به نام خدا

تمرین اول درس جستجو و بازیابی اطلاعات در وب، «روش‌های سنتی بازیابی اطلاعات»

استاد درس: دکتر ممتازی

پاییز ۱۴۰۲– دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر

نکاتی در مورد این تمرین که نیاز به توجه و دقت دوستان دارد:

۱- برای ارسال پاسخ تمرین‌های این درس، **مجموعا ۱۰ روز** زمان تاخیر مجاز در نظر گرفته شده‌است و در صورت تجاوز مجموع زمان تاخیرها از مقدار در نظر گرفته‌شده، پاسخ ارسال‌شده مورد بررسی قرار نخواهد گرفت.

۲- برای طرفین مشارکت‌کننده در هرگونه کپی‌کردن، بدون اغماض، نمره **منفی ۱۰۰** در نظر گرفته می‌شود.

۳- آخرین مهلت ارسال تمرین، ساعت **۲۳:۵۵** روز **شنبه ۶ آبان ۱۴۰۲** می‌باشد. این زمان با توجه به شرایط، جمع‌بندی‌ها و زمان لازم برای سایر تمرین‌ها در نظر گرفته شده‌است و **قابل تمدید نمی‌باشد**.

۴- دوستان فایل ارسالی خود را به صورت فشرده و به صورت «شماره دانشجویی_HW1» مانند HW1_400131123 نام‌گذاری کنید. در این فایل باید مواردی نظیر کدها، فایل گزارش و سایر موارد موردنیاز در هنگام بررسی و نمره‌دهی وجود داشته باشد و تنها این فایل جهت نمره‌دهی در نظر گرفته می‌شود.

۵- زبان برنامه‌نویسی پاسخ این تمرین تنها می‌تواند **پایتون** باشد.

۶- به صورت مناسب کامنت‌های لازم را در کدهای خود قرار دهید. به صورتی که بتوان حداقل روال اجرا و موارد مورد نیاز را درک کرد.

۷- سعی کنید ابتدا تمامی سوالات و بخش‌ها را مطالعه کنید.

۸- استفاده از کتابخانه‌های آماده به جز موارد مطرح شده در تمرین مجاز **نمی‌باشد** و شما باید موارد خواسته‌شده را پیاده‌سازی نمایید.

۹- در صورت هرگونه سوال یا مشکل می‌توانید با تدریس‌یار درس از طریق ایمیل زیر در ارتباط باشید.

mohammad.naeimi+ir@aut.ac.ir

بخش اول: معرفی مجموعه‌داده

مجموعه‌داده[1] ارائه شده در این تمرین، شامل ۳ فایل queries، docs، qrels می‌باشد. این مجموعه‌داده مربوط به وظیفه پاسخ به سوال[2] است که مجموعه ای از مقالات دانشگاهی در مورد COVID-19 و تحقیقات مرتبط با ویروس کرونا می‌باشد. هدف ما در این تمرین این است که برای هر پرسش[3] منحصر به فرد در فایل queries، ۱۰ تا از سند[4]های مرتبط به آن، از بین سندهای موجود در فایل docs، استخراج شوند. در واقع تمام سندهای فایل docs فضای جست‌جوی شما هستند. در مجموعه‌داده به ازای هر پرسش در فایل queries، تعداد ۱۵ سند در فایل qrels، به عنوان اسناد مرتبط مشخص شده‌اند، که حکم داده طلایی[5] جهت ارزیابی را دارند.

فایل queries (شامل ۵۰ ورودی)

ویژگی	توضیحات
query_id	شماره‌ی یکتای پرسش
query	متن پرسش

فایل docs (شامل ۷۵۰ ورودی)

ویژگی	توضیحات
doc_id	شماره‌ی یکتای سند
document	متن سند

فایل qrels (شامل ۷۵۰ ورودی)

ویژگی	توضیحات
query_id	شماره‌ی یکتای پرسش
doc_id	شماره یکتای سند مرتبط

بخش دوم: بازیابی با استفاده از مدل فضای‌برداری (۳۵ امتیاز)

با استفاده سندهای موجود در فایل docs، سندها را به صورت بردار TF-IDF نمایش بدهید. این کار را برای پرسش‌های فایل queries نیز انجام دهید. سپس با استفاده از معیار فاصله‌ی کسینوسی[6]، ۱۰ سند مرتبط با پرسش را مشخص کنید.

- طول بردار باید حداقل برابر با ۱۰۰۰ باشد به این معنا که ۱۰۰۰ کلمه‌ی برتر را برای ساخت بردار ملاک قرار دهید.
- به یاد داشته باشید که بردار سازنده‌ی پرسش‌ها و سندها، هردو بر اساس فایل docs ساخته شوند.
- درصورتی که پرسش دارای واژه‌ی جدیدی بود که در سندهای docs وجود نداشت، آن کلمه را نباید در بردار درنظر بگیرید.

---

[1] Dataset
[2] Question Answering
[3] Query
[4] Document
[5] Gold
[6] Cosine Similarity

بخش سوم: بازیابی با استفاده از مدل احتمالاتی BIM (۲۰ امتیاز)

با استفاده از مدل BIM، ۱۰ سند مرتبط برای هر پرسش را مشخص کنید.

- مقدار p_t را با در نظر گرفتن سه مقدار ثابت ۰.۳، ۰.۵، ۰.۷ امتحان کنید.
- مقدار u_t را برابر با $\frac{df_t}{N}$ در نظر بگیرید. (N تعداد کل سندها و df_t تعداد سندهایی است که در آن‌ها واژه‌ی t وجود دارد.)

بخش چهارم: بازیابی با استفاده از مدل احتمالاتی BM25 (۲۰ امتیاز)

با استفاده از مدل BM25، ۱۰ سند مرتبط برای هر پرسش را مشخص کنید.
تاثیر ابرپارامترهای مختلف مانند b و k_1 را با انتخاب سه مقدار مختلف برای هر کدام بررسی نموده و در گزارش ذکر کنید.

بخش پنجم: استفاده از روش‌های ارزیابی (۲۵ امتیاز)

روش‌های پیاده‌سازی‌شده در بالا را با استفاده از معیارهای ارزیابی P@5، P@10، MAP و MRR ارزیابی نموده و نتایج به‌دست‌آمده را در گزارش خود ذکر کنید.

- تمامی معیارهای ارزیابی مورد نظر را باید پیاده‌سازی نمایید.
- برای محاسبه معیارهای ارزیابی از فایل qrels به عنوان برچسب درست استفاده کنید، در واقع شما به ازای هر پرسش در فایل qrels، سندهای بازیابی شده، اگر در این فایل مقابل سند مربوطه بود، به عنوان پاسخ صحیح لحاظ شود و اگر نبود به عنوان پاسخ نادرست درنظر گرفته شود.
- با توجه به این که تعداد کل سندهای مرتبط برای هر پرسش در فایل qrels مشخص است، برای محاسبه‌ی معیار AP، مخرج کسر را برابر با تعداد کل سندهای مرتبط قرار دهید.

بخش آخر: برخی نکات در مورد گزارش و تمرین

- دادگان مطرح شده در این تمرین و تمامی بخش‌ها همراه با صورت تمرین در سایت درس قرارداده شده است.
- در این تمرین شما مجاز به استفاده کتابخانه‌های زیر و موارد مشابه و هم‌کاربرد با آن‌ها می‌باشد:

numpy, scipy, pandas, genism, pickle, tensorflow, pytorch, keras

- در این تمرین سعی شده است علاوه بر آشنایی شما با کاربرد مباحث ارائه‌شده در کلاس و لمس بهتر آن‌ها، خلاقیت و حل چالش شما نیز ارزیابی شود. لذا در صورتی که در این تمرین چالشی وجود دارد که شما راه‌حلی برای آن ارائه دادید و استفاده کردید، آن را در گزارش بیان کنید. اما اگر مشکلی بزرگ وجود دارد که نیاز به بررسی مجدد دارد، آن را از طریق ایمیل با تدریس‌یاران درس مطرح کنید.
- در صورتی که هر گونه پیش‌پردازش بر روی دادگان انجام دادید آن را در گزارش خود بیان کنید.
- این تمرین **۱ نمره** از بارم کلی تمرین‌های شما را با توجه به پوشش مباحث و حجم تمرین دارد. امتیاز این تمرین از ۱۰۰ محاسبه می‌شود که بارم هر بخش مشخص شده است.
- در تمامی بخش‌ها، میزان نتایج در ارزیابی شما تاثیر چندانی ندارند (مگر اینکه بسیار دور باشد). بلکه میزان تسلط، دیدگاه و پیاده‌سازی، تحلیل‌ها و خلاقیت شماست که در نمره شما تاثیر مستقیم دارد و بر اساس این موارد مورد ارزیابی قرار می‌گیرید.

موفق باشید

محمد نعیمی