به نام خدا

دانشگاه تهران

دانشکده مهندسی برق و کامپیوتر

درس شبکه‌های عصبی و یادگیری عمیق

تمرین چهارم

پرسش ۱	نام دستیار طراح	امیرحسین صفدریان
	رایانامه	Safdarian2000@gmail.com
پرسش ۲	نام دستیار طراح	مائده صادقی
	رایانامه	Maisa.sadeqi@gmail.com
	مهلت ارسال پاسخ	۱۴۰۳.۰۹.۲۹

فهرست

قوانین

قبل از پاسخ دادن به پرسش‌ها، موارد زیر را با دقت مطالعه نمایید:

- از پاسخ‌های خود یک گزارش در قالبی که در صفحه‌ی درس در سامانه‌ی Elearn با نام ***REPORTS_TEMPLATE.docx*** قرار داده شده تهیه نمایید.
- پیشنهاد می‌شود تمرین‌ها را در قالب گروه‌های دو نفره انجام دهید. (بیش از دو نفر مجاز نیست و تحویل تک نفره نیز نمره‌ی اضافی ندارد) توجه نمایید الزامی در یکسان ماندن اعضای گروه تا انتهای ترم وجود ندارد. (یعنی، می‌توانید تمرین اول را با شخص A و تمرین دوم را با شخص B و ... انجام دهید)
- **کیفیت گزارش شما در فرآیند تصحیح از اهمیت ویژه‌ای برخوردار است**؛ بنابراین، لطفا تمامی نکات و فرض‌هایی را که در پیاده‌سازی‌ها و محاسبات خود در نظر می‌گیرید در گزارش ذکر کنید.
- در گزارش خود مطابق با آنچه در قالب نمونه قرار داده شده، برای شکل‌ها زیرنویس و برای جدول‌ها بالانویس در نظر بگیرید.
- الزامی به ارائه توضیح جزئیات کد در گزارش نیست، اما باید نتایج بدست آمده از آن را گزارش و تحلیل کنید.
- **تحلیل نتایج الزامی می‌باشد، حتی اگر در صورت پرسش اشاره‌ای به آن نشده باشد.**
- **دستیاران آموزشی ملزم به اجرا کردن کدهای شما نیستند**؛ بنابراین، هرگونه نتیجه و یا تحلیلی که در صورت پرسش از شما خواسته شده را به طور واضح و کامل در گزارش بیاورید. در صورت عدم رعایت این مورد، بدیهی است که از نمره تمرین کسر می‌شود.
- **کدها حتما باید در قالب نوت‌بوک با پسوند .ipynb تهیه شوند، در پایان کار، تمامی کد اجرا شود و خروجی هر سلول حتما در این فایل ارسالی شما ذخیره شده باشد.** بنابراین برای مثال اگر خروجی سلولی یک نمودار است که در گزارش آورده‌اید، این نمودار باید هم در گزارش هم در نوت‌بوک کدها وجود داشته باشد.
- **در صورت مشاهده‌ی تقلب امتیاز تمامی افراد شرکت‌کننده در آن، 100- لحاظ می‌شود.**
- تنها زبان برنامه نویسی مجاز **Python** است.
- **استفاده از کدهای آماده برای تمرین‌ها به هیچ وجه مجاز نیست. در صورتی که دو گروه از یک منبع مشترک استفاده کنند و کدهای مشابه تحویل دهند، تقلب محسوب می‌شود.**

- نحوه محاسبه تاخیر به این شکل است: پس از پایان رسیدن مهلت ارسال گزارش، حداکثر تا یک هفته امکان ارسال با تاخیر وجود دارد، پس از این یک هفته نمره آن تکلیف برای شما صفر خواهد شد.
  - سه روز اول: بدون جریمه
  - روز چهارم: ۵ درصد
  - روز پنجم: ۱۰ درصد
  - روز ششم: ۱۵ درصد
  - روز هفتم: ۲۰ درصد
- حداکثر نمره‌ای که برای هر سوال می‌توان اخد کرد ۱۰۰ بوده و اگر مجموع بارم یک **سوال** بیشتر از ۱۰۰ باشد، در صورت اخد نمره بیشتر از ۱۰۰، اعمال نخواهد شد.
  - برای مثال: اگر نمره اخذ شده از سوال ۱ برابر ۱۰۵ و نمره سوال ۲ برابر ۹۵ باشد، نمره نهایی تمرین ۹۷.۵ خواهد بود و نه ۱۰۰.
- لطفا گزارش، کدها و سایر ضمایم را به در یک پوشه با نام زیر قرار داده و آن را فشرده سازید، سپس در سامانه‌ی Elearn بارگذاری نمایید:

HW[Number] _[Lastname]_[StudentNumber]_[Lastname]_[StudentNumber].zip

(مثال: HW1_Ahmadi_810199101_Bagheri_810199102.zip)

- برای گروه‌های دو نفره، بارگذاری تمرین از جانب یکی از اعضا کافی است ولی پیشنهاد می‌شود هر دو نفر بارگذاری نمایند.

پرسش ۱. تشخیص هرزنامه

هدف از این سوال آشنایی با وظیفه تشخیص هرزنامه بر روی متن فارسی می‌باشد. تشخیص هرزنامه شامل شناسایی و دسته‌بندی پیام‌ها به دو دسته‌ی هرزنامه و غیر هرزنامه است. برای آشنایی با روند سوال، این مقاله را مطالعه کنید.

۱-۱. مجموعه‌داده

(۱۰ نمره)

برای شروع، این مجموعه داده را از Kaggle دریافت کنید. توجه داشته باشید که این مجموعه داده با داده مورد استفاده در مقاله متفاوت است. در صورت عدم دسترسی، میتوانید از فایل فشرده این مجموعه داده که پیوست شده است استفاده کنید. کلاسهای موجود در ستون labelو تعداد نمونه‌های هر کلاس را به کمک یک نمودار میله‌ای نمایش دهید.

۱-۲. پیش‌پردازش داده‌ها

(۲۰ نمره)

پیش‌پردازش متن در پردازش زبان طبیعی **(NLP)** یک مرحله‌ی رایج برای آماده‌سازی داده‌ها است که معمولاً برای هماهنگ‌سازی فرمت داده‌ها و حذف نویزها انجام می‌شود. در این تمرین، شما باید مراحل زیر را به‌منظور بررسی تأثیر پیش‌پردازش بر عملکرد مدل اجرا کنید:

مراحل پیش‌پردازش:

برای پیش‌پردازش داده‌ها، مراحل زیر باید اعمال شود:

- حذف**URL** ها (لینک‌ها) :حذف لینک‌های موجود در متن به دلیل اینکه اغلب ارتباط مستقیمی با موضوع تحلیل ندارند.
- حذف آدرس‌های ایمیل :حذف ایمیل‌ها از متن به دلیل اینکه معمولاً اطلاعات معنایی خاصی برای مدل فراهم نمی‌کنند.
- حذف شماره‌های تلفن :حذف اعداد طولانی یا شماره‌های تماس.
- حذف تکرار حروف :کاهش تکرار حروف به یک حرف ساده (مثلاً "عاااالی" به "عالی").

- حذف کلمات توقف: حذف کلماتی مانند "و"، "به"، "از" که به‌تنهایی اطلاعات معنایی خاصی ندارند.

۱-۳. نمایش ویژگی

(۲۰ نمره)

برای این بخش:

۱. داده‌های متنی پیش‌پردازش‌شده را با استفاده از توکن‌ساز **ParsBERT** به اعداد تبدیل کنید.

۲. **Padding**: تمام سطرها باید طول یکسانی داشته باشند. طول جملات را برابر با ۳۲ در نظر بگیرید.

۳. بردار تعبیه: با استفاده از مدل از پیش آموزش‌دیده **ParsBERT**، بردار تعبیه (**Embedding**) را برای ورودی‌ها به‌دست آورید.

۴. ابعاد بردار تعبیه را به ۱۲۰ کاهش دهید.

سؤال:

- ابعاد پیش‌فرض بردار تعبیه در **ParsBERT** چقدر است؟
- تعداد ابعاد این بردار بیانگر چیست؟
- مفهوم بردار تعبیه را توضیح دهید و بیان کنید کدام کلمات موجود در مجموعه داده ممکن است تعبیه‌ای نزدیک به هم داشته باشند؟

۱-۴. ساخت مدل

(۳۵ نمره)

وظایف:

۱. داده‌ها را با نسبت ۷۰-۳۰ به دو دسته‌ی آموزش و تست تقسیم کنید. از ۲۰٪ داده‌های آموزش به‌عنوان مجموعه اعتبارسنجی استفاده کنید.

الگوریتم جستجوی حریصانه برای یافتن بهترین ترکیب هایپرپارامترها برای مدل **CNN-LSTM** در فضای زیر اعمال کنید:

- **batch_sizes = [8, 64]**
- **learning_rates = [0.001, 0.0001]**

- **optimizers = [Adam, SGD]**

۲. با استفاده از بهترین هایپرپارامترها، مدل CNN-LSTM را بسازید و آموزش دهید.

۳. مدل‌های ساده‌ی CNN و LSTM را نیز با هایپرپارامترهای بهینه ایجاد و آموزش دهید.

سؤال:

- نقاط قوت و ضعف هر یک از مدل‌ها CNN و LSTM چیست؟
- ادغام این دو مدل با چه هدفی انجام می‌شود؟

۱-۵. ارزیابی

(۱۵ نمره)

دادههای تست را به کمک معیارهای ارزیابی ذکر شده در مقاله ارزیابی کنید و یک جدول مشابه جدول ۳ مقاله برای مدل‌های LSTM،CNN و CNN-LSTM چاپ کنید.

۱-۵. امتیازی

(۵ نمره)

از روش کیسه کلمات (Bag of Words) برای نمایش ویژگی استفاده کنید.

از بین مدل‌های سنتی یادگیری ماشین که مقاله در جدول ۳ به آن‌ها اشاره کرده ۴ مورد را به انتخاب خود با استفاده از کتابخانه‌ی sklearn آموزش دهید و روی داده‌های تست ارزیابی کرده و نتایج را به جدول بخش قبل اضافه کنید.

پرسش ۲ – پیش‌بینی ارزش نفت

۲-۱. مقدمه

از رایج ترین کابرد شبکه های حافظه دار می‌توان به پیش بینی سری های زمانی اشاره کرد. در این سوال با نحوه ی پیش‌بینی ارزش نفت خام[1] با استفاده از چهار روش متفاوت آشنا خواهید شد.

۲-۲. مجموعه دادگان و آماده‌سازی

مجموعه داده ی مورد استفاده در این سوال،CL=F، را از سال ۲۰۱۰ تا کنون را از Yahoo Finance دانلود کنید. از بین ویژگی های داده شده ستون Adj Close به عنوان ویژگی اصلی مد نظر قرار دهید. در برخی روزها داده ای ثبت نشده است که به عنوان داده ی null تلقی میشوند.

- علاوه بر داده های null موجود، ده درصد داده های ثبت شده را به صورت رندم حذف کنید.
- سپس، روش هایی برای جایگزینی داده های ناموجود ارائه دهید و داده ها را تکمیل کنید. **(۱۰ نمره)**
- طبق نسبت موجود در مقاله داده هارا به دو دسته ی آموزشی و آزمایشی تقسیم کرده و نرمال کنید. **(۵ نمره)**
- مشابه شکل ۶ داخل مقاله، هیستوگرام توزیع قیمت را نمایش دهید.**(۵ نمره)**

[1] https://www.ijournalse.org/index.php/ESJ/article/view/2149

۲-۳. پیاده‌سازی مدل ها

در مقاله ی داده شده پیش‌بینی سری زمانی توسط سه مدل LSTM، Bi-LSTM و GRU انجام شده‌است. ضمن در نظر گرفتن میانگین مربعات خطا[1] به عنوان تابع خطا، طبق هایپرپارامتر های جدول ۴ موجود در مقاله این مدل هارا آموزش دهید و موارد خواسته شده را گزارش کنید.(۳۰ نمره)

Table 4. Hyperparameters of LSTM, GRU, and Bi-LSTM modelling

Learning Rate	0.0010
Batch Size	100
Optimizer	Adam
Epochs	50
Units	512 (LSTM & GRU); 1024 (Bi-LSTM)

- برای هر سه مدل داده شده، نتایج پیش‌بینی شده را همراه مقادیر واقعی نمایش دهید. (۱۵ نمره)
- ابتدا به طور مختصر در مورد معیار های داخل مقاله، MAE، RMSE، R-Squared و MAPE توضیح دهید. سپس مقادیر را گزارش کرده و نتایج را تحلیل و مقایسه کنید. (۱۵ نمره)

۲-۴. ARIMA

(۲۵ نمره)

در این قسمت از سوال با مدل کلاسیک ARIMA [2] و SARIMA [3] آشنا خواهید شد. در ابتدا تفاوت این دو مدل را بیان کنید.

- مزایا و محدودیت های مدل ARIMA را مختصر ذکر کنید.

[1] Mean Square Error
[2] Autoregressive Integrated Moving Average
[3] Seasonal ARIMA

- مدل ARIMA پارامتر هایی دارد، مفهوم ریاضی این مدل را با ذکر پارامتر ها شرح دهید.
- پارامتر های بهینه ی این مدل را بدست آورده و گزارش کنید.
- ضمن ارائه ی جدولی مشابه جدول شماره ۶، نتایج را با نتایج داخل مقاله مقایسه کنید.