به نام خدا

پروژه فهم زبان

دانشگاه صنعتی امیرکبیر

در این پروژه هدف تولید خودکار شعر که یکی از کاربردهای تولید خودکار متن (تولید زبان طبیعی[1] در فارسی است، می‌باشد. تولید متن زیرمجموعه‌ای از پردازش زبان طبیعی (NLP) بوده که در آن ماشین سعی در تقلید رفتار انسانی در نوشتن دارد. در این پروژه ۲ هدف اصلی دنبال می‌شود و در پایان این پروژه شما با مدل زبانی GPT2 و همچنین T5 خواهید توانست که متن (در اینجا شعر) تولید کنید. برای این منظور از مجموعه داده اشعار حافظ استفاده خواهد شد که در مجموعه فایل‌ها در پیوست پروژه قرار دارد. این مجموعه داده شامل ۲ دسته‌ی آموزشی و ارزیابی است. شما در روند آموزش مدل‌ها و تنظیم دقیق تنها حق استفاده از داده آموزشی را داشته و از داده ارزیابی تنها برای بررسی نتایج استفاده خواهید کرد. از آنجایی که اشعار حافظ به تنهایی برای آموزش کافی نیست، اشعار دیگر شاعران نیز در اختیارتان قرار داده شده است که از آنها می‌توانید استفاده کنید، اما برای بدست آوردن معیارهای ارزیابی فقط از دادگان حافظ استفاده کنید.

۱- استفاده از مدل زبانی GPT2

در این بخش هدف تنظیم دقیق مدل GPT2 به منظور تولید شعر با شرایط زیر است:

- در ابتدا مدل را به نحوی تنظیم کنید که یک مصراع دریافت و مصراع بعدی را تولید کند.
- در مرحله بعد این کار را برای یک بیت تکرار کنید. در واقع در این حالت مدل با دیدن یک بیت، باید بیت بعدی را تولید کند. در دنباله ورودی و خروجی باید یک توکن برای مرز بین مصراع‌ها در بیت استفاده شود.

در مرحله‌ی تست، برای هر دو بخش (۵۰۰ بیت و یا ۱۰۰۰ مصرع) نمونه تولید کرده و معیار BLUE را محاسبه کنید. بدین‌منظور باید یک مصرع و یا یک بیت از دادگان تست به مدل داده و خروجی تولید شده مدل را بدست آوردید. سپس با مقایسه بین خروجی مدل و مصرع دوم واقعی معیار را حساب کنید. یک مرتبه هم نزدیک‌ترین مصراع از دادگان آموزشی حافظ به مصراع تولید شده را پیدا کرده و سپس معیار را با آن محاسبه کنید. اینجوری می‌توان قدرت حفظ و نوآوری مدل را با هم مقایسه کرد. برای مدل پایه، می‌توانید از مدل‌های چندزبانه بر پایه GPT استفاده کنید و یا اینکه از مدلی که برای فارسی پیش آموزش شده است استفاده کنید. دقت کنید که یک مدل برای تولید شعر از این وجود دارد که حق استفاده از آن را ندارید!

۲- استفاده از مدل M5

مانند بخش قبل باید یکی از مدل‌های برمبنای T5 را برای تولید شعر آموزش دهید. می‌توانید از مدل چندزبانه آن (mT5) استفاده کنید و یا اینکه مدل‌هایی که برای زبان فارسی آموزش داده شده‌اند. فرآیند آموزش و ارزیابی دقیقاً مثل بخش قبل خواهد بود.

---

[1] Natural Language Generation (NLG)

نمره اضافه:

در داده‌های آموزشی قالب شعر نیز وجود دارد. در صورتی که بتوانید مدل‌ها را با دیتای تمام شاعرها به نحوی آموزش دهید که بتواند با قالب‌های مختلف شعر تولید کند تا ۲۰ درصد نمره اضافه می‌توانید بسته به کیفیت کار بگیرید.

برای پروژه باید یک گزارش مناسب و استاندارد بنویسید. در تمامی بخش‌ها تحلیل‌ها تا حد امکان کافی و مناسب باشد. برای هر بخش علاوه بر معیارهایی عددی، از نمونه‌های تولید شده باید حداقل ۵ نمونه در گزارش قرار گیرد و روی آن بحث شود. مثلاً اینکه مدل کجا خوب بوده و کجا بد و ...