در این پروژه قرار است با استفاده از شبکه‌های عصبی به دسته‌بندی ۶ نوع ویروس مختلف بپردازید. ژنوم هر ویروس توسط دنباله‌ای از نوکلئوتیدها نشان داده می‌شود که طول‌های متفاوتی دارند. نام اصلی هر ویروس در این پروژه مهم نیست و به همین دلیل نام‌ها با Class1 تا Class6 نشان داده شده‌اند. دادگان آموزشی شامل ۱۳۲۰ نمونه است که به صورت مساوی از هر نوع ویروس ۲۲۰ نمونه داریم. این داده‌ها در یک فایل متنی csv ذخیره شده است.

علاوه‌بر داده‌های آموزش، یک داده محدودی شامل ۳۰ نمونه از هر کلاس به عنوان داده توسعه (development) به شما داده شده است. با استفاده از این دادگان شما باید دقت مدل خودتان را بدست آورید و چک کنید که درست کار می‌کند یا نه. فرمت این دادگان مثل دادگان آموزش است. دقت کنید که استفاده از داده توسعه برای آموزش به هر نحوی (مثلاً برای داده اعتبارسنجی در آموزش مدل و یا اضافه کردن آن به داده آموزش) مجاز نیست و صرفاً باید از آن برای ارزیابی استفاده کنید.

برای داده‌های ارزیابی (تست) یک فایل مجزا شامل ۴۰۰ تست بدون برچسب به شما داده خواهد شد. شما باید این فایل را خط به خط پردازش و کلاس معادل آنها را بدست آورید. برای سادگی کار این کلاس‌ها را در یک لیست از اعداد ۱ تا ۶ قرار دهید. از آنجایی که در کوئرا باید یک فایل پایتون آپلود کنید، این فایل باید شامل این لیست از اندیس‌ها باشد. ورودی هر تست‌کیس به صورت ساده یک اندیس از این لیست است. شما باید با توجه به آن اندیس یکی از خانه‌های لیست را انتخاب و کلاس معادل آن را در کنسول چاپ کنید. بدین ترتیب به راحتی خروجی این کد با برچسب‌های واقعی قابل مقایسه است. بعد از آپلود این کد پایتون کوچک خروجی را به صورت درصد خواهید دید. برای اینکه مشکلی در فهم این قسمت نداشته باشید نمونه کد آن به شما داده شده است.

قوانین انجام پروژه:

۱. انتخاب نوع شبکه عصبی از بین MLP، LSTM و CNN به عهده خود شماست. از هر روشی استفاده کنید قابل قبول است. فقط دقت خروجی برای پروژه مهم است.

۲. از آنجایی که دنباله هر نمونه طول‌های متفاوتی دارد و طول آنها هم زیاد است باید از روشی برای تبدیل آن به طول ثابت و یا کاهش طول آن استفاده کنید. اینکه از چه روشی استفاده کنید نیز به خود شما وابسته است. جهت پیشنهاد می‌توانید از روش ساده k-mer که در مقاله زیر توضیح داده شده و خیلی شبیه به مطالبی است که در درس خوانده‌اید استفاده کنید.

https://academic.oup.com/gigascience/article/7/12/giy125/5140149?login=false

۳. دقت کنید که تعداد نمونه‌ها زیاد نیست. به همین خاطر شبکه مورد استفاده خیلی زود می‌تواند به داده‌های آموزشی بیش‌برازش شود. هم‌چنین اگر از روش k-mer استفاده کنید، به راحتی می‌توانید شبکه را روی CPU هم آموزش دهید. اگر هم خواستید از شبکه بزرگتری استفاده کنید باید از Colab استفاده کنید.

۴. برای انجام پروژه حتماً باید از پایتون استفاده کنید و تمام کدهای شما باید در یک فایل باشد که به راحتی قابل آپلود در کوئرا باشد. دقت کنید که قرار نیست کد شما روی کوئرا اجرا شود و آنجا فقط شباهت و ... کدها چک می‌شود و در نهایت کدها

نیز به صورت دستی بررسی خواهند شد. خروجی را خودتان تولید و آپلود خواهید کرد. در صورت نیاز باید پروژه را به دستیارن تحویل اسکایپی دهید. پس سعی نکنید که روش‌های غیرمجاز استفاده کنید.

۵. برای انجام پروژه از هر جعبه ابزار شبکه عصبی‌ای که خواستید می‌توانید استفاده کنید. پیشنهاد اول PyTorch و پیشنهاد دوم TensorFlow است ولی از دیگر جعبه ابزارها نیز می‌توانید استفاده کنید. فقط الزام به استفاده از پایتون است.

۶. از آنجایی که تعدادی از دانشجویان درخواست داشتند که پروژه نداشته باشیم، پروژه اختیاری است. در صورتی که پروژه را انجام ندهید نمره پروژه روی باقی نمرات پخش خواهد شد، یعنی ابتدا نمره شما از ۱۸ حساب شده و سپس به ۲۰ اسکیل می‌شود. پس شما خودتان باید تصمیم بگیرید که می‌خواهید پروژه را انجام دهید یا نه.

شیوه نمره‌دهی:

نمره پروژه به این صورت است که ده درصد اول بر اساس دقت، نمره مثبت خواهند گرفت. باقی دانشجویان نیز بر اساس میزان دقت و همچنین کیفیت گزارش نمره‌دهی خواهند شد. نمره مثبت بخش کدنویسی و انجام کار حداکثر ۲۰ درصد است و با توجه به فاکتورهای مختلفی داده خواهد شد.

برای پروژه علاوه‌بر کدنویسی باید یک گزارش حداقل ۳ صفحه‌ای نیز ارسال کنید. در گزارش باید مراحل انجام کار، ساختار شبکه مورد استفاده و نتایج بدست آمده بر اساس دادگان توسعه و تست (از کوئرا خواهید دید) را ذکر کنید. در صورتی که شبکه‌های مختلفی انجام داده باشید خوب است که مقایسه بین مدل‌ها را نیز در گزارش بیاورید. نمره گزارش بخشی از نمره پروژه است و اگر کیفیت گزارش خوب نباشد نمره این بخش را از دست خواهید داد.

نکات تحویل پروژه

1) پروژه را به صورت انفرادی انجام دهید.
2) گزارش پروژه را در یک فایل با نام "StudentNumber_FirstnameLastname.pdf" در سایت بارگذاری نمایید.
3) در صورت پیداشدن هرگونه کپی نمره‌ی هر دو نفر 100- درنظر گرفته خواهد شد. این نکته را جدی بگیرید تا بعداً به مشکل نخورید.
4) اشکالات خود را با ایمیل زیر با آقای نفر در میان بگذارید. در صورتی که پاسخی دریافت نکردید با ایمیل استاد آن را مطرح کنید.
   mmnafar57@gmail.com
5) نکته مهم: نمرات درس باید ۱۵ تیر ثبت نهایی شود و این یعنی اینکه باید روز ۱۳ تیر نمرات ثبت موقت شود. از این رو ددلاین پروژه ثابت است و امکان تمدید آن وجود ندارد. **ددلاین پروژه روز ۱۳ تیر ساعت ۸ صبح است**. پس به این زمان دقت کنید.