به نام خدا

تمرین دوم درس جستجو و بازیابی اطلاعات در وب، «روش‌های مبتنی بر مدل‌زبانی و بردار معنایی عصبی»

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 ۱۱ آذر 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_HW2» مانند HW2_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] استفاده‌شده در این تمرین همان مجموعه‌داده استفاده‌شده در تمرین اول است. این مجموعه‌داده، 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

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

بخش دوم: بازیابی با استفاده از مدل‌های زبانی (۴۰ امتیاز)

۱. **یونیگرام**: یک مدل زبانی یونیگرام برای سندهای موجود در فایل docs بسازید و با استفاده از آن ۱۰ سند مرتبط برای هر پرسش را بازیابی نموده و معیارهای ارزیابی را گزارش کنید. برای هموارسازی این مدل‌زبانی از روش Jelinek-Mercer استفاده کنید. پارامتر این روش هموارسازی، ضریب ثابت λ_1 است. تلاش کنید مقدار بهینه این پارامتر را با استخراج و ارزیابی ۱۰ سند مرتبط برای هر پرسش به‌دست آورید. (حداقل ۵ مقدار مختلف را آزمایش نمایید.)

۲. **بایگرام**: با استفاده از یک مدل بایگرام به بازیابی ۱۰ سند مرتبط برای هر پرسش و گزارش معیارهای ارزیابی بپردازید. در یک مدل بایگرام میتوان مقدار $P(Q|D)$ را از رابطه‌ی زیر به‌دست 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:

$$P(q_i|q_{i-1}, D) = \lambda_1 \frac{TF_{q_i,q_{i-1},D}}{TF_{q_{i-1},D}} + \lambda_2 \frac{TF_{q_i,D}}{|D|} + (1 - \lambda_1 - \lambda_2) \frac{CF_{q_i}}{|C|}$$

در این رابطه $TF_{q_i,q_{i-1},D}$ تعداد رخدادهای بایگرام q_i, q_{i-1} در سند D و $TF_{q_{i-1},D}$ تعداد رخدادهای واژه q_{i-1} در سند D می‌باشد. همچنین $|D|$ تعداد واژگان سند، $|C|$ تعداد کل سندها و CF_{w_i} تعداد سندهای شامل واژه q_i می‌باشند. پارامتر λ_1، همان ضریب ثابت سوال اول است. در این روش نیز مقدار بهینه ضریب ثابت λ_2 را با استخراج و ارزیابی ۱۰ سند مرتبط برای هر پرسش به‌دست آورید.

بخش سوم: بازیابی با استفاده از بردارهای معنایی عصبی (۴۰ امتیاز)

۱. می‌خواهیم با استفاده از تعبیه واژگان سندها را بازیابی کنیم. برای به‌دست آوردن بردار معنایی هر پرسش یا سند، از بردارهای معنایی word2vec (در حالت skip-gram) کلمات موجود در آن‌ها میانگین بگیرید. سپس با استفاده از معیار شباهت کسینوسی[۶]، برای هر پرسش ۱۰ سند مرتبط را مشخص کرده و نتایج معیارهای ارزیابی را گزارش کنید. برای این سوال نتایج را در دو قسمت، با استفاده از روش‌های زیر میانگین را محاسبه نمایید. (استفاده از genism پیشنهاد می‌شود.)

الف) میانگین حسابی[۷].

ب) میانگین وزنی[۸]، با در نظرگرفتن مقدار TF-IDF هر واژه به عنوان وزن واژه.

۲. در این قسمت از مدل پیش آموزش‌دیده[۹] BERT برای به‌دست‌آوردن بردارهای معنایی استفاده کنیم. ابتدا متن سندهای فایل docs را در فضای‌برداری بازنمایی کنید. سپس با استفاده از معیار شباهت کسینوسی، ۱۰ سند مرتبط با هر پرسش فایل queries را از میان سندهای فایل docs به‌دست بیاورید و نتایج معیارهای ارزیابی را گزارش کنید. (برای مدل BERT، استفاده از Hugging Face transformers پیشنهاد می‌شود.)

---

[6] Cosine Similarity
[7] Arithmetic Mean
[8] Weighted Mean
[9] Pretrained

بخش چهارم: بازیابی با استفاده از مدل ترجمه (۲۰ امتیاز، اختیاری)

برای بازیابی سندهای مرتبط با هر پرسش از مدل ترجمه[10] مطرح‌شده در درس استفاده می‌کنیم. در مدل ترجمه، واژگانی که میزان شباهت کسینوسی نرمال‌شده میان بردارهای معنایی word2vec آن‌ها با هر واژه پرسش بیشتر از مقدار ۰.۷ است را استفاده کنید.

الف) با استفاده از این مدل، برای هر پرسش ۱۰ سند مرتبط را مشخص کرده و نتایج معیارهای ارزیابی را گزارش کنید.

ب) واژگان استخراج شده برای هر واژه‌ی پرسش را در گزارش خود قرار دهید.

بخش پنجم: استفاده از روش‌های ارزیابی (۲۰ 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. تحلیل خود از نتایج به دست آمده را نیز به صورت مختصر بیان 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.

[10] Translation Model

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