان داده مانند میسرا-گریس، طرحوارهٔ شمارش، و طرحوارهٔ شمارش-کمینه با حافظه ثابت و تنها یک پویش از داده، تخمینی از فراوانی اقلام ارائه می‌دهند. بررسی کنید چگونه می‌توان از این الگوریتم‌ها به عنوان مرحله پیش‌پالایش در روش A-Priori استفاده کرد و مزایا و معایب هر ترکیب را ارزیابی کنید.
جهت پیشبرد کار مجموعه داده‌ای مصنوعی و در عین حال واقع‌گرایانه از تراکنش‌های خرید تولید کنید، سپس سه روش ترکیبی متفاوت را بر پایه الگوریتم‌های Misra-Gries میسرا-گریس، طرحوارهٔ شمارش، و طرحوارهٔ شمارش-کمینه پیاده‌سازی کرده و با

روش اصلی A-Priori مقایسه نمایید. هدف نهایی، تعیین شرایطی (محدودیت حافظه، تعداد پویش‌ها، دقت مورد نیاز) است که هر روش ترکیبی بر دیگری برتری دارد.

نحوه تهیه و ایجاد مجموعه داده

برای اینکه تمرین قابل اجرا و تکرارپذیر باشد، شما خود باید یک مجموعه داده تصنعی با ویژگی‌های مشخص تولید کنید. ویژگی‌های پیشنهادی به شرح زیر است:

تعداد کل تراکنش‌ها، یک میلیون عدد در نظر گرفته شود. هر تراکنش شامل تعدادی محصول است که شناسه آن‌ها عددی صحیح و مثبت فرض می‌شود. تعداد کل محصولات منحصربه‌فرد، پنجاه هزار عدد باشد. میانگین طول هر تراکنش (تعداد محصولات درون یک سبد) ده مورد در نظر گرفته شود. همچنین توزیع فراوانی محصولات باید به گونه‌ای باشد که تعداد کمی محصول (مثلاً یک درصد محصولات) بسیار پرتکرار (بیش از یک درصد تراکنش‌ها) و بقیه محصولات کم‌تکرار باشند. روش‌های ریاضیاتی نمره ویژه خواهند داشت. آستانه پشتیبانی مورد نظر برای پربسامدی را ۰.۵ درصد از کل تراکنش‌ها تعیین کنید. بنابراین با یک میلیون تراکنش، هر محصولی که در حداقل پنج هزار تراکنش ظاهر شود، محصولی پربسامد به حساب می‌آید.

مراحل

گام نخست، تولید مجموعه داده: مطابق با ویژگی‌های بالا،دانشجو کد تولید داده را نوشته و فایل خروجی را با نام trakonesh.txt ذخیره کنید. همچنین باید توزیع واقعی فراوانی محصولات را برای ارزیابی نهایی محاسبه و ذخیره نماید تا بتواند صحت نتایج روش‌های تقریبی را بسنجد.

گام دوم، پیاده‌سازی روش اصلی A-Priori به عنوان مبنا است. در این روش، بدون استفاده از هیچ تقریبی، ابتدا در یک پویش کامل، فراوانی تک‌محصولات محاسبه شده و محصولات با فراوانی کمتر از پنج هزار حذف می‌شوند. فرض می‌شود تعداد محصولات باقیمانده چندصد عدد باشد. در صورت نیاز به یافتن جفت‌های پربسامد، پویشی دوم برای شمارش دقیق جفت‌های حاصل از محصولات پربسامد انجام می‌گیرد. میزان حافظه مصرفی (تعداد دفعاتی که همزمان شمارنده‌ها در حافظه نگهداری شده‌اند) و تعداد کل پویش‌های انجام شده بر روی فایل داده را گزارش کنید.

گام سوم، پیاده‌سازی سه روش ترکیبی: با استفاده از الگوریتم‌های پربسامد به عنوان پالایش است. برای هر کدام از الگوریتم‌های میسرا-گریس، طرحوارهٔ شمارش، و طرحوارهٔ شمارش-کمینه ، ساختار با پارامترهایی طراحی کنید که تضمین کند هیچ محصول پربسامدی حذف نشود. فرضیات لازم را دربارهٔ اندازهٔ آرایه‌های روش‌ها اعمال و تحلیل کنید. در هر سه روش، ابتدا پویش کامل روی داده انجام داده و خلاصه را می‌سازید. سپس از روی خلاصه، فهرست محصولات پربسامد استخراج شود. در مرحله بعد، پویش دوم برای محاسبه فراوانی دقیق همین کاندیدها انجام دهید. در نهایت، در صورت نیاز به جفت‌های پربسامد، از میان محصولاتی که پس از پویش دوم تأیید شدند، جفت‌ها ساخته شده و با پویش سوم شمارش دقیق می‌شوند.

گام چهارم، ارزیابی و مقایسه: جدولی شامل موارد زیر برای هر چهار روش (یک روش اصلی و سه روش ترکیبی) تهیه کنید: تعداد پویش‌های کامل انجام شده، حداکثر حافظه مصرفی همزمان بر حسب تعداد زوج‌های (کلید، مقدار) ذخیره شده، تعداد محصولات نامزد در مرحله اول، دقت در تشخیص محصولات واقعاً پربسامد (نسبت مثبت صادق به کل مثبت‌ها)، و در صورت پیاده‌سازی جفت‌ها، دقت مربوط به جفت‌ها. همچنین باید زمان اجرای هر روش (به صورت نسبی) را گزارش دهید.

ب- همچنین، موارد زیر را با استفاده از نتایج حاصل گزارش کنید.

- استفاده از هر یک از سه روش ترکیبی، تعداد پویش‌ها را نسبت به روش اصلی A-Priori افزایش داده یا کاهش؟ توضیح دهید که چرا برای یافتن تنها محصولات پرتکرار، تعداد پویش‌ها در روش ترکیبی بیشتر از روش اصلی نیست، اما برای یافتن جفت‌ها ممکن است پویش اضافی لازم شود.

- نشان دهید که چگونه روش‌های ترکیبی به ویژه در مرحله اول، حافظه بسیار کمتری نسبت به روش اصلی مصرف می‌کنند. در مقابل، آیا در مراحل بعدی (ذخیره‌سازی جفت‌های نامزد) این مزیت از بین می‌رود؟

- بررسی کنید که هیچ یک از سه روش ترکیبی، محصولی پربسامد را از دست داده است. در صورت مثبت بودن پاسخ، کدام روش و با چه پارامترهایی این چنین عمل کرده است. همچنین میزان بیش- یا کم-تخمیتی هر روش را نسبت به فراوانی محصولات محاسبه کنید.

- با توجه به نتایج خود، توضیح دهید که آیا همیشه روش ترکیبی بهتر از روش اصلی A-Priori است یا خیر. شرایطی را نام ببرید که در آنها روش اصلی علی‌رغم مصرف حافظه بیشتر، به دلیل تعداد پویش کمتر یا سادگی پیاده‌سازی، گزینه بهتری محسوب می‌شود.

تمرین‌ها در گروه‌های یک یا دو نفره ممکن است. مراجع در صورت استفاده باید دقیق باشد و صرفا به مدل‌های زبانی تکیه نشود.