به نام ایزد منان

تمرین اول درس بازیابی اطلاعات، «روش‌های سنتی بازیابی اطلاعات»

استاد درس: دکتر ممتازی

پاییز ۱۴۰۱– دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر

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

۱- در جدول زیر نحوه اعمال جریمه تاخیر در ارسال تمرین‌ها ذکر شده است.

میزان تاخیر (روز)	میزان جریمه
۱ الی ۲ روز	هر روز ۵٪
۲ الی ۶ روز	هر روز ۱۰٪

در صورتی که بین ۷ تا ۱۴ روز تاخیر داشته باشید، نمره شما از ۵۰٪ محاسبه می‌شود و پس از این بازه با توجه به سایر تمرین‌ها و زمان تحویل، به تمرین ارسالی نمره‌ای تعلق نمی‌گیرد.

۲- هرگونه کپی کردن باعث عدم تعلق نمره به تمامی افراد مشارکت کننده در آن می‌شود.

۳- آخرین مهلت ارسال تمرین، ساعت **۲۳:۵۵** دقیقه روز **۱ آذر** می‌باشد. این زمان با توجه به جمع‌بندی‌های صورت گرفته، شرایط و با توجه به سایر تمرین‌ها در نظر گرفته شده است و **قابل تمدید نمی‌باشد**.

۴- دوستان فایل ارسالی خود را به صورت فشرده و به صورت «شماره دانشجویی_HW01» مانند HW01_99131123 نام گذاری کنید. در این فایل باید مواردی نظیر کدها، فایل گزارش و سایر موارد مورد نیاز در هنگام بررسی وجود داشته باشد و صرفا این فایل در روز ارائه در نظر گرفته می‌شود.

۵- این تمرین دارای تحویل در محیط گوگل میت می‌باشد. زمان آن پس از یک هفته از پایان مهلت تمرین از طریق مودل درس اعلام می‌شود.

۶- زبان برنامه‌نویسی این تمرین می‌تواند پایتون، سی‌پلاس‌پلاس و یا جاوا باشد. (پیشنهاد ما پایتون است).

۷- کدهای خود را به صورت مناسب کامنت گزاری کنید. به صورتی که بتوان حداقل روال اجرا و موارد مورد نیاز را درک کرد.

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

۹- استفاده از هیچ کتابخانه آماده‌ای به جز موارد مطرح شده در تمرین مجاز **نمی‌باشد** و شما باید تمامی موارد را پیاده‌سازی کنید.

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

aliaz2197@gmail.com

بخش اول – معرفی دادگان

دادگان[1] ارائه شده در این تمرین شامل ۳ فایل `train`، `validation`، `test` می‌باشد. این دادگان مربوط به تسک *paraphrasing* است که در آن به ازای هر جفت متن سوال، وضعیت *paraphrase* بودن این دو با مقادیر ۰ و ۱ مشخص شده است. هدف ما در این تمرین این است که برای هر یک از *qid1* منحصر به فرد در فایل test، ۱۰ سوال *paraphrase* شبیه به آن، از *qid2*های فایل *train* استخراج شود. در واقع تمام *qid2*های فایل *train* فضای جست‌جوی شما می‌باشد، توجه کنید که باید مقادیر یکتای این ستون را برای جست‌وجو در نظر بگیرید.

مجموعه دادگان

ویژگی	توضیحات
ID	شماره‌ی یکتای پرس‌وجو
qid1	آیدی سوال اول
qid2	آیدی سوال دوم
Question1	متن سوال اول
Question2	متن سوال دوم
is_duplicate	برچسب *paraphrase*

بخش دوم – بازیابی با استفاده از مدل فضای برداری[2] (۳۰ امتیاز)

با استفاده از فایل *train*، متن سوالات*qid2* را به صورت بردار *TF-IDF* نمایش بدهید.سپس با استفاده از معیار فاصله‌ی کسینویسی[3]، ۱۰ سوال مشابه با سوالات موجود در ستون *qid1* فایل *test* را با توجه به بردارهای استخراج شده از فایل *train* بدست بیاورید.

*: طول بردار باید برابر با ۲۰۰۰ باشد به این معنا که ۲۰۰۰ کلمه‌ی برتر را برای ساخت بردار ملاک قرار دهید.

*: به یاد داشته باشید که بردار سازنده‌ی سوالات در هر سه فایل، بر پایه‌ی فایل *train* ساخته می‌شوند.

[1] Dataset
[2] Vector Space
[3] Cosine Similarity

❋: در صورتی که در فایل *validation* و *test* واژه‌ی جدیدی بود که در فایل *train* وجود نداشت، آن کلمه را در نظر نگیرید.

بخش سوم – بازیابی سندها، بر پایه‌ی مدل‌زبانی(۴۰ امتیاز)

در این بخش در دو قسمت با استفاده از مدلهای زبانی یونیگرام و بایگرام به بازیابی سندها میپردازیم.

۱. یک مدل زبانی یونیگرام برای سوالات *qid2* موجود در فایل *train* بسازید. برای هموارسازی این مدل زبانی از روش *Dirichlet* استفاده کنید. پارامتر این روش هموارسازی، μ است. تلاش کنید مقدار بهینه این پارامتر را با استفاده از فایل *validation* بدست آورید.

۲. در بخش دوم با استفاده از یک مدل بایگرام به بازیابی سوالات بپردازید. در یک مدل بایگرام میتوان مقدار $P(Q|D)$ را از رابطه‌ی زیر به دست آورید:

$$P(Q|D) = P(q_1|D) \times \prod_{i=2}^{n} P(q_i|q_{i-1}, D)$$

در این رابطه $P(q_1|D)$ با استفاده از احتمال یونیگرام هموارشده در بخش پیشین محاسبه می‌شود و $P(q_i|q_{i-1},D)$ از رابطهی هموارشده‌ی زیر محاسبه می‌شود:

$$P(q_i|q_{i-1},D)=\lambda\frac{C_{q_i,q_{i-1},D}}{C_{q_{i-1},D}} + \lambda P_{smoothed\ Unigram}(q_i|D)$$

در این رابطه $C_{qi,qi-1,D}$ تعداد رخدادهای بایگرام q_i,q_{i-1} در سند D و $C_{qi-1,D}$ تعداد رخدادهای واژه q_{i-1} در سند D می‌باشد.همچنین $P_{smoothed\ Unigram}$ همان احتمال هموار شده‌ی بخش نخست است. در این روش نیز مقدار بهینه ضریب ثابت λ را با استفاده از داده‌های فایل *validation* بدست آورید.

بخش چهارم – کار با روش‌های ارزیابی (۳۰ امتیاز)

هر سه روش مطرح شده در بالا را با استفاده از معیارهای ارزیابی *P@10*، *P@5*، *MAP* و *MRR* ارزیابی و گزارش کنید. هر سه روش مورد نظر را باید پیاده‌سازی کنید.

❋:برای محاسبه معیار ارزیابی از ستون *qid2* فایل تست به عنوان برچسب درست استفاده کنید، در واقع شما به ازای هر سوال در ستون *qid1* فایل تست، تمام سوالات مشابهی که بازگردانده می‌شوند، اگر در ستون*qid2* بود، به عنوان پاسخ صحیح لحاظ شود و اگر در این ستون نبود به عنوان پاسخ نادرست درنظر گرفته شود.

❋: با توجه به این که تعداد کل سوالات مرتبط برای هر سوال در فایل *test* مشخص شده است، بنابراین برای محاسبه‌ی معیار *AP* مخرج کسر را برابر با تعداد کل سوالات مرتبط در فایل *test* 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
```

- در این تمرین سعی شده است علاوه بر آشنایی شما با کاربرد مباحث ارائه شده در کلاس و لمس بهتر آن، خلاقیت و حل چالش شما نیز ارزیابی شود. لذا در صورتی که در این تمرین چالشی وجود دارد که شما راه حلی 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 پیش پردازش بر روی دادگان انجام دادید آن را در گزارش خود بیان کنید.

- این تمرین **۱** نمره از بارم کلی شما از تمرینات 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 شما در ارزیابی شما تاثیر چندانی ندارند (مگر اینکه بسیار دور باشد). بلکه میزان تسلط، دیدگاه و پیاده سازی، تحلیل‌ها و خلاقیت شماست که در نمره شما تاثیر مستقیم دارد و بر اساس این موارد مورد ارزیابی قرار می‌گیرید.

موفق باشد - عزیزی