به نام خدا

مهارت‌های پیشرفته کار با کامپیوتر (بهار ۱۴۰۴)

تمرین کامپیوتری ۴

مهلت ارسال: ۱۴۰۴/۰۳/۲۸

استاد درس: دکتر دوستی

دستیاران طراح: شهزاد ممیز، فرشته باقری

بازبینی: علی خرم‌فر

قوانین و ملاحظات

نحوه ارسال تمرین:

- تمامی فایل‌ها باید در یک فایل فشرده با نام ECS-CA4-StudentID ارسال شوند.
- کدهای مربوط به هر بخش را با نام مناسب بر اساس جدول انتهای همین فایل ذخیره کرده و همراه گزارش ارسال کنید.
- تمامی کدهای ارسال‌شده باید امکان اجرای مجدد داشته باشند. اگر تنظیمات خاصی برای اجرا نیاز است، آن را ذکر کنید.
- کدهای ارسال‌شده باید توسط خودتان اجرا شده باشند و نتایج اجرا در فایل ارسالی مشخص باشد.

رعایت اصول آکادمیک و صداقت علمی:

- این تمرین می‌تواند به صورت **گروهی (دو نفر)** انجام شود. اما ارزیابی آن به صورت فردی خواهد بود.
- در صورت مشاهده تشابه در پاسخ، تمامی افراد درگیر نمره صفر دریافت خواهند کرد و موضوع به استاد گزارش خواهد شد.

استفاده از ابزارهای هوش مصنوعی:

استفاده از ابزارهایی مانند ChatGPT ، Gemini، Copilot و موارد مشابه مجاز است، اما تحت شرایط زیر:

- نحوه استفاده از این ابزارها را در گزارش خود توضیح دهید (ابزارهای استفاده‌شده، کاربردهای مشخص و موارد مرتبط).
- تمامی پرامپت‌ها و لینک‌های استفاده‌شده را در انتهای گزارش قرار دهید.
- عدم ارائه این اطلاعات به‌منزله سرقت علمی محسوب شده و منجر به نمره صفر خواهد شد.

مهلت ارسال و جریمه تأخیر:

- امکان ارسال تمرین با تأخیر تا ۲ روز و به ازای هر روز تاخیر ۱۰ درصد جریمه وجود دارد.
- تاخیر به صورت ساعتی محاسبه شده و پس از دو روز تأخیر، تمرین پذیرفته نخواهد شد.

ارزیابی حضوری:

- ارزیابی تمرین به‌صورت حضوری انجام خواهد شد.
- محل ارزیابی: آزمایشگاه NLP ، طبقه منفی یک، دانشکده مهندسی برق و کامپیوتر شماره ۲.

توضیحات تمرین

- در صورت بروز مشکل با ایمیل‌های زیر در ارتباط باشید:

سوال ۱ و ۲ : shmomayez@gmail.com

سوال ۳ : fereshte12bagheri@gmail.com

سوال ۱. REGULAR EXPRESSION (۴۵ نمره + ۱۰ نمره امتیازی)

بخش اول) سیستم اعتبارسنجی برای ورودی فرم‌ها:

در این پروژه، قصد داریم یک سیستم اعتبارسنجی داده‌ها را برای فرم‌های ورودی مختلف ایجاد کنیم. در دنیای برنامه‌نویسی، اعتبارسنجی ورودی‌ها یکی از مهم‌ترین وظایف در هر برنامه است. فرض کنید شما مدیر یک سیستم هستید و قصد دارید مقادیر وارد شده در فرم‌های مختلف را اعتبارسنجی کنید تا اطمینان حاصل شود که داده‌ها صحیح و مطابق با قوانین تعیین شده هستند.

در این پروژه، از متا پروگرمینگ برای ایجاد توابع اعتبارسنجی استفاده می‌شود که کد آن به شما داده شده است و وظیفه ی شما تنها تکمیل TO DO# های موجود در کد هست که مربوط به Regular expression می باشد. متا پروگرمینگ به معنای نوشتن کدی است که خود به طور دینامیک کدهای دیگری را تولید می‌کند. این به ما این امکان را می‌دهد که توابع اعتبارسنجی را بدون نوشتن هر بار کد تکراری، به صورت خودکار برای هر فیلد ایجاد کنیم. در این سیستم، از تکنیک‌های متا پروگرمینگ برای تعریف قوانین اعتبارسنجی به صورت داخلی در یک کلاس پایه استفاده شده است، به طوری که برای هر فیلد (مثل ایمیل، شماره تلفن، پسورد و ...) تابعی مجزا به طور خودکار ساخته می‌شود.

در اینجا ما از یک متا کلاس به نام ValidatorMeta استفاده می‌کنیم که وظیفه‌ی ایجاد توابع اعتبارسنجی مختلف را به صورت خودکار به عهده دارد. این متا کلاس به کمک قوانین اعتبارسنجی که در یک دیکشنری قرار داده شده‌اند، توابعی ایجاد می‌کند که به راحتی می‌توانند داده‌ها را اعتبارسنجی کنند.

۱. Email: ایمیل باید فرمت صحیحی داشته باشد و شامل مجموعه‌ای از حروف، اعداد و کاراکترهای "@" و "." در موقعیت‌های مناسب باشد. همچنین، باید با پسوندهای .com یا .org پایان یابد.
(ورودی صحیح: "example@example.com"، ورودی نادرست: "example.com")

۲. Phone_Number: شماره تلفن باید با +۹۸ شروع شده و به دنبال آن ۸ رقم وارد شود.

۳. Password: رمز عبور باید بین ۸ تا ۱۲ کاراکتر باشد و حداقل شامل یک حرف بزرگ، یک حرف کوچک، یک عدد و یک کاراکتر خاص از مجموعه‌ی {"*","@", "٪", "!"} باشد.

۴. Product_Code: کد محصول باید از ۲ حرف بزرگ، ۲ تا ۴ رقم و یک حرف کوچک اختیاری تشکیل شده باشد. به صورت اختیاری در برخی محصولات version آن به صورت -v نیز می‌تواند اضافه شود (محدوده نسخه‌ها از v1 تا v99).(ورودی صحیح: "AB1234-v1"، ورودی نادرست: "A123B")

۵. Stop_Word: این اعتبارسنجی باید کلمه "Stop" یا "stop" را به عنوان یک کلمه مجزا شناسایی کند، و اطمینان حاصل کند که این کلمات جزئی از یک کلمه بزرگتر نیستند و به هیچ علامت نگارشی قبل یا بعد خود متصل نمی‌شوند.(ورودی صحیح: "Stop"، ورودی نادرست:"Stopped")

۶. Repeated_Phrase: این اعتبارسنجی باید عبارات "some students" یا "many employees" را شناسایی کند و اجازه دهد که چندین بار "students" یا "employees" تکرار شوند. با این حال، عبارت‌هایی مانند " some students collaborate with many employees" باید رد شوند، زیرا دومین بار باید مشابه اولین بار باشد. همچنین، می‌تواند فضاها یا علائم نگارشی بین کلمات وجود داشته باشد.(ورودی صحیح: "some students like some students")

۷. Date: تاریخ باید به فرمت YYYY/MM/DD باشد و شامل اعتبارسنجی اضافی برای مدیریت صحیح ماه‌ها (هیچ ۳۱ در ماه‌هایی که کمتر از ۳۱ روز دارند) باشد. (ورودی صحیح: "۱۴۰۴/۰۳/۲۰" ، ورودی نادرست: "۱۴۰۴/۱۳/۳۲")

۸. Quotation: متن باید داخل کوتیشن‌های تک یا دوتایی باشد و فقط شامل حروف، اعداد و فضاها باشد. همچنین کوتیشن‌ها باید متوازن باشند، به این معنا که هیچ کوتیشن تو در تو نباید وجود داشته باشد.

۹. Parenthesis: پرانتزهای باز و بسته باید جفت و متوازن باشند. هیچ پرانتز اضافی یا بدون جفت نباید در متن وجود داشته باشد. ورودی صحیح: "This is a valid parentheses"، ورودی نادرست: "(This is an invalid parentheses")

نحوه استفاده از توابع اعتبارسنجی:

برای هر فیلد یک تابع اعتبارسنجی به صورت خودکار با استفاده از متا پروگرمینگ ایجاد می‌شود. این توابع به شما این امکان را می‌دهند که داده‌ها را به راحتی بررسی کنید. برای مثال، برای اعتبارسنجی ایمیل، می‌توانید تابع validate_email را فراخوانی کنید. این کار را برای سایر فیلد ها نیز انجام دهید.

بخش دوم) Web scrapping (۱۰ نمره امتیازی)

یکی از کاربردهای اصلی عبارات منظم (regex) در وب‌اسکرپینگ است که در زمینه دیتا ساینس نیز به‌طور گسترده‌ای مورد استفاده قرار می‌گیرد. در این تمرین، سعی کنید با استفاده از کتابخانه‌ی re در پایتون و بدون استفاده از کتابخانه‌های دیگر، اقدام به وب‌اسکرپینگ از سایت https://www.imdb.com/chart/top/ نمایید. هدف این است که اطلاعاتی از جمله Title، URL، Description،Rating، Rating Count، Content Rating،Genre وDuration را استخراج کنید. در حین انجام این کار، باید چالش‌ها و موانع احتمالی استفاده از regex برای وب‌اسکرپینگ را مطرح کنید و راه‌حل‌های ممکن برای مقابله با این مشکلات را نیز بیان نمایید.

بخش سوم) سوالات تشریحی:

۱-پنج مورد از کاربردهای Regex را نام ببرید.

۲- در صورتی که درحین انجام پروژه محدودیتی در استفاده از Regex وجود داشت، ذکر کنید و همچنین محدودیت‌های کلی استفاده از Regex را بیان نمایید.

سوال ۲. DEBUGGING: سوال تشریحی (۵ نمره)

برنامه شما در محیط توسعه به درستی کار می‌کند اما در سرور یا محیط‌های دیگر خطا می‌دهد. برای دیباگ کردن این مشکل چه کاری باید انجام دهید؟

سوال ۳. PROFILING (۵۰ نمره)

در این بخش شما با یک پایپ‌لاین پیش‌پردازش داده‌ها کار خواهید کرد و یاد می‌گیرید که چگونه عملکرد آن را پروفایل کنید، گلوگاه‌ها را شناسایی کرده و تکنیک‌های بهینه‌سازی را اعمال کنید. این پروژه به سه قسمت تقسیم شده است.

کد مورد نظر از لینک زیر قابل دسترسی است:

ECS-CA4-P3.ipynb

(برای اینکه بتوانید در سیستم خود کد را ران کنید می‌توانید دیتاست‌ها را دانلود کرده و از سیستم لوکال خود load کنید)

بخش اول: پروفایلینگ پایپ‌لاین پیش‌پردازش داده‌ها

شما یک پایپ‌لاین پیش‌پردازش داده‌ها دریافت می‌کنید که شامل مراحل مختلفی از جمله پاکسازی داده‌ها، تغییر شکل و استخراج ویژگی‌ها است. شما باید با استفاده از ابزارهایی که در کلاس یاد گرفته‌اید، عملکرد این پایپ‌لاین را بررسی کنید.

- cProfile: برای نظارت بر زمان اجرای هر تابع.
- line_profiler: برای شناسایی اینکه کدام خطوط کد بیشترین زمان را مصرف می‌کنند.
- memory_profiler: برای نظارت بر مصرف حافظه در طول اجرای پایپ‌لاین.
- py-spy یا perf: برای نظارت بر استفاده از CPU و حافظه در زمان واقعی.
- مدت زمان هر فراخوانی تابع.
- شناسایی گلوگاه‌ها و قسمت‌هایی از کد که بیشترین منابع (هم از نظر زمان و هم از نظر حافظه) را مصرف می‌کنند.

عملکرد هر مرحله از پایپ‌لاین را خلاصه کنید و توابعی که بیشترین زمان و حافظه را مصرف کرده‌اند، شناسایی کنید. نتایج پروفایلینگ شما و همچنین تحلیل شما از آن باید به صورت کامل در گزارش آورده شوند.

بخش دوم: Optimization

در مورد راهکارهای بهبود عملکرد کد قسمت بالا تحقیق کنید و حداقل ۳ مورد از آن‌ها را در کد استفاده کنید و بعد از آن نتایج profiling را با قسمت قبل مقایسه کنید. شما باید دو نسخه کد را از نظر زمان اجرا توابع مختلف، مصرف حافظه و همچنین مصرف CPU مقایسه کنید.

بخش سوم: مقایسه روش‌های مختلف برای اعمال شرط بر روی ستون DataFrame

در این بخش، شما یک شرط را روی یک ستون از DataFrame با استفاده از سه روش مختلف اعمال خواهید کرد:

- استفاده از `pandas apply`
- استفاده از `pandas map`
- استفاده از `numpy where`

فرض کنید در حال نوشتن یک library برای پایتون هستید و سه متد مختلف برای انجام یک کار را پیاده‌سازی کردید. چگونه آن ها را با هم مقایسه می‌کنید؟ توضیح دهید برای مقایسه این سه حالت، چه منابعی از سیستم و به طور کلی چه metric هایی باید با همدیگر مقایسه شوند.

با توجه به توضیحات خودتان این سه روش را برای اعمال یک شرط با هم مقایسه کنید و نتیجه و تحلیل خود را در گزارش خود بیاورید.

بخش چهارم: سوالات تشریحی

پروفایلینگ در محیط Production چگونه انجام می‌شود؟ در مورد مشکلات احتمالی و مواردی که باید در نظر بگیریم توضیح دهید.

پروفایلینگ عملکرد در برنامه‌هایmultithreded چگونه انجام می‌شود؟ تفاوت پروفایلینگ منابع را با یک برنامه تک ریسمانه توضیح دهید.

گزارش نهایی:

پس از اتمام سه بخش، شما باید یک گزارش نهایی شامل موارد زیر ارسال کنید:

۱. **خلاصه‌ای از نتایج پروفایلینگ** برای پایپ‌لاین اصلی و بهینه‌شده.
۲. **مقایسه عملکرد** روش‌های مختلف برای اعمال شرط بر روی ستون DataFrame.
۳. **پیشنهادات** بر اساس تحلیل داده‌های پروفایلینگ، با تمرکز بر بهبود عملکرد و بهترین شیوه‌ها برای کار با داده‌های بزرگ در پایتون.

این پروژه به شما تجربه عملی در پروفایلینگ و بهینه‌سازی کد پایتون می‌دهد و به شما کمک می‌کند تا نحوه استفاده از عملیات وکتوریزه و بهترین تکنیک‌ها برای اعمال شرط به‌طور بهینه را یاد بگیرید.

نحوه تحویل تمرین کامپیوتری ۴

فایلها را به صورت زیر نام‌گذاری کرده و همه را در یک فایل zip در سامانه ارسال کنید.

سوال	بخش	نام فایل‌ها
۱	الف	`CA4_Question1.ipynb` توضیحات را میتوانید به صورت فایل `pdf` جداگانه یا داخل ژوپیتر نوت بوک قرار دهید.
	ب	`CA4_Question1.ipynb` توضیحات را میتوانید به صورت فایل `pdf` جداگانه یا داخل ژوپیتر نوت بوک قرار دهید
۲	تمام بخش‌ها	موارد خواسته شده در صورت سوال در گزارش ارائه شود.
۳	تمام بخش‌ها	موارد خواسته شده در صورت سوال (کد `optimize` شده و گزارش)