به نام خدا

دانشگاه صنعتی شاهرود

دانشکده مهندسی کامپیوتر

مبانی بازیابی اطلاعات و جستجوی وب

تمرین سری دوم

مهلت تحویل: ۱ آذر ۱۴۰۰

- لطفا تمرین‌ها را در قالب یک فایل پی دی اف صرفا در بخش ارسال تمرین‌های سیستم مدیریت آموزش مجازی دانشگاه آپلود کنید.
- هر سری از تمرین‌ها دارای مهلت تحویل بوده و باید در زمان معین بارگذاری شوند .به ازای هر روز تاخیر در ارسال تمرین‌ها، مقداری از نمره کسر خواهد شد.
- تمامی تمرین‌ها به صورت تایپ شده یا به صورت اسکن شده از دست نویس (خوانا و خوش خط) تحویل گرفته خواهد شد. برای هر سری از تمرین‌ها، حتما صفحه اولی که در آن نام، نام خانوادگی، شماره دانشجویی و عنوان تمرین وجود دارد، در نظر گرفته شود.
- متن تمرین را پس از مطالعه درس به زبان خودتان بنویسید (لطفا کپی نکنید!)

۱- در نمایه Permuterm هر permuterm vocabulary term به عبارت (یا عبارت‌های) واژگان اصلی [1] که از آن مشتق شده است، اشاره می‌کند (در واقع لیست پست این کلمه شامل کلمه اصلی است که از آن مشتق شده است). حال چند عبارت واژگان اصلی می‌تواند در لیست پست یک permuterm vocabulary term وجود داشته باشد؟

۲- کلمات دیکشنری نمایه permuterm را برای کلمه sing بنویسید و بگویید برای پرس و جوی "s*ng" چه چیزی جستجو شود؟

[1] **Vocabulary term(s)**

۳- داده ساختارها اصلی برای جستجوی کلمات نمایه (دیکشنری) را نام ببرید و توضیح دهید چه ضوابطی را هنگام استفاده از آنها باید در نظر گرفت؟

۴- فاصله کلمات ''roof'' و ''road'' را توسط الگوریتم levenshtein بیابید.

۵- ضریب جاکارد را برای دو جمله ی زیر بدست آورید.

Query: shahrood university of technology

Document: the main field of our university is technology.

۶- Variable byte code را برای posting list زیر حساب کنید. در صورت امکان به جای DocID از gap ها استفاده کنید.

Posting list(777,17743,294068,3125136)

۷- میزان فضای مورد نیاز مجموعه داده Reuters با بلاک هایی با اندازه k=8 وk=16 با در نظر گرفتن دیکشنری به صورت رشته با بلوک را حساب کنید.

تمرین پیاده‌سازی

فایل های دیتاست همراه تمرین را به وسیله Lucene ایندکس کنید. این دیتاست شامل چکیده و مقدمه تعدادی مقاله علمی در سایت arxiv می باشد. همه حروف را کوچک کنید و کلمات ایست را حذف کنید. ریشه یابی را به دلخواه خود می توانید انجام دهید. سپس حداقل پنج پرس و جو مطرح کنید که عملگرهای and، or، not و جستجوی عبارت و پرس و جویwildcard را شامل شوند و نتایج را بررسی کرده و گزارش کنید که آیا نتایج و رتبه بندی آنها منطقی هستند یا خیر. می توانید برای سادگی تنها بخش چکیده (abstract) فایل ها را ایندکس نمایید.