پروژه اول درس داده کاوی پیشرفته – زمستان ۱۴۰۰

دانشکده علوم رایانه و فن آوری اطلاعات
دانشگاه تحصیلات تکمیلی علوم پایه زنجان

در یک پروژه تحقیقاتی، محققان دو گروه از افراد شامل افراد مبتلا به چاقی و افراد نرمال را مورد مطالعه قرار داده‌اند. در این پروژه، یکی از انواع داده، داده‌های ژنتیکی، است. پس از تحلیل این داده‌ها، جهش‌های ژنتیکی که در هر یک از افراد جمعیت مورد مطالعه وجود دارد، به دست آمده است. شما بعنوان یک تحلیلگر داده قرار است جهش‌هایی که در جمعیت نرمال و همچنین جمعیت چاق وجود دارد را بطور جداگانه استخراج کنید و جهش‌های با فراوانی بالا را نشان دهید.

فایل‌های داده شده LIHC.csv و LIHC-Normal.csv به ترتیب مربوط به گروه نرمال و مبتلا به چاقی هستند. فرمت داده‌ها با ذکر مثال در ادامه توضیح داده می‌شود. این مثال چند سطر اول از فایل گروه نرمال است:

```
SMARCA1-X-128645791-Missense_Mutation-(TCGA-DD-AAVP-01A-11D-A40R-10)
GCM1-6-52993306-Missense_Mutation-(TCGA-DD-AAVP-01A-11D-A40R-10)
TNFRSF21-6-47251829-Missense_Mutation-(TCGA-DD-AAVP-01A-11D-A40R-10)
NLRP14-11-7078963-Missense_Mutation-(TCGA-DD-AAVP-01A-11D-A40R-10)
```

قسمت سبزرنگ نام جهش را نشان می‌دهد. سپس یک خط تیره و سپس در داخل پرانتز، شناسه‌ی بیمار قرار دارد. در مثال بالا هر چهار جهش مربوط به یک بیمار هستند. وظیفه شما یافتن جهش‌های پرتکرار در این داده است. منظور از تکرار جهش، وقوع آن در بیش از یک بیمار است.

توضیحاتی درباره انجام پروژه:

- اگر در مقطع کارشناسی ارشد هستید می‌توانید در گروه‌های دو نفره پروژه را انجام دهید.
- برای انجام این پروژه می‌توانید از زبان R، پایتون یا متلب استفاده کنید (توصیه به پایتون است). توصیه می‌شود زبانی را انتخاب کنید که هر دو نفر گروه به آن تسلط دارید زیرا در هنگام تحویل پروژه ممکن است از شما خواسته شود تغییراتی در کد انجام داده و کد را مجدد اجرا کنید.
- در اجرای پروژه مجاز به استفاده از کامپیوتری هستنید که حداکثر دارای 16G حافظه RAM است.

- در تحویل هم کد (دارای کامنت و خوانا) و هم مستندات پروژه شامل روش حل مسأله، و توضیحات مناسبی از خروجی لازم است. مستند نهایی به صورت pdf و با فونت مناسب تحویل داده شود.
- در صورتیکه سوالی داشتید می‌توانید در گروه WhatsApp درس سوال کنید تا همه از پاسخ آن استفاده کنند.
- مهلت تحویل پروژه روز یکشنبه ۱۵ اسفندماه ساعت ۲۳:۵۹ است.
- پروژه‌های خود را طبق روال قبلی به حل تمرین‌های درس ارسال نمایید. جزئیات ایمیل ارسالی توضیح داده خواهد شد.
- تمامی فایل‌های خود شامل کد و مستندات را در یک فایل زیپ قرار دهید. اسم فایل را برابر با نام خانوادگی اعضای گروه قرار دهید. برای مثال اگر گروهی متشکل از آقای شامی و خانم صدیقیان باشد نام فایل زیپ، shami_sadighian.zip خواهد بود.

موفق باشید
زهرا نریمانی