بسمه تعالی
هوش مصنوعی و سیستم‌های خبره
پاییز ۱۴۰۰

تمرین سری **دو کامپیوتری**
دانشکده مهندسی و علوم کامپیوتر

مدرس: دکتر مهرنوش شمس‌فرد
تاریخ تحویل: ۱۷ آذر ۱۴۰۰

تحلیل لاگ‌های سرور به کمک کتابخانه‌های پایتون

هدف اصلی این تمرین تقویت مهارت های برنامه‌نویسی شما در سه کتابخانه‌ی مهم در حوزه یادگیری ماشین و علوم داده به نام‌های NumPy، Pandas و Matplotlib می‌باشد.

ابتدا فایل اولیه LogServerAnalysis.ipynb برای شروع این تمرین را دانلود کنید. وظیفه شما برای این تمرین، کامل کردن کد های داخل این فایل است. به این صورت که ۱۰ قسمت اصلی در این فایل قرار دارد و شما می‌بایستی فقط قسمت‌هایی که با *Your code here* مستندسازی شده‌اند را کامل کنید.

لطفا به نکات زیر حتما توجه کنید:

۱. در باقی کدهایی که در این فایل نوشته شده اند و امضای توابع نوشته شده در این فایل تغییری ایجاد نکنید.

۲. برای این تمرین می‌بایستی یک گزارش متنی برای سوالاتی که از شما خواسته‌شده تا تحلیلی داشته باشید تهیه فرمایید. نام این فایل گزارش را Report.pdf قرار دهید. همچنین می‌توانید نمودارهایی که در حین تکمیل کدها ایجاد کرده اید را وارد گزارش متنی نمایید.

۳. پس از تکمیل کدها، تمامی توابع پیاده‌سازی شده(آنهایی که با *Your code here* مستندسازی شده اند) را داخل یک فایل پایتون به نام Utils.py کپی پیست کنید. دقت فرمایید که در صورت انجام ندادن این کار نمره‌ای دریافت نخواهید کرد.

۴. تمیز بودن و زیبا و خوانا بودن نمودارهایی که رسم می‌کنید به شدت تاثیر مستقیمی در نمره شما خواهد داشت. همانطور که بالاتر گفته شده است، هدف اصلی این تمرین ارتقای دانش برنامه‌نویسی در کتابخانه‌های علوم داده در پایتون است.

۵. فایل نهایی که آپلود می‌کنید باید **تنها** شامل **سه** فایل باشد: Report.pdf، LogServerAnalysis.ipynb و Utils.py.

دیتاست

برای این تمرین از دیتاست لاگ های nginx سرور وبسایت فروشگاهی ایرانی zanbil.ir استفاده شده است. حجم این دیتاست 3.3GB است که شامل بیش از یک میلیارد نمونه از لاگ‌های به فرمت رایج nginx می‌باشد. یک نمونه از این داده‌ها را در اینجا می‌بینید:

```
31.56.96.51 - - [22/Jan/2019:03:56:16 +0330] "GET /image/60844/productMode
l/200x200 HTTP/1.1" 200 5667 "https://www.zanbil.ir/m/filter/b113" "Mozill
a/5.0 (Linux; Android 6.0; ALE-L21 Build/HuaweiALE-L21) AppleWebKit/537.36
(KHTML, like Gecko) Chrome/66.0.3359.158 Mobile Safari/537.36" "-"
```

هدف این تمرین **تمیزسازی**، **استخراج ویژگی** و **مهندسی ویژگی** این دیتاست خام است. در علم داده، فرآیندی قبل از هر تسک تحلیل داده وجود دارد که به آن Exploratory Data Analysis یا به اختصار EDA گفته می شود. با انجام EDA بر روی هر دیتاستی، می‌توان:

- بیشترین درک از دیتاست را داشت.
- از ساختارهای نهان موجود در دیتاست پرده پوشانی کرد.
- ویژگی‌های مهم استخراج کرد.
- تشخیصی از ناهنجاری ها و داده های پرت داشته باشیم.
- فرضیات اولیه‌مان را تست کنیم.

در واقع با رسم نمودارهای گرافیکی، استخراج جداول مهم و تحلیل ویژگی های آماری دیتاست خود را آماده برای پیاده سازی یک مدل یادگیری ماشین می‌کنیم.

دیتاست خام را می‌توانید از https://www.kaggle.com/eliasdabbas/web-server-access-logs دانلود کنید. فایلی که باید دانلود شود access.log نام دارد که حجم فشرده سازی شده‌ی آن تقریبا 264MB است.

صورت سوالات و باقی توضیحات در فایل LogServerAnalysis.ipynb به صورت شفاف و واضح توضیح داده شده است.

لطفا به نکات زیر توجه فرمایید:

- فرمت نام‌گذاری فایلی که آپلود می‌کنید **حتماً** به صورت**[student id][student name]** باشد.
- در صورت مشاهده هرگونه تقلب نمره **صفر** برای تکلیف در نظر گرفته می‌شود.
- می‌توانید سوالات و ابهامات خود را از **mohammad99hashemi@gmail.com** بپرسید.