به نام او

متن پیش رو شامل دو دسته سوال است. در دسته اول باید به تمامی سوالات پاسخ دهید و در دسته دوم یکی از سه سوال را انتخاب کرده و به آن پاسخ دهید.

دقت داشته باشید که همه پاسخ‌ها باید با مستندات لازم همراه باشد و کد پایتون را ضمیمه کنید.

به تمامی سوالات پاسخ دهید.

سوال ۱: خوشه بندی

خوشه‌بندی یکی از حوزه‌های مهم در علم داده است و هدف آن این است که داده‌ها را در تعداد معینی خوشه گروه‌بندی کند به نحوی که داده‌های درون یک خوشه بیشترین شباهت و داده‌های دو خوشه متفاوت بیشترین عدم شباهت را به هم داشته باشند. حال فرض کنید n داده در اختیار داریم و می‌خواهیم آنها را به k خوشه دسته‌بندی کنیم. عدد d را به‌گونه‌ای بیابید که فاصله زوجی هر دو داده از هم از این مقدار بزرگ‌تر باشد. به بیان دیگر، به دنبال یافتن حد پایین فاصله داده‌ها از هم هستیم. مشخصات حل مسئله به صورت زیر است:

ورودی: خط اول شامل تعداد داده‌هاست (n). در هر n خط بعدی مشخصات داده mام در قالب (x_i, y_i) داده می‌شود. در خط آخر تعداد خوشه (k) ارائه خواهد شد.

داده‌ها به صورت عدد بوده و $2 \leq k \leq n \leq 200; -103 \leq x_i, y_i \leq 103$.

خروجی: بیشترین مقدار d به عنوان خروجی گزارش شود. این عدد باید با ۷ رقم اعشار نمایش داده شود.

محدودیت زمانی: پاسخ کمتر از ۳ ثانیه ارائه شود.

مثال ۱

ورودی

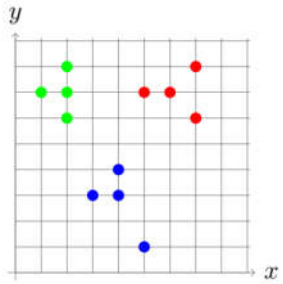
```
12
7 6
4 3
5 1
1 7
2 7
5 7
3 3
7 8
2 8
4 4
6 7
2 6
3
```

خروجی

```
2.828427124746
```

توضیحات

پاسخ $\sqrt{8}$ است. نمایش داده‌ها در سه دسته به شکل روبرو است.



مثال ۲

ورودی

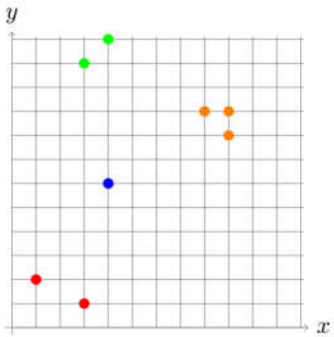
```
8
3 1
1 2
4 6
9 8
9 9
8 9
3 11
4 12
4
```

خروجی

```
5.000000000
```

توضیحات

پاسخ 5 است. نمایش داده‌ها در چهار دسته به شکل روبرو است.



مسئله ۲: مسئله فروشنده دوره‌گرد

فرض کنید می‌خواهید برای یک راننده کامیونی کوتاه‌ترین مسیر را به صورت زیر بیابید: مسیر از یک انبار آغاز می‌شود و کامیون باید به فروشگاه‌های مختلف مراجعه کرده و کالاها را تحویل دهد. در انتها کامیون به انبار باز می‌گردد. بنابراین شما باید کوتاهترین مسیری را بیابید که از انبار آغاز شده، حداقل یکبار به همه فروشگاه‌ها مراجعه کرده و مجدد به انبار باز می‌گردد.

ورودی: ورودی مسئله در دو بخش به شما ارائه می‌شود. بخش اول شامل توصیفی از مسیرهاست و بخش دوم شامل پرس‌وجوهاست که به صورت مجزا به شما داده خواهند شد. ابتدا بخش نخست (گراف شبکه) به عنوان ورودی ارائه می‌شود و شما فرصت دارید که این گراف را پردازش کنید. پس از پایان پردازش باید عبارت Ready پرینت شود (فایل پایتون ارائه شده به شما کمک خواهد کرد). پس از آنکه عبارت ready چاپ شد، بخش دوم ورودی (پرس‌وجو) به عنوان ورودی به شما داده می‌شود.

ورودی بخش نخست: خط اول این ورودی شامل دو عدد n (تعداد گره‌ها در شبکه) و m (تعداد لینک‌ها یا مسیرها در شبکه) است. هر یک از m خط بعدی در ورودی شامل سه عدد u، v و l است که وجود مسیری با طول l بین دو گره (u, v) را نشان می‌دهد.

ورودی بخش دوم: خط اول ورودی بخش دوم شامل عدد q یا تعداد پرس‌وجوها برای محاسبه فاصله است. هر یک از q خط بعدی با عدد k آغاز می‌شود که نشان دهنده تعدادگره انبار و فروشگاه‌هایی است که کامیون باید به آنها مراجعه کند. k عدد دیگر نیز در همان خط ارائه می‌شود که اولین آنها شماره گره مربوط به انبار بوده و بقیه مربوط به شماره گره‌های فروشگاه‌هاست. به عنوان نمونه در ورودی مانند 2 1 2 اولین عدد ۲ بیانگر دو گره (یک انبار و یک فروشگاه) است که شماره گره انبار، ۱ و شماره گره فروشگاه، ۲ تعیین شده است.

در این مسئله: $1 \le n \le 110000;\ 1 \le m \le 250000;\ 1 \le u, v \le n;\ 1 \le l \le 100000;$ و

$$1 \le q \le 100;\ 1 \le k \le 20$$

خروجی: پس از پردازش وردی بخش نخست باید عبارت ready چاپ شود و سپس ورودی بخش دوم پردازش شود و برای هر ورودی این بخش یک خروجی چاپ شود. اگر هیچ مسیری بین انبار و فروشگاه‌ها وجود نداشت خروجی 1- چاپ شود و درغیر اینصورت، خروجی شامل طول کوتاه‌ترین مسیری است که از انبار شروع شده و به هر فروشگاه حداقل یکبار مراجعه کرده و مجددا به انبار ختم می‌شود.

محدودیت زمانی: برای پردازش وردی اول زمان بایدکمتر از ۵۰ ثانیه باشد و برای پردازش ورودی دوم زمان کمتر از ۴۰ ثانیه باشد.

مثال

ورودی

```
4 5
1 2 1
2 3 1
3 4 1
4 1 1
2 1 1
3
2 1 2
2 1 3
4 1 2 3 4
```

خروجی

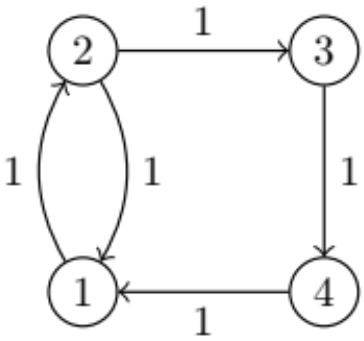
```
Ready
2
4
4
```

توضیحات

ورودی اول:

خط اول به این معناست که شبکه ۴ گره و ۵ مسیر دارد.

خط دوم شامل مسیری است که از گره ۱ آغاز می‌شود به گره ۲ می‌رود با طول ۱. خط سوم شامل مسیری است که از گره ۲ آغاز می‌شود به گره ۳ می‌رود با طول ۱. و به همین ترتیب کل شبکه ترسیم می‌شود (شکل روبرو).



ورودی دوم:

خط اول شامل تعداد پرس‌وجوهاست.

خط دوم شامل پرس‌وجویی است که از ۲ گره تشکیل شده است. گره ۱ برای انبار و گره ۲ برای فروشگاه. مسیر از گره ۱ آغاز می‌شود به گره ۲ رفته و مجدد به گره ۱ بازمی‌گردد، پس طول مسیر برابر ۲ است.

خط سوم شامل پرس‌وجویی است که از ۴ گره تشکیل شده است. گره ۱ برای انبار و گره ۳ برای فروشگاه. مسیر از گره ۱ آغاز می‌شود به گره ۲ و ۳ رفته و برای بازگشت به گره مجدد به گره ۱ باید به گره ۴ برود و سپس به گره ۱. پس طول مسیر برابر ۴ است.

با کمک فایل پایتون ضمیمه ورودی را خوانده و روش حل مسئله خود را در قسمت خالی کدنویسی کنید.

مسئله ۳: بزرگترین (طولانی‌ترین) زیررشته مشترک

فرض کنید دو رشته s و t در اختیار دارید و هدف پیدا کردن رشته w است به طوری که طول این رشته برابر با طول بیشترین زیررشته مشترک s و t باشد. درواقع به دنبال پیداکردن طول بیشترین رشته مشترک بین دو رشته در زمان خطی هستیم.

ورودی: هر خط ورودی شامل دو رشته ورودی s و t است که از حروف کوچک لاتین با طول کمتر از ۱۰۰ هزار کاراکتر تشکیل شده است.

خروجی: برای هر دو رشته، طولانی‌ترین (بزرگ‌ترین) زیررشته مشترک را بیابید. خروجی شامل ۳ عدد باید باشد. عدد اول، نقطه آغاز زیررشته در s، دومین عدد، نقطه آغاز زیررشته در t و سومین عدد نشانگر طول زیررشته است. اگر چندین پاسخ وجود داشت، یکی را گزارش کنید.

محدودیت زمانی: پاسخ کمتر از ۱۵ ثانیه پاسخ چاپ شود.

مثال

ورودی	خروجی
cool toolbox aaa bb aabaa babbaab	1 1 3 0 1 0 0 4 3

توضیحات

طولانی‌ترین زیر رشته مشترک در خط اول "ool" بوده که از نقطه ۱ در رشته اول و دوم آغاز شده و طول آن ۳ است.

در خط دوم، هیچ رشته مشترکی وجود ندارد و طول برابر ۰ است. در این حالت هر عددی را می‌توانید برای نقطه آغاز رشته‌ها ارائه دهید.

در خط سوم، زیررشته "aab" زیررشته مشترک بوده که از نقطه ۰ در رشته اول و ۴ در رشته دوم آغاز شده و طول آن برابر ۳ است. زیر رشته "baa" نیز زیررشته مشترک دیگری با طول ۳ است که از نقطه ۲ در رشته اول و ۳ در رشته دوم آغاز شده و طول آن برابر ۳ است (هرکدام از این دو می‌تواند به عنوان خروجی چاپ شود).

به یکی از ۳ سوال زیر پاسخ دهید.

مسئله ٤: تحلیل احساس کاربر

تحلیل احساسات یکی از زیرمجموعه‌های روش‌های دسته‌بندی است که می‌توان آن را با کمک روش‌های مختلفی انجام داد. به عنوان نمونه، فرض کنید سه دسته احساس مثبت (خوشایند)، خنثی و منفی (ناخوشایند) به صورت زیر تعریف شده باشد:

- نظراتی مثبت هستند که کاربر به آنها ۴ و یا ۵ ستاره داده است.
- نظراتی منفی هستند که کاربر به آنها ۲ یا ۱ ستاره داده است.
- نظرات خنثی نظراتی هستند که کاربر به آنها ۳ ستاره داده است.

حال با کمک برچسب‌های فوق و الگوریتم‌هایی مانند لجستیک و یا kNN می‌توان نظرات را دسته‌بندی کرد.

با این منطق، می‌خواهیم با کمک مجموعه داده amazon_baby سیستمی برای تحلیل احساس کاربر توسعه دهیم. برای توسعه چنین سیستمی به سوالات زیر پاسخ دهید. پاسخ‌ها در زبان برنامه نویسی Python کدنویسی شده باشند.

۱. ۱۰ کالای پراستناد (معروف) کدامند؟
۲. نحوه رتبه دهی کاربران به معروف‌ترین کالا چگونه است؟
۳. نظرات کاربران را تحلیل احساس کنید.
۴. نظرات کاربران را دسته‌بندی کنید (می‌توانید از منطق متفاوتی برای برچسب‌گذاری استفاده کنید).
۵. نظرات مثبت و منفی چه ویژگی‌هایی دارند. آنها را تحلیل کنید.
۶. برای یک محصول خاص مثبت‌ترین و منفی‌ترین نظرات را تحلیل کنید.

مسئله ۵: سیستم پیشنهاددهنده

می‌خواهیم با استفاده از مجموعه داده song_data یک سیستم پیشنهاددهنده توسعه دهیم. این مجموعه داده از ۶ ویژگی تشکیل شده است: (۱) شناسه کاربر (user_id)، (۲) شناسه آهنگ (song_id)، (۳) تعداد دفعاتی که کاربر به آهنگی خاص گوش داده است (listen_count)، (۴) عنوان (title) و (۵) خواننده (artist) و (۶) ترکیب خواننده و عنوان آهنگ (song). با توجه به این مجموعه داده به سوالات زیر پاسخ دهید (پاسخ در زبان برنامه نویسی Python کدنویسی شده باشد).

۱. ۱۰ آهنگ‌های پربازدید کدامند.
۲. تعداد کاربران منحصربفرد این مجموعه داده را مشخص کنید.
۳. یک سیستم پیشنهادهنده توسعه دهید که بر اساس خواننده محبوب کاربر به او پیشنهادهایی را ارائه کند.

مسئله ٦: پیش‌بینی قیمت

می خواهیم با استفاده از مجموعه داده home_data یک سیستم پیش‌بینی کننده قیمت مسکن (Price) توسعه دهیم. به این مسئله در سه مرحله پاسخ دهید.

مرحله اول: یک مدل پیش‌بینی کننده خطی توسعه دهید که براساس مساحت قابل سکونت (sqft_living) قیمت را پیش‌بینی کند سپس با کمک معیار RMSE (root means squared error) مدل خود را ارزیابی کنید.

مرحله دوم: یک مدل پیش‌بینی کننده خطی براساس ویژگی‌های زیر توسعه دهید:

- Number of bedrooms (bedrooms)
- Number of bathrooms (bathrooms)
- square foot of living space (sqft_living)
- square foot of the lot (sqft_lot)
- number of floors (floors)
- ZIP code (zipcode)

مرحله سوم: می‌خواهیم قیمت ۳ خانه را با کمک دو مدل ساخته شده در بالا پیش‌بینی کنیم. مشخصات این خانه‌ها به شرح زیر است (تصویر این خانه‌ها در فایل House-Data_Image.zip موجود است):

- خانه شماره ۱: id=5309101
- خانه شماره ۲: id=1925069082

- خانه شماره ۳: اطلاعات این خانه به شرح زیر است:
  - Number of bedrooms (bedrooms): 8
  - Number of bathrooms (bathrooms): 25
  - square foot of living space (sqft_living): 50000
  - square foot of the lot (sqft_lot): 225000
  - number of floors (floors): 4
  - ZIP code (zipcode): 98039

در این مرحله به دو سوال زیر پاسخ دهید:

۱. قیمت پیشنهادی شما توسط دو مدل ساخته شده در مرحله ۱ و مرحله ۲ چقدر است؟

۲. به نظر شما چرا در مورد خانه شماره ۱ مدل مرحله ۱ بهتر از مدل مرحله ۲ پیش‌بینی می‌کند اما در مورد خانه‌های شماره ۲ و ۳ (خصوصا شماره ۳) پیش‌بینی مدل شماره ۲ بهتر است.