به نام خداوند جان و خرد

پروژه – فاز ۲

نام درس : مبانی بازیابی اطلاعات

استاد درس: دکتر زهرا زجاجی

حل تمرین‌ها: محمدرضا توکلیان

مهلت تحویل: ۱۴۰۰/۰۴/۰۳

سامانه تحویل: lms.ui.ac.ir

در دومین فاز پروژه بازیابی اطلاعات شما باید مطالبی که در فصل ranking آموخته‌اید را پیاده‌سازی کنید.

در این فاز شما باید الگوریتم BM25 را برای scoring در سیستم بازیابی اطلاعات پیاده‌سازی کنید. در فایل ارسال شده حاوی کدهای فاز ۱ و ۲ در پوشه ranker سه فایل idf.py ، scorer.py و ranker.py قرار داده شده است.

مقدار score برای الگوریتم BM25 برای داکیومنت D و کوئری Q از طریق فرمول زیر محاسبه می‌شود.

$$\text{score}\ (D,Q) = \sum_{i=1}^{n} \text{IDF}\ (q_i) \cdot \frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D)+k_1\cdot\left(1-b+b\cdot\frac{|D|}{\text{avgdl}}\right)}$$

در فرمول بالا $f(q_i,D)$ مقدار term frequency برای q_i در داکیومنت D است. $|D|$ اندازه داکیومنت D است و avgdl میانگین اندازه داکیومنت‌هاست. مقادیر b و k_1 در این فرمول باید توسط شما تنظیم شوند . $\text{IDF}\ (q_i)$ مقدار IDF برای q_i است که از فرمول زیر بدست می‌آید.

$$\text{IDF}\ (q_i) = \ln\left(\frac{N-n(q_i)+0.5}{n(q_i)+0.5}+1\right)$$

که در فرمول بالا نیز N تعداد کل داکیومنت‌هاست. $n(q_i)$ تعداد داکیومنت‌هایی است که شامل q_i هستند.

نکات اجرایی:

- از کدهای نوشته شده در فاز ۱ بهتر است برای راحتی کار استفاده کنید.

- پروژه را در قالب فایل‌های آورده شده تحویل دهید اما در ساختار کلاس‌های نوشته شده آزادید که تغییر ایجاد کنید.