یا علیم

مهلت ارسال: ۱۴۰۴/۰۲/۲۶

مبانی بازیابی اطلاعات و جستجوی وب

تمرین عملی دوم

- تمرینات عملی می تواند به صورت گروهی (حداکثر دو نفر) انجام شود. در این صورت اسم هر دو نفر در پاسخ درج شده و فقط یکی از آنها تمرین را در lms بازگذاری نماید.
- به تاریخ تحویل درج شده در این فایل (و نه تاریخ LMS) توجه فرمایید.
- پاسخ تمرین باید شامل بحث و بررسی نتایج باشد و صرف اشاره به کد یا نتایج کافی نیست.

پیاده‌سازی مدل TF-IDF و محاسبه شباهت کسینوسی بین اسناد

در این تمرین، با مفاهیم نمایش برداری متون (TF-IDF) و محاسبه شباهت کسینوسی(Cosine Similarity) آشنا خواهید شد. این روش یکی از پایه‌های مهم در بازیابی اطلاعات، جستجوی متون و سیستم‌های پیشنهاددهنده است.

شما باید یک برنامه بنویسید که موارد زیر را انجام دهد:

۱. یک مجموعه شامل چند سند متنی را دریافت کرده و آن‌ها را پیش‌پردازش کند (کوچک‌سازی، توکن‌سازی، حذف علائم، و در صورت نیاز حذف کلمات توقف).

نکات:

- می‌توانید از متن‌های تمرین قبلی(مربوط به Elasticsearch) استفاده کنید.
- اسناد می‌توانند به زبان فارسی یا انگلیسی باشند.
- هر سند حداقل در حد یک پاراگراف باشد.

۲. برای هر سند، مقادیر TF (Term Frequency)و سپس TF-IDFرا محاسبه کند.

۳. شباهت کسینوسی بین تمامی جفت اسناد را محاسبه کرده و به صورت زوج‌زوج نمایش دهد.

٤. سه کوئری (Query) براساس متن هایتان بنویسید و برای هر یک:

- ابتدا خودتان به صورت دستی و شهودی اسناد را بر اساس میزان ارتباط مرتب کنید.
- سپس، با استفاده از برنامه‌ی خود شباهت پرس‌وجو را با هر سند محاسبه کرده و خروجی را بر اساس میزان شباهت TF-IDF و شباهت کسینوسی مرتب کنید.

٥. برای هر کوئری، خروجی مدل را با لیست دستی خودتان مقایسه کرده و با استفاده از آن مقادیر زیر را محاسبه کنید:

- Precision

- Recall
- F1-Score

نکات:

۱. استفاده از کتابخانه‌هایی مثل ,math, numpy یا collections مجاز است.

۲. استفاده از TfidfVectorizerبرای این تمرین مجاز **نیست**، هدف پیاده‌سازی دستی است. البته در نظر داشته باشید که ممکن است به دلیل تفاوت در نحوه پیش پردازش متن و تفاوت در فرمول های محاسباتی به نتایج یکسانی نرسید.