دانشکده مهندسی کامپیوتر

مبانی پردازش زبان و گفتار

نیم‌سال دوم ۱۴۰۱-۱۴۰۲

امتحان عملی

استاد: سید صالح اعتمادی

اساتید حل تمرین

هادی شیخی

مصطفی رستم‌خانی

غزل زمانی‌نژاد

عرفان موسوی

صادق جعفری

پنج‌شنبه ۴ خرداد ماه ۱۴۰۲

مدت امتحان ۱۸۰ دقیقه

قوانین امتحان

- صفحه نمایش شما در طول امتحان بصورت کامل باید ضبط شده و در انتهای امتحان روی فلش تحویل داده شود.
- استفاده از کدهای تمرین‌ها و ورک‌شاپ‌ها که خودتان زدید یا در زمان ورک‌شاپ در اختیار شما قرار داده شده، مجاز است.
- فقط و فقط منابع آنلاین زیر قابل استفاده هستند. استفاده از هر منبع دیگری تقلب محسوب می‌شود.
  - https://huggingface.co/docs
  - https://docs.python.org
  - https://numpy.org/doc
  - https://pytorch.org/docs/2.0
  - https://nlp.stanford.edu/projects/glove
- کپی کردن هر قطعه کدی از هر جایی تقلب محسوب می‌شود.
- هر گونه تماس با هر روشی از جمله گیت با هر کسی تقلب محسوب شده و منجر به مردودی در درس خواهد شد.
- امتحان از طریق GradeScope بوده و راس ساعت ۱۲:۳۰ بسته می‌شود. برای تمرین‌هایی که نوت‌بوک دارد، نوت بوک اجرا شده بدون پاک کردن خروجی را با همین اسم ارسال کنید. برای بخش‌های مرتبط با تمرین A3 و A4 فقط فایل‌های تغییر یافته را ارسال کنید. در نهایت همه فایل‌ها را یکجا در گریداسوپ آپلود کنید.

سوال ۱ (۵ نمره)

در فایل نوت‌بوک Q1_Embedding.ipynb تابع find_k_nearest_neighbors را برای پیدا کردن نزدیک‌ترین کلمات با معیار فاصله اقلیدسی پیاده‌سازی کرده و ۱۰ کلمه نزدیک به یک کلمه انتخابی خود را چاپ کنید.

سوال ۲ (۱۰ نمره)

برای لینک داده شده در نوت‌بوک Q2_Crawling.ipynb تمام جواب‌ها را پیدا کرده و ۵۰ کاراکتر ابتدای آنها را چاپ کنید.

سوال ۳ (۱۰ نمره)

با توجه به آشنایی که در تمرین A1 با امبدینگ‌های GloVe پیدا کردید و آشنایی که با نسبت کلمات به هم دارید، در نوت بوک Q0_A3.ipynb:

۱. کد لازم برای پیدا کردن نسبت بیمارستان (hospital) به آمبولانس (ambulance) را بنویسید. مثال: نسبت مرد به شاه مثل نسبت زن است به ملکه.

۲. کد لازم برای پیدا کردن یک مثال از تبعیض جنسیتی در رابطه با اسباب‌بازی (toy) پیدا کنید.

با توجه به نیاز به بارگذاری GloVe بهتر است نوت‌بوک را در گوگل‌کولب اجرا کنید.

سوال ۴ (۱۵ نمره)

در تمرین A3 در مدل طرح شده برای dependency parsing، یک لایه مخفی به همراه تابع فعال سازی ReLU بعد از اولین لایه و قبل از لایه آخر قرار دهید و تعداد نرون‌ها در این لایه را برابر با hidden_size قرار دهید. تغییرات لازم برای این کار را انجام دهید. (توجه داشته باشید که ممکن است بیش از یک تابع را تغییر دهید). برای اطمینان نسبی از درستی کد زده شده، میتوانید دستور python run.py را اجرا کنید و مشاهده کنید که آیا مدل شما آموزش میبیند یا دچار خطا می‌شود.

دقت کنید که چنانچه تمرین ۳ را انجام داده باشید، بدون هیچ تغییری دستور `run.py` با موفقیت اجرا می‌شود. لذا اجرا موفقیت‌آمیز آن نشانه حل درست این سوال نیست. کد شما برای حل این سوال بررسی و تصحیح می‌شود. کدهای اولیه لازم برای این سوال را از تمرین `A3` که فرستاده‌اید بردارید.

سوال ۵ (۲۰ نمره)

در تمرین `A4`، برای `project` کردن دو بردار به فضای مشترک، از `concat` کردن دو بردار استفاده شده است (در بردار `U_t` طبق نوشتار تمرین). حال شما به جای استفاده از `torch.cat` از `torch.matmul` استفاده کنید تا دو بردار را به یک فضای مشترک `project` کنید (به ابعاد بردار های ورودی توجه کنید و به این نکته نیز توجه کنید که ممکن است نیاز باشد بیش از یک تابع را تغییر دهید). برای اطمینان نسبی از درستی پیاده سازی انجام شده، از دستور `python sanity_check.py 1f` طبق تمرین استفاده کنید. مجددا دقت کنید که چنانچه تمرین `A4` را انجام داده باشید، بدون هیچ تغییری دستور بالا با موفقیت اجرا می‌شود. لذا اجرا موفقیت‌آمیز آن نشانه حل درست این سوال نیست. کد شما برای حل این سوال بررسی و تصحیح می‌شود. کدهای اولیه لازم برای این سوال را از تمرین `A4` که فرستاده‌اید بردارید.

سوال ۶ (۲۰ نمره)

برای این سوال از نوت‌بوک `Q6_Hugging_Face.ipynb` استفاده کنید. این همان فایل ورک‌شاپ است که در انتهای آن بخش‌ها امتحان اضافه شده است. قبل از ارسال «حتما» همه بخش‌های قبل از "`Exam`" را پاک کنید.

۱. دیتاست `https://huggingface.co/datasets/poem_sentiment` را بارگذاری کنید
۲. مدل و `tokenizer` با عنوان '`bert-base-uncased`' را برای تسک متناسب با دیتاست انتخاب و تنظیم کنید.
۳. عملیات `tokenization` را روی دیتاست انجام دهید.
۴. از `Trainer` برای تنظیم دقیق (`fine-tune`) کردن مدل برای ۵ ایپاک استفاده کنید.
۵. دقت مدل را روی بخش تست دیتاست محاسبه و گزارش کنید.

برای اجرای بدون مشکل لازم است نسخه `4.28` ترنسفومرز نصب شده باشد. لذا از اجرای سلول زیر اطمینان حاصل کنید.

```
!pip install -q transformers==4.28.0
!pip install -q datasets
```

سوال ۷ (۲۰ نمره)

برای این سوال از نوت‌بوک `Q7_Numpy_HuggingFace.ipynb` استفاده کنید. برای بخش اول تابع `softmax` را با استفاده از `NumPy` با `Temperature`های مختلف مطابق کدهای خواسته شده پیاده‌سازی، اجرا و تحلیل کنید.

در بخش دوم مطابق نوت‌بوک مدل زبانی `GPT2` را بارگذاری کرده و روش‌های `decoding` مختلف برای `generation` را روی آن آزمایش کرده و راه‌حلی برای عدم تکرار `n-gram`های دوتایی پیدا کنید و در نهایت نتایج آنرا با هم مقایسه و تحلیل کنید. برای تحلیل بهتر می‌توانید از متن اولیه‌های مختلفی استفاده کنید.

در ادامه بخش دوم، مدل `GPT2-Large` را در `HuggingFace` با عنوان `Exam-Part7-GPT2-Large` ارسال کنید.

برای نمره مثبت روش استفاده از پارامتر `Temperature` را در `generation` با مثال مطابق نوت‌بوک توضیح دهید.