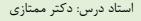
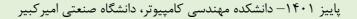
به نام ایزد منان

تمرین دوم درس بازیابی اطلاعات، «روشهای مبتنی بر شبکه عصبی برای بازیابی اطلاعات»









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.

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

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 نمی گیرد.

- ۲- هرگونه کپی کردن باعث عدم تعلق نمره به تمامی افراد مشارکت کننده در آن میشود.
- ۳- آخرین مهلت ارسال تمرین، ساعت **۱۳:۵۵** دقیقه روز ۱۰ **دی ۱۴۰۱** میباشد. این زمان با توجه به جمعبندیهای صورت گرفته، شرایط و با توجه به سایر تمرینها در نظر گرفته شده است و قابل تمدید نمی باشد.
- ۴- دوستان فایل ارسالی خود را به صورت فشرده و به صورت «شماره دانشجویی_۱۳۷۱» مانند ۱۳۷۱_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، سی پلاس پلاس 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.

kasra96.d@gmail.com

بخش اول – معرفی دادگان

دادگان ۱ ارائه شده در این تمرین دقیقا مشابه تمرین اول است. این دادگان شامل ۳ فایل validation ،train، 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 ۱۰٬ test سوال paraphrase شبیه به آن، از qid2های فایل train استخراج شود. در واقع تمام qid2های فایل train فضای جستجوی شما میباشد، توجه کنید که باید مقادیر یکتای این ستون را برای جستوجو در نظر بگیرید.

مجموعه دادگان

ویژگی	توضيحات
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 بردارهای معنایی (۲۵ امتیاز)

با استفاده از مدل BERT متن سوالات qid2 از فایل train را در فضای برداری بازنمایی کنید. سپس با استفاده از معیار فاصلهی کسینویسی^۲، ۱۰ سوال مشابه با سوالات موجود در ستون *qid1* فایل *test* را با توجه به بردارهای استخراج شده از فایل train بدست بیاورید.

¹ Dataset

² Cosine Similarity

بخش سوم - بازیابی با آموزش مدل شبکه عصبی (۴۵ امتیاز)

در این بخش باید با طراحی یک شبکه Siamese مدلی مبتنی بر شبکه عصبی را برای بازیابی سوالات آموزش در این بخش باید با طراحی یک شبکه Siamese مدلی مبتنی بر شبکه عصبی را برای بازیابی سوالات آموزش در در از دادههای موجود در فایل rain و train بدست بیاورید. سوال موجود در ستون qid1 فایل ۱۰، test سوال با بیشترین شباهت را از فایل train بدست بیاورید. * برای طراحی شبکه عصبی موردنظر آزاد هستید و می توانید از ساختارهای پیشنهادی زیر استفاده کنید:

- ورودی شبکه بازنمایی word2vec یا fasttext کلمات باشد.
 - لایههای شبکه با مدلهای LSTM یا BiLSTM باشد.
- ترکیب انتهای شبکه برای مقایسه دو متن فاصله کسینوسی باشد و یا کانکت دو بردار به یک لایه شبکه عصبی ساده داده شود.
 - سعی کنید با تست کردن حالتهای مختلف ذکر شده بهترین نتیجه را بدست آورید.

بخش چهارم – کار با روشهای ارزیابی (۳۰ امتیاز)

هر سه روش مطرح شده در بالا را با استفاده از معیارهای ارزیابی P@5 P@5 P@5 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 می شوند، اگر در ستون qid2 بود، به عنوان پاسخ صحیح لحاظ شود و اگر در این ستون نبود به عنوان پاسخ نادرست در نظر گرفته شود. *: با توجه به این که تعداد کل سوالات مرتبط برای هر سوال در فایل test مشخص شده است، بنابراین برای محاسبه 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 هم کاربرد با آنها میباشد:

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

- 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 تدریس یاران 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 می شود 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 می گیرید.

مو فق باشد - در ویشی