# Home Work

## Generative AI

## python review

1. فایل تمرین را در پنل خود آپلود کنید.
2. title فایل تمرین به صورت (نام تمرین+نام و نام خانوادگی) به انگلیسی باشد.
3. فایل‌های پیوست را می‌توانید از اینجا دانلود کنید.
4. در صورتی که سوال و یا ابهامی دارید در گروه چت تلگرامی بپرسید.

## ۱. پردازش متن و تحلیل فراوانی کلمات

در این تمرین، شما باید یک برنامه‌ی پایتونی بنویسید که عملیات زیر را روی یک فایل متنی انجام دهد:

1. خواندن محتوای یک فایل متنی: برنامه‌ی شما باید یک فایل متنی را بخواند و محتوای آن را دریافت کند
2. حذف آدرس‌های ایمیل و وبسایت‌ها: از پترن‌های Regex برای حذف آدرس‌های ایمیل و آدرس‌های وبسایت استفاده کنید
   - پترن برای حذف ایمیل‌ها:

```
\b[A-Za-z0-9._%+-]+@[A-Za-z0-9.-]+\.[A-Z|a-z]{2,}\b
```

- پترن برای حذف آدرس‌های وبسایت‌ها (URL):

```
https?://\S+|www\.\S+
```

3. تقسیم متن به کلمات: بعد از حذف ایمیل‌ها و آدرس‌های اینترنتی، متن را به کلمات تقسیم کنید.
4. حذف کلمات متداول (Stop Words) فارسی: از لیست کلمات پرتکرار فارسی برای حذف کلمات بی‌اهمیت استفاده کنید. کلمات پرتکرار در زبان فارسی شامل کلماتی هستند که به‌طور معمول در جملات ظاهر می‌شوند ولی هیچ اطلاعات خاصی ندارند.
   - لیست کلمات پرتکرار فارسی (Stop Words):

```
و, در, به, از, که, این, را, با, است, برای, آن, یک, هم, تا, نیز, اما, یا, بر, اگر, هر, چون, باید, می, شد, کند, کرد, شده, دیگر, همه, نیک, که, اینجا, اینها, آنان, خود
```

5. حذف علائم نگارشی: تمام علائم نگارشی (مثل نقطه، ویرگول، علامت سوال، و غیره) باید از متن حذف شوند.
6. محاسبه‌ی فراوانی کلمات: بعد از پردازش متن، تعداد تکرار هر کلمه را محاسبه کنید و آن را در یک دیکشنری ذخیره کنید.
7. ذخیره‌ی نتایج در فایل JSON: نتایج حاصل از شمارش کلمات را در یک فایل JSON ذخیره کنید. این فایل باید شامل کلمات و تعداد تکرار آن‌ها به‌صورت دیکشنری باشد.

ساختار کلاس:

- __init__ → مقداردهی اولیه مسیر فایل و خروجیJSON
- read_file → خواندن فایل متنی
- clean_text → حذف لینک‌ها، ایمیل‌ها و علائم نگارشی با Regex
- remove_stopwords → حذف کلمات پرتکرار فارسی

- count_word_frequencies → محاسبه تعداد تکرار هر کلمه
- save_to_json → ذخیره دیکشنری کلمات در یک فایل JSON
- process → اجرای تمام مراحل به صورت یکپارچه

یک نمونه از این کلاس به فرمت py. در اختیار شما قرار داده شده است.

خروجی مورد انتظار:

برنامه‌ی شما باید یک فایل JSON به نام word_frequencies.json تولید کند که در آن کلمات و تعداد تکرار آن‌ها به‌صورت یک دیکشنری ذخیره شده باشند.

کد پایتون خود را در قالب یک فایل text_processor.py ارسال کنید.