به نام خدا

دانشگاه تهران

دانشکده مهندسی برق و کامپیوتر

درس داده کاوی

تمرین دوم

طراحان	دیبا رشیدی Diba.rashidi@ut.ac.ir
تاریخ بارگذاری	۱۴۰۲/۱۲/۲۵
مهلت ارسال	۱۴۰۲/۰۱/۱۷

فهرست

بخش تشریحی

سوال اول

یک data warehouse برای ذخیره اطلاعات پروازهای خارجی از سه بعد: مبدا، مقصد و زمان پرواز و دو معیار قیمت بلیط و تعداد مسافران ایجاد شده است.

الف) شمای ستاره‌ای متناظر با این انبار داده را رسم کنید.

ب) با شروع از cuboid پایه [مبدا، مقصد ،زمان] برای مقایسه میانگین قیمت بلیط‌های تهران به میلان در خرداد ماه سال ۱۴۰۲ با میانگین قیمت بلیط‌های تهران به آمستردام در فروردین ۱۴۰۱ چه عملیات‌های OLAP نیاز است انجام شود.

سوال دوم

یک cuboid پایه با سه بعد A, B,C را در نظر بگیرید.تعداد سلول‌های هر بعد برابر است با:

$|A| = 1{,}000{,}000 \;;\; |B| = 100 \;;\; |C| = 1{,}000$

و هر بعد به صورت مساوی به ۱۰ قسمت تقسیم شده است.

الف) ترتیب بهینه‌ی پیمایش chunk ها در cuboid پایه به منظور ساختن cuboid های دوبعدی را بیان کنید.

ب) اگر هر سلول cube، یک معیار را در داخل ۴ بایت ذخیره کند؛ در بهترین حالت چه مقدار فضا در حافظه‌ی اصلی برای ساختن cuboid های دو بعدی مورد نیاز است؟

ج) بهترین روش برای محاسبه هر کدام از cuboid های یک بعدی، با استفاده از cuboid های دو بعدی چیست؟

سؤال سوم

فرض کنید که cuboid پایه‌ی یک a دارای دو سلول زیر است و می‌دانیم:

$$a_i \neq b_i$$

$$\left(a_1, \underline{a_2}, b_3, a_4, \underline{a_5}, b_6, a_7, \underline{a_8}, b_9\right) : 15$$

$$\left(b_1, \underline{a_2}, a_3, b_4, \underline{a_5}, a_6, b_7, \underline{a_8}, a_9\right) : 10$$

الف) چند cuboid در این data cube وجود دارد؟

ب) این data cube، چند سلول aggregate غیرتهی دارد؟

ج) چند سلول بسته‌ی غیر تهی در این data cube موجود است؟ چه تعداد از آنها aggregated است؟

د) اگر شرط minimum support = 20 را در نظر بگیریم، تعداد سلول‌های aggregate غیر تهی در iceberg cube متناظر چقدر خواهد بود؟

سوال چهارم

مجموعه داده زیر با ۳ بعد، اطلاعات مشتریان یک فروشگاه را نمایش می‌دهد. هدف ما، محاسبه یک iceberg cube با استفاده از الگوریتم BUC است.

Gender	education	occupation
female	college	teacher
female	college	programmer
female	college	programmer
female	graduate	teacher
female	high school	CEO
male	high school	programmer
male	high school	CEO
male	college	teacher
male	college	programmer
male	graduate	doctor

الف) ترتیب پردازش ابعاد را به گونه‌ای تعیین کنید که الگوریتم BUC بهترین کارایی را داشته باشد. دلایل خود را برای انتخاب این ترتیب در گزارش بنویسید.

ب) با توجه به ترتیبی که برای بررسی ابعاد در بخش الف مشخص کردید، الگوریتم BUC را روی مجموعه‌داده‌ی فوق اجرا کنید و iceberg cube را با شرط minimum support = 3 محاسبه نمایید

بخش عملی

پیش‌نیازها

برای پاسخ به این تمرین عملی باید از زبان برنامه‌نویسی Python استفاده‌کنید و نیاز است که پیش از شروع، یک سرور Jupyter بر روی سیستم نصب و راه‌اندازی شود تا بتوانید بر روی یک فایل ipynb. کدهای خود را اجرا کنید، همچنین راه حل جایگزین آن استفاده از Google Colab است.

استفاده از کتابخانه‌های Pandas، Numpy و Datetime می‌تواند گزینه‌ی مناسبی برای حل مسائل پیشرو باشد.

شرح دادگان

این مجموعه داده با نام weatherAUS در فایل فشرده dataset.zip قرار داده شده و شامل اطلاعات مربوط به آب و هوای استرالیا است. این مجموعه داده شامل ۲۳ ستون است و اطلاعات مربوط به هر ستون در جدول زیر آورده شده‌است.

جدول ۱. توضیحات مجموعه دادگان

#	Column name	Description
1	Date	The date of observation
2	Location	The common name of the location of the weather station
3	MinTemp	The minimum temperature in degrees Celsius or Fahrenheit
4	MaxTemp	The maximum temperature in degrees Celsius or Fahrenheit
5	Rainfall	The amount of rainfall recorded for the day in mm
6	Evaporation	The so-called Class A pan evaporation (mm) in the 24 hours to 9am
7	Sunshine	The number of hours of bright sunshine in the day.
8	WindGustDir	The direction of the strongest wind gust in the 24 hours to midnight
9	WindGustSpeed	The speed (km/h) of the strongest wind gust in the 24 hours to midnight
10	WindDir9am	Direction of the wind at 9am
11	WindDir3pm	Wind speed (km/hr) averaged over 10 minutes prior to 9am
12	WindSpeed9am	Wind speed (km/hr) averaged over 10 minutes prior to 9am
13	WindSpeed3pm	Wind speed (km/hr) averaged over 10 minutes prior to 3pm
14	Humidity9am	Humidity (percent) at 9am
15	Humidity3pm	Humidity (percent) at 3pm
16	Pressure9am	Atmospheric pressure (hpa) reduced to mean sea level at 9am
17	Pressure3pm	Atmospheric pressure (hpa) reduced to mean sea level at 3pm
18	Cloud9am	Fraction of sky obscured by cloud at 9am. This is measured in "oktas", which are a unit of eigths. It records how many eigths of the sky are obscured by

		cloud. A 0 measure indicates completely clear sky whilst an 8 indicates that it is completely overcast.
19	Cloud3pm	Fraction of sky obscured by cloud (in "oktas": eighths) at 3pm. See Cload9am for a description of the values
20	Temp9am	Temperature (degrees C) at 9am
21	Temp3pm	Temperature (degrees C) at 3pm
22	RainToday	Boolean: 1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0
23	RainTomorrow	The amount of next day rain in mm. Used to create response variable RainTomorrow. A kind of measure of the "risk".

سوالات

۱. در ابتدا پیش پردازش‌های لازم بر روی داده‌ها را انجام دهید و مراحل آن را به طور کامل توضیح دهید. (برای این کار می توانید از مراحل کار تمرین گذشته استفاده کنید)

۲. مقادیر زیر را از مجموعه داده به دست آورید:

الف) میانگین بارش در استرالیا

ب) تعداد روزهای بارانی در شهر Watsonia در سال ۲۰۱۵

ج) بیشترین رطوبت ثبت شده در شهر Townsville

د) اختلاف میانه‌های MinTemp و MaxTemp در ماه ژانویه (ماه اول میلادی)

ه) سردترین ۵ روز شهر MountGinini در سه ماه اول سال

۳. روشی کارآمد برای محاسبه هر کدام از سنجه‌های سوال بالا ارائه دهید. به طوری که قابلیت افزایشی داشته باشد.

ملاحظات

تمامی نتایج شما باید در یک فایل فشرده با عنوان DM_CA2_StudentID تحویل داده شود.

خوانایی و دقت بررسی‌ها در گزارش نهایی از اهمیت ویژه‌ای برخوردار است. به تمرین‌هایی که به صورت کاغذی تحویل داده شوند یا به صورت عکس در سایت بارگذاری شوند، ترتیب اثری داده نخواهد شد.

بخش اصلی نمره به گزارش شما تعلق می‌گیرد و دستیاران الزامی برای اجرای تمام کدهای شما در صورتی که در گزارش به آن‌ها اشاره‌ای نکرده باشید ندارند. لطفا تمام موارد مورد نیاز را در گزارش ذکر کنید.

کدهای نوشته شده برای هر بخش را با نام مناسب مشخص کرده و به همراه گزارش تکلیف ارسال کنید. همه‌ی کدهای پیوست گزارش بایستی قابلیت اجرای مجدد داشته باشند. در صورتی که برای اجرا مجدد آنها نیاز به تنظیمات خاصی می‌باشد بایستی تنظیمات مورد نیاز را نیز در گزارش خود ذکر کنید.

برای تحویل تمارین از چارچوب قرارداده شده در سامانه و کانال تلگرام استفاده کنید.

در صورت قصد ارسال تمرین به صورت دیگر (انگلیسی، latex و ...)، لطفا پیش از ارسال با دستیارمسئول تمرین هماهنگ کنید.

توجه کنید این تمرین باید به صورت تک نفره انجام شود و پاسخ‌های ارئه شده باید نتیجه فعالیت فرد نویسنده باشد (هم‌فکری خارج از چارچوب و به اتفاق هم نوشتن تمرین نیز ممنوع است). در صورت مشاهده تقلب برای همه‌ی افراد مشارکت کننده، نمره تمرین، صفر در نظر گرفته خواهد شد.

در صورت بروز هرگونه مشکل با ایمیل زیر در ارتباط باشید:

mailto:diba.rashidi@ut.ac.ir

مهلت تحویل: ۱۷ فروردین ۱۴۰۳

مهلت تحویل با تاخیر: ۲۴ فروردین ۱۴۰۳