دانشکده مهندسی کامپیوتر

درس پردازش زبان طبیعی

تمرین دوم

مهلت ارسال تمرین:

۱۴۰۴/۰۱/۲۲

استاد:

دکتر مرضیه داوودآبادی

همیار استاد:

محمدامین عباسی

نکات تمرین

- خلاقیت نمره اضافی دارد
- جوابها از لحاظ شباهت بررسی می شوند و جواب های مشابه مورد قبول نیستند.
- به ازای هر روز تاخیر ۵۰ درصد از نمره تمرین کسر خواهد شد.
- کدهای این تمرین باید در قالب jupyter notebook به همراه یک گزارش با فرمت PDF ارائه شود.
- تمامی فایل های این تمرین در قالب یک فایل فشرده (rar یا zip) با نام گذاری زیر ارسال شود:

StudentNumber_FirstName_LastName_HW2.zip

شما به‌تازگی به‌عنوان مهندس پردازش زبان طبیعی (NLP) در تیم داده‌ی فروشگاه اینترنتی دیجی‌کالا مشغول به کار شده‌اید. هر روز هزاران کاربر نظر خود را درباره‌ی محصولات در وب‌سایت ثبت می‌کنند؛ نظراتی که گاه بسیار مثبت‌اند و نشانه‌ی رضایت، و گاه به‌شدت منفی و بیانگر نارضایتی از کیفیت محصول یا خدمات. مدیریت و تیم پشتیبانی از شما می‌خواهند سیستمی طراحی کنید که بتواند به‌صورت خودکار احساسات موجود در این نظرات را تشخیص دهد؛ تا تیم‌های مختلف بتوانند به سرعت واکنش نشان دهند، کیفیت خدمات را ارزیابی کنند و حتی تصمیمات استراتژیک‌تری درباره‌ی محصولات بگیرند. برای رسیدن به این هدف، وظیفه‌ی شماست که با کمک ابزارهای یادگیری ماشین و کتابخانه‌ی قدرتمند PyTorch، مدلی بسازید که زبان فارسی را درک کند و بتواند احساس هر نظر را به‌درستی تشخیص دهد. در این پروژه، با استفاده از داده‌های واقعی نظرات کاربران دیجی‌کالا، مرحله‌به‌مرحله وارد فرآیند طراحی یک مدل تحلیل احساسات خواهید شد. از پاک‌سازی داده‌ها و ساخت واژگان گرفته تا طراحی شبکه‌های عصبی و آموزش آن‌ها، شما در نقش یک مهندس NLP واقعی قدم به دنیای پردازش زبان فارسی می‌گذارید.

مأموریت شما آغاز شده است...

مراحل انجام پروژه

۱. داده

در این پروژه، از یک مجموعه‌داده واقعی شامل نظرات کاربران فروشگاه دیجی‌کالا استفاده کنید. این دیتاست شامل هزاران نمونه متن فارسی به همراه برچسب احساس (مثبت یا منفی) است.
https://www.kaggle.com/datasets/soheiltehranipour/digikala-comments-persian-sentiment-analysis

۲ .پیش‌پردازش داده‌ها

با استفاده از کتابخانه‌هایی نظیر هضم یا ابزارهای مشابه، گام‌های زیر را روی داده‌ها اعمال کنید:

- نرمال‌سازی متون فارسی
- حذف علائم نگارشی و نویزهای متنی
- توکن‌سازی جملات به کلمات
- حذف کلمات توقف(Stopwords)
- در صورت نیاز، انجام ریشه‌یابی Stemming یا Lemmatization

۳ .آماده‌سازی داده برای مدل

- تعریف واژگان (Vocabulary) از روی توکن‌ها
- نگاشت کلمات به اندیس‌های عددی
- تبدیل متن‌ها به دنباله‌هایی از اعداد
- یکنواخت‌سازی طول دنباله‌ها با Padding
- پیاده‌سازی کلاسDataset سفارشی در PyTorch
- استفاده از DataLoader برای ساخت Batch های آموزشی

۴ .تعریف مدل درPyTorch

یک مدل ساده طبقه‌بندی متن با استفاده از ساختار زیر طراحی کنید:

- لایه Embedding برای تبدیل اندیس‌ها به بردارهای تعبیه
- شبکه بازگشتی LSTM یا GRU برای تحلیل توالی کلمات
- لایه Fully Connected برای پیش‌بینی برچسب متن

۵ .آموزش مدل

- تعریف تابع زیان CrossEntropyLoss
- استفاده از بهینه‌ساز مانند Adam
- پیاده‌سازی حلقه‌ی آموزش شامل:
  - محاسبه خروجی مدل
  - محاسبه زیان
  - انجام Backpropagation و به‌روزرسانی وزن‌ها

۶ .ارزیابی مدل

- تقسیم داده‌ها به دو بخش آموزش و اعتبارسنجی(Train/Validation)
- محاسبه معیارهایی مانند :
  - دقت(Accuracy)
  - F1-Score ،Recall ،Precision
- نمایش پیش‌بینی مدل روی چند نمونه از داده‌های تست

۷ .تحلیل نتایج و پیشنهاد بهبود

- تحلیل عملکرد مدل و گزارش مشکلات احتمالی
- پیشنهاد برای بهبود نتایج
  - بهبود پیش‌پردازش

- تغییر ساختار مدل (مثلاً تعداد لایه‌ها)
- استفاده از embedding های پیش‌آموزش‌دیده مانند FastText

مأموریت شما به‌عنوان مهندس NLP دیجی‌کالا در این مرحله به نقطه پایان خود نزدیک شده است. سیستمی طراحی کرده‌اید که دیگر صرفاً به داده‌های خام و پراکنده اکتفا نمی‌کند؛ بلکه زبان کاربران را می‌فهمد، احساس آن‌ها را تشخیص می‌دهد و می‌تواند به تصمیم‌سازی در سطوح مختلف سازمان کمک کند. مدلی که ساختید، اکنون قادر است در کسری از ثانیه صدها نظر را پردازش کرده و آن‌ها را بر اساس احساس موجود در متن دسته‌بندی کند. این یعنی یک گام بزرگ به‌سوی اتوماسیون هوشمند تحلیل بازخورد کاربران، به‌ویژه در زبان فارسی که چالش‌های خاص خود را دارد. در این مسیر، شما نه‌تنها با مفاهیم یادگیری عمیق، مدل‌سازی و پردازش زبان طبیعی آشنا شدید، بلکه نقش واقعی یک متخصص NLP را تجربه کردید: از مواجهه با داده‌های خام و واقعی تا ساخت مدلی قابل اعتماد و کاربردی. اما این فقط آغاز راه است. آنچه امروز ساختید، می‌تواند پایه‌ای باشد برای پروژه‌های پیشرفته‌تر: از تشخیص موضوع، خلاصه‌سازی خودکار، تحلیل چندزبانه، تا ترکیب با مدل‌های بزرگ زبانی.

آینده، منتظر مهندسانی است که زبان را به داده و داده را به درک تبدیل می‌کنند.

و حالا، یکی از آن‌ها شما هستید.