دانشگاه صنعتی امیرکبیر

دانشکده مهندسی کامپیوتر

پروژه درس رایانش عصبی و یادگیری عمیق

پروژه‌ی سوم

هدف: استفاده از نقشه‌های خودسازمانده برای تولید تقویت شده با بازیابی

کد: پیاده سازی این پروژه را به زبان پایتون انجام دهید؛ در این فعالیت مجاز به استفاده از tensorflow یا pytorch یا jax می‌باشید. فایل‌های کد خود را بر اساس شماره سوال و زیر قسمت خواسته شده‌ی آن نام گذاری کنید (برای مثال می‌توان نام گذاری قسمت اول برای سوال سوم تمرین را بصورت P3_a_ preprocessing.py در نظر گرفت). فایل‌های ارسالی‌تان بایستی با فرمت py. یا ipynb.(با حفظ خروجی هر سلول) باشد. همچنین برای استفاده از شبکه‌های خودسازمانده از کتابخانه‌ی minisom استفاده کنید.

گزارش: ملاک اصلی انجام فعالیت، گزارش آن است و ارسال کد بدون گزارش فاقد ارزش است. برای این فعالیت یک فایل گزارش در قالب pdf تهیه کنید که دارای فهرست بوده و پاسخ‌ها بترتیب در آن قرار گرفته‌اند و نام، نام خانوادگی و شماره دانشجویی‌تان در قسمت چپ سربرگ تمامی صفحات تکرار شده است.

تذکر: مطابق قوانین دانشگاه هر نوع کپی برداری و اشتراک کار دانشجویان غیر مجاز بوده و با تمامی طرفین برخورد خواهد شد. استفاده از کدها و توضیحات اینترنت به منظور یادگیری صرفا با ارجاع به آن بلامانع است، اما کپی کردن آن غیرمجاز است.

راهنمایی: در صورت نیاز میتوانید سوالات خود را در خصوص پروژه از تدریسیارهای درس، از طریق ایمیل زیر یا در گروه تلگرامی بپرسید.(لینک گروه تلگرامی)

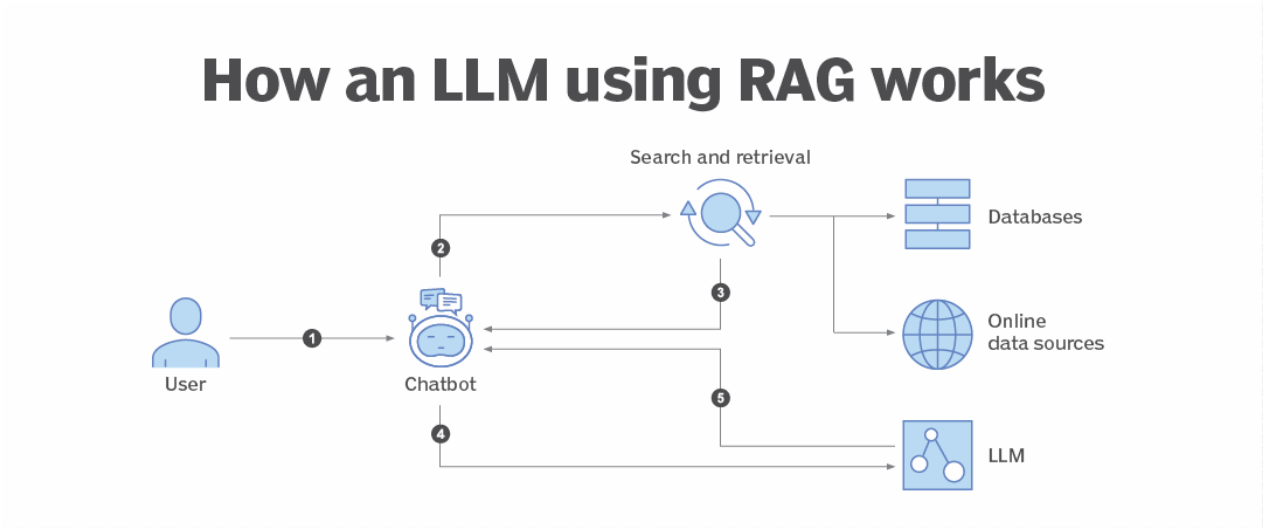
Email: ann.ceit.aut@gmail.com CC: m.ebadpour@aut.ac.ir

توجه: می‌توانید از منابع و بسترهای سخت افزاری برخط رایگان نظیر Google Colab یا Kaggle استفاده نمایید.

تاخیر مجاز: در طول ترم، ده روز زمان مجاز تاخیر برای ارسال پروژه‌ها در اختیار دارید(بدون کسر نمره). این تاخیر را می‌توانید بر حسب نیاز بین پروژه‌های مختلف تقسیم کنید که مجموع آن نباید بیشتر از ده روز شود. پس از استفاده از این تاخیر مجاز، هر روز تاخیر باعث کسر ۱۰٪ نمره‌ی کسب شده‌ی آن تمرین خواهد شد.

ارسال: فایل های کد و گزارش خود را در قالب یک فایل فشرده با فرمت StudentID_HW03.zip تا تاریخ ۱۴۰۲/۰۲/۰۶ صرفا از طریق سایت کورسز ارسال نمایید. ارسال از طریق تلگرام، ایمیل و سایر راه‌های ارتباطی مجاز نبوده و تصحیح صورت نخواهد گرفت.

برای آموزش مدل‌های زبانی بزرگ[1] که حاوی میلیون‌ها و میلیاردها پارامتر هستند از حجم قابل توجهی داده استفاده می‌شود. اما در تمامی این مدل‌ها یک تاریخ قطع آموزش وجود دارد که مدل زبانی هیچ اطلاعاتی در خصوص داده‌های تولید شده‌ی پس از این زمان ندارد. به عنوان مثال تاریخ قطع آموزش مدل GPT-۳.۵-turbo-instruct سپتامبر ۲۰۲۱ است و از همین رو این مدل ممکن است به سوالات مربوط به رویدادهای سال‌های ۲۰۲۲، ۲۰۲۳ و ۲۰۲۴ پاسخ صحیحی ندهد. چنین داده‌هایی که بعد از تاریخ قطع آموزش تولید شده‌اند و یا بخشی از داده‌ی آموزشی اولیه‌ی مدل زبانی نیستند را داده‌ی خارجی می‌گوییم. تکنیک **تولید تقویت شده با بازیابی**[2](RAG) رویکردی است که با استخراج داده‌ی خارجی متناسب با فرمان[3] دریافت شده و افزودن آن به عنوان ورودی به مدل زبانی تلاش می‌کند که فرمان ورودی را تقویت کرده و به مدل زبانی کمک کند تا جواب مرتبط و متناسبی بسازد. به عنوان مثال در پاسخ به یک فرمان متنی مانند «چه کسی شرکت توییتر را در سال ۲۰۲۲ خرید؟» تمامی داده‌های خارجی متناسب با این فرمان را استخراج می‌کند و آنها را به عنوان ورودی به مدل زبانی GPT-۳.۵-turbo-instruc می‌دهد تا مدل زبانی بتواند با دانش دریافت شده پاسخ متناسبی تولید کند. این رویکرد نیاز به آموزش مجدد و یا باز-تنظیم[4] مدل زبانی را برطرف می‌سازد. در این پروژه می‌خواهیم با استفاده از **شبکه‌های خودسازمانده** این تکنیک را پیاده سازی کنیم.



شکل ۱: فرآیند کلی RAG در یک مدل زبانی بزرگ

وظیفه‌ی اصلی RAG جستجوی معنایی[5] در پایگاه داده‌های اطلاعاتی و بازیابی اطلاعات خارجی دارای تناسب محتوایی با فرمان داده شده به یک مدل زبانی است. برای تسهیل جستجوی معنایی، ابتدا داده‌های خارجی استخراج شده به بازنمایی‌های عددی یا بردار تبدیل می‌شوند که به این بازنمایی، تعبیه‌ی متن[6] می‌گوییم. در زمان بازیابی نیز ابتدا فرمان متنی به بازنمایی برداری تبدیل

[1] Large Language Models
[2] Retrieval-Augmented Generation
[3] Prompt
[4] Fine-Tuning
[5] Semantic Search
[6] Text Embedding

می‌شود و سپس نزدیکترین بردارهای داده‌ی خارجی متناسب با آن استخراج می‌شود. شکل ۱ دیاگرام کلی این فرآیند را نشان می‌دهد. چالش اصلی این رویکرد این است که جستجوی معنایی ذکر شده به دلیل نیازمندی به محاسبه‌ی فاصله‌ی بردار فرمان با حجم عظیمی از بردارهای داده‌ی خارجی، به منابع پردازشی و محاسباتی زیاد و زمان قابل توجهی نیاز دارد. بنابراین پیدا کردن رویکردی که جستجوی معنایی را به صورت کارا **انجام** دهد بسیار حائز اهمیت است.

برای افزایش کارایی جستجوی معنایی، یک رویکرد رایج این است که بردارهای داده‌های خارجی را خوشه‌بندی کنیم و در زمان جستجو نیز ابتدا خوشه‌ی مشابه با بردار فرمان ورودی را پیدا می‌کنیم و سپس شباهت بردارهای داده‌های خارجی متعلق به آن خوشه با بردار فرمان را محاسبه می‌کنیم و اگر شباهت بردارها از یک آستانه بیشتر باشد آنها را به عنوان اطلاعات مرتبط در نظر می‌گیریم.

1. در این پروژه قصد داریم برای خوشه بندی داده‌های خارجی از شبکه‌های خود سازمانده استفاده کنیم. بررسی کنید که این شبکه‌ها نسبت به سایر روش‌های خوشه بندی که در یادگیری ماشین به کار گرفته می‌شوند چه مزایا و معایبی دارند؟ به نظر شما چرا استفاده از شبکه‌های خود سازمانده در حل این چالش رویکرد مناسبی است؟(۱۰ امتیاز)
2. با توجه به اینکه یادگیری در شبکه‌های خودسازمانده به صورت با نظارت صورت نمی‌گیرد، فرآیند یادگیری این مدل‌ها را توضیح دهید.(۱۰ امتیاز)
3. مجموعه داده‌ی ارائه شده در این پروژه شامل رویدادهای سه سال متوالی از ۲۰۲۲ تا ۲۰۲۴ است که از سایت ویکی پدیا جمع آوری شده است. داده‌ی مربوطه را بارگذاری کنید و پیش پردازش‌های متنی شامل حذف کلمات ایست[7]، واحدسازی[8] کلمات و تبدیل به بردارهای Glove را روی آن **انجام** دهید.(۱۰ امتیاز)
4. پارامترهای ورودی مدل minisom را توضیح دهید. پارامترهای شبکه‌ی خودسازمانده را تنظیم کنید و شبکه را بر روی داده‌های مربوطه آموزش دهید.(مقادیر تمامی پارامترها را در گزارش خود اضافه کنید.)سپس به ازای هر داده‌ی ورودی واحد منطبق[9] با آن را به دست آوردید و به عنوان نمایه‌ی داده‌ی مربوطه ذخیره کنید.(۲۰ امتیاز)
5. برای ۵۰ رویداد که به صورت تصادفی از مجموعه داده انتخاب شده اند نقشه‌ی خروجی را رسم کنید. نقشه‌ی به دست آمده را تفسیر کنید.(۱۰ امتیاز)
6. فرآیند جستجو را به صورت زیر برای سه رویداد دلخواه از سه سال گذشته **انجام** دهید.(می‌توانید از پرسش های موجود در فایل sample_questions.txt کمک بگیرید.) و خروجی مربوطه را در گزارش خود اضافه کنید.(۲۵ امتیاز)
   - تبدیل پرسش به بردار.
   - پیدا کردن نمایه‌ی متناسب با پرسش مربوطه.

[7] Stop-words
[8] Tokenization
[9] Best Matching Unit

- پیدا کردن تمامی داده‌های خارجی نمایه‌ی مورد نظر.
- محاسبه‌ی معیار شباهت کسینوسی و خروجی دادن بردارهای داده‌های خارجی با شباهت بیشتر از آستانه. (چرا معیار کسینوسی در این مساله انتخاب مناسبی است؟)