باسمه تعالی

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده مهندسی برق و کامپیوتر

درس بازیابی اطلاعات

پروژه پایانی – پاییز ۱۴۰۰

مقدمه:

زمینه مطالعاتی پاسخ‌دهی پرسش[1] به عنوان یکی از زمینه های مشترک بازیابی اطلاعات و پردازش زبان طبیعی[2] به حساب می‌آید. امروزه و با گسترش منابع اطلاعاتی در دسترس، جستجو و یافتن پاسخ مناسب و صحیح، فرآیندی زمانبر و پرهزینه خواهد بود. در این راستا این امکان وجود دارد که با بهره‌گیری از سیستم‌های پرسش و پاسخ، هدف هایی همچون رتبه بندی پاسخ های مرتبط، بازیابی مناسب‌ترین پاسخ، شناسایی پاسخ‌های مرتبط و موارد مشابه حاصل شوند.

در این پروژه شما عمل بازیابی پاسخ را با هدف رتبه بندی پاسخ های مرتبط انجام خواهید داد. مجموعه داده‌ای شامل ۴۴K زوج پرسش-پاسخ می باشد. در این مجموعه به ازای هر پرسش ده پاسخ با برچسب مناسب وجود دارد. این مجموعه داده به سه قسمت train, dev, test تقسیم شده است که داده‌های موجود در مجموعه تست بدون برچسب می‌باشند. هدف رتبه‌بندی ده پاسخ متناظر با هر پرسش است به گونه‌ای که پاسخ‌های با برچسب good در رتبه‌بندی بالاتر از پاسخ‌های با برچسب bad و potentially useful قرار گیرند.

گام اول: پیش‌پردازش

در ابتدا باید اسناد موجود را پیش پردازش کنید. در این مرحله اسناد را tokenize کنید، stop wordها را حذف کنید، کلمات نهایی را stem کنید. برای این کار می‌توانید از NLTK استفاده کنید. برای دیدن نمونه‌ای

[1] Question answering
[2] Natural language processing (NLP)

از کارهایی که می‌توانید با NLTK انجام دهید به https://www.nltk.org/book/ch01.html مراجعه نمایید. (برخی از این اعمال ممکن است در گام چهارم و پنجم نیازی نباشد.)

گام دوم: استخراج ویژگی

با توجه به آموخته‌های خود در کلاس درس مجموعه‌ای از ویژگی‌ها را استخراج کنید تا به کمک آنها بتوانید یک شبکه عصبی را آموزش دهید. این ویژگی‌ها می‌توانند مبتنی بر پرسش، مبتنی بر پاسخ، مبتنی بر ارتباط بین پرسش و پاسخ و یا سایر اطلاعات موجود در مجموعه داده باشند. ویژگی های خود را به همراه دلیل انتخاب هر ویژگی بیان کرده و تاثیر آن را در عملکرد نهایی رتبه‌بندی بررسی نمایید.

گام سوم: بازیابی پاسخ با کمک یک شبکه پرستپرون چندلایه [3]

به ازای هر پاسخ کاندیدای $C_i, i = 1,2, \dots ,10$ ، بردار ویژگی $x_i \in \mathbb{R}^d$ که یک بردار با d بعد مربوط به d ویژگی گام قبل است را بدست آورید و یک شبکه عصبی پیشخور را با کمک آنها و با توجه به داده‌های train، آموزش دهید. تنظیمات نهایی را براساس مجموعه داده dev اعمال نمایید. سپس از مدل نهایی بدست آمده برای بازرتبه‌بندی پاسخ‌های مرتبط با هر پرسش در مجموعه‌داده test استفاده نمایید.

گام چهارم: بازنمایی طیفی کلمات

یک روش جاسازی کلمات[4] مانند Glove را برای بازنمایی کلمات موجود در مجموعه داده‌ای بکار برید و با کمک آن یک شبکه مناسب را آموزش دهید. سپس از مدل بدست آمده برای بازرتبه‌بندی پاسخ‌های مرتبط با هر پرسش در مجموعه‌داده تست استفاده نمایید. (امکان انتخاب هر نوع شبکه ای در این بخش برای شما مقدور است. با ذکر این نکته که لازم است دلیل انتخاب شما برای انتخاب شبکه پیشنهادی توضیح داده شود. در این راستا می توانید از شبکه هایی همچون MLP، LSTM و یا موارد دیگر استفاده کنید.)

[3] Multi layer perceptron (MLP)
[4] Word embedding

گام پنجم(امتیازی): BERT finetuning

با کمک یک روش بازنمایی مبتنی بر بافتار مانند BERT بازیابی پاسخ را انجام دهید. برای اینکار زوج پرسش-پاسخ را در قالب ورودی به شبکه BERT که متناسب با هدف این پروژه بهینه شده است اعمال کنید . سپس از مدل بدست آمده برای بازرتبه‌بندی پاسخ‌های مرتبط با هر پرسش در مجموعه‌داده تست استفاده نمایید. در صورت نیاز می‌توانید از گونه‌های مختلف مدل BERT از جمله DistilBERT و RoBERTa نیز استفاده نمایید.

https://www.coursera.org/lecture/attention-models-in-nlp/fine-tuning-bert-EMBvt

https://huggingface.co/transformers/training.html

توضیحات:

- تحلیل نتایج حاصل از روش‌های فوق، مقایسه و بررسی نقاط ضعف و قوت آنها بخش مهمی از این پروژه را شامل می‌شود.

- معیار MAP برای نتایج حاصل از خروجی تست برای شما محاسبه خواهد شد.

- تمامی نتایج شما باید در یک فایل فشرده با عنوان IR_FIN_PRJ-studentNum تحویل داده شود. این پوشه باید شامل موارد زیر باشد:

۱- گزارش به فرمت PDF شامل شرح آزمایشهای انجام شده، پارامترهای آزمایش، نتایج و تحلیلها. (از توضیح دادن کد در گزارش خودداری نمایید. در صورت نیاز به توضیح بر روی کد کامنت بگذارید)

۲- یک پوشه به نام کد، که در آن فایل کدهای پروژه با نامگذاری و ساختار مناسب قرار می‌گیرند.

۳- یک پوشه به نام ranking که در آن رتبه‌بندی پیش‌بینی شده برای مجموعه تست قرار داده می‌شوند. فرمت خروجی لازم است که به شکل زیر باشد. در این فرمت

ستون اول=شماره سوال

ستون دوم=شماره پاسخ

ستون سوم= رتبه پیشنهادی (ترتیب نوشتن پاسخ‌ها در خروجی متناسب با ترتیب پاسخ‌ها فایل تست باشد)

ستون چهارم= میزان امتیاز محاسبه شده برای پاسخ مربوطه

ستون پنجم= برچسب true یا False متناظر با پاسخ می باشند. (true متناظر با برچسب good، و false متناظر با برچسب bad و PotentiallyUseful می باشند.)

Q1	Q1_C1	1	1	true
Q1	Q1_C2	2	0.9	false
Q1	Q1_C3	3	0.8	false
Q1	Q1_C4	4	0.7	false
Q1	Q1_C5	5	0.6	false
Q1	Q1_C6	6	0.5	false
Q1	Q1_C7	7	0.35	false
Q1	Q1_C8	8	0.24	false
Q1	Q1_C9	9	0.12	false
Q1	Q1_C10	10	0.1	false

- گزارش نهایی خود را حتما به صورت PDF در سایت درس بارگذاری نمایید.
- توجه کنید این تمرین باید در گروه‌های دو نفره انجام شود و پاسخ‌های ارئه شده باید نتیجه فعالیت گروه نویسنده باشد. (همفکری و به اتفاق هم نوشتن نیز ممنوع است.) در صورت مشاهده تقلب متاسفانه همه افراد متقلب نمره‌ی منفی دریافت می کنند.
- در صورت وجود ابهام در پروژه می توانید از طریق ایمیل‌های زیر با کمک مدرس‌های درس ارتباط برقرار کنید:

ghasemi.shima@gmail.com

golnar.afzali@ut.ac.ir

- مهلت تحویل : ۲۴ دی ماه

با آرزوی موفقیت و کامیابی