فاز اول | پردازش زبان طبیعی

چالش یادگیری ماشین

Machine Learning Challenge

برگزارکننده:

دستیاران آموزشی درس هوش مصنوعی دانشگاه صنعتی شریف

بهار ۱۴۰۲

مقدمه

در چالش پردازش زبان طبیعی، هدف اصلی طراحی و آموزش یک مدل دسته‌بند است که قادر باشد موضوع متن موجود در یک صفحه وب را تشخیص دهد. برای این منظور، مجموعه‌ای از متن صفحات وب به همراه برچسب موضوعی آنها در اختیار شما قرار خواهد گرفت و از شما خواسته خواهد شد تا مدلی آموزش دهید که برای متن صفحات بدون برچسب، برچسب موضوع را به صورت عددی تشخیص دهد.

توضیحاتی درباره مدل‌های زبانی

از جمله مدل‌هایی که می‌توانید از آن‌ها برای آموزش دسته‌بند استفاده کنید، مدل‌های زبانی هستند. این مدل‌ها شبکه‌های عصبی هستند که سعی دارند توزیع احتمالی زبان طبیعی را یاد بگیرند یا به عبارتی پیش‌بینی می‌کنند کلمه بعدی در یک جمله یا متن، بر اساس کلمات قبلی چه خواهد بود. از جمله مهم‌ترین انواع آن‌ها، مدل‌های زبانی بر پایه تبدیل‌گر[1] است که برخلاف مدل‌هایی مانند LSTM و GRU که از معماری توالی برای پردازش متن استفاده می‌کنند، از مکانیزم توجه (Attention Mechanism) برای درک بهتر ساختار و معنی متن استفاده می‌کنند. از مهم‌ترین‌های این نوع مدل‌زبانی، می‌توان به BERT اشاره کرد که قادر است از دو جهت (چپ به راست و راست به چپ) متن را پردازش کند و درک زبان طبیعی را بهبود بخشد.

فاز اول

در فاز اول، متن صفحات به همراه برچسب هر یک که می‌تواند یکی از ۱۶ موضوع موجود در مجموعه داده باشد در قالب یک دیتافریم پایتون (مانند نمونه ی زیر) در اختیار شما قرار می‌گیرد و شما می‌بایست با استفاده از آن‌ها، مدل دسته‌بند خود را آموزش دهید. به مجموعه داده یادگیری می‌توانید از این لینک دسترسی داشته باشید. سپس تا پایان زمان برگزاری این فاز، مهلت دارید مدل را برای ارزیابی در سامانه بارگذاری کنید. پس از اتمام این فاز، در مرحه ارزیابی مدل، دستیاران از معیار Weighted F1 Score برای بررسی عملکرد مدل شما استفاده می‌کنند؛ به این صورت که با استفاده از داده تستی که در اختیار شما قرار داده نشده است، مدل شما را اجرا کرده و با استفاده از برچسب‌هایی که مدل شما برای هر نمونه از داده تست ایجاد می‌کند، مقدار معیار Weighted F1 Score محاسبه می‌شود. سپس گروه‌های شرکت کننده در این فاز بر اساس مقدار این معیار به صورت نزولی مرتب می‌شوند و ۸ گروه اول به فاز دوم چالش دعوت خواهند شد.

---

[1] Transformer

اگر شما به نتیجه Weighted F1 Score = 0.6 برسید، نمره درسی این بخش را دریافت می‌کنید.

text	class_id	class_name
...	...	...
تداوم فاز انتظاری در بازار خودرو/دنا پلاس امروز چند قیمت خورد؟+ جدول رصد بازار خودروهای داخلی ...	21	خودرو
...	...	...

نحوه ارسال

در وهله‌ی اول ما از شما می‌خواهیم تا نوت‌بوکی که به شما تحویل داده شده است را تکمیل کنید. تمام مراحل آموزش مدل و اجرای کد روی دادگان باید در این نوت‌بوک آورده شده باشد.

سپس در کنار نوت‌بوک، یک فایل به نام `test.py` (به صورت کد زیر) حاوی کلاس `ClassificationModel` باید ارسال شود:

test.py:

```
class ClassificationModel():
    def __init__(self):
        # code for model initialization

    # test_dataframe consists of a column ‘text’ with N rows
    # each row contains a string.
    # returns an N list
    def classify_text(self, test_dataframe):
        # computation of the model’s output
        return labels
```

این کلاس شامل تابع `classify_text` (کپی این تابع از نوت‌بوک) است. دستیاران از این تابع برای بررسی عملکرد مدل شما به صورت زیر استفاده خواهند کرد:

```
from test import ClassificationModel

classifier = ClassificationModel()
results = classifier.classify_text(test_dataframe)
```

بنابراین اگر نیاز به بارگذاری مدل و کارهای دیگر است، باید در این کلاس و در این فایل به صورت کامل لحاظ شود تا با اجرای تابع `classify_text(test_dataframe)` از کلاس، خروجی `results` به صورت آرایه ای از اعداد که نشان دهنده برچسب پیش‌بینی شده هر متن ورودی به ترتیب ظهور آن‌ها در `test_dataframe` است، بازگردانده شود.

در نهایت فایل `requirements.txt` شامل نام و نسخه تمام کتاب‌خانه‌هایی که در کدتان استفاده می‌کنید را تولید کنید تا در زمان ارزیابی از آن برای تولید virtual environment پایتون استفاده شود.

بنابراین فایل‌های `test.py`، `AI_Course_Challenge_NLP_Phase1.ipynb` و `requirements.txt` را به صورت زیپ‌شده برای ما ارسال کنید.

محدودیت‌های منابع

مدل شما نمی‌تواند بیشتر از ۵ گیگابایت حجم داشته باشد. همچنین حداکثر مقدار حافظه و زمان قابل استفاده برای تولید برچسب برای یک نمونه ورودی با تعداد ۵۱۲ توکن به صورت زیر است:

```
GPU Memory: 4GB
RAM: 4GB
Time: 1s
```

می‌توانید مدل خود را در مکان‌های عمومی مانند Hugginface یا Google Drive آپلود کنید و در کلاس `ClassificationModel` آن‌را بارگذاری کنید.

محدودیت‌های پیاده‌سازی

- کد پیاده‌سازی شما حتما باید به زبان پایتون باشد.
- برای پیاده‌سازی مدل حتما از کتابخانه PyTorch استفاده کنید.
- استفاده از کتابخانه‌های با پیاده‌سازی آماده (مانند transformers در پایتون) مجاز می‌باشد.
- شما مجاز به استفاده از LangChain، OpenAI-API و غیره نیستید.