به نام خدا

تمرین اول درس پردازش زبان طبیعی

«آشنایی با مدل‌های زبانی و کاربردهای آن»

استاد درس: دکتر ممتازی

دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر

زمستان ۱۴۰۰

برای ارسال تمرین به نکات زیر توجه کنید.

۱- در جدول زیر نحوه اعمال نمره منفی برای تاخیر در ارسال تمرین‌ها ذکر شده‌است.

میزان تاخیر (روز)	میزان جریمه
۱ الی ۲ روز	هر روز ۵٪
۲ الی ۶ روز	هر روز ۱۰٪

در صورتی که برای ارسال تمرین‌ها بین ۷ تا ۱۴ روز تاخیر داشته باشید، نمره شما از ۵۰٪ محاسبه می‌شود و پس از این بازه به تمرین ارسالی نمره‌ای تعلق نمی‌گیرد.

۲- هرگونه کپی‌برداری در انجام تمرین‌ها موجب کسر نمره خواهد شد.

۳- آخرین مهلت ارسال تمرین، **ساعت ۲۳:۵۵ روز جمعه ۱۹ فروردین** می‌باشد.

۴- فایل های ارسالی خود شامل فایل‌های پیاده‌سازی و گزارش را فشرده کنید و با عنوان «**شماره دانشجویی_HW1**» مانند `HW1_97131022` ارسال کنید.

۵- زبان برنامه‌نویسی برای انجام تمرین‌ها، پایتون یا جاوا در نظر گرفته شده‌است.

۶- کدهای ارسالی خود را برای افزایش خوانایی و درک بهتر به صورت مناسب کامنت‌گذاری کنید.

۷- برای مدل‌های زبانی آماری استفاده از کتابخانه‌های آماده مجاز نیست. برای بخش ساخت مدل زبانی با استفاده از شبکه‌های عصبی، می‌توانید از کتابخانه‌های تنسورفلو (کراس) و یا پایتورچ استفاده کنید.

۸- در صورت هرگونه سوال یا مشکل می‌توانید با تدریس‌یاران درس از طریق ایمیل زیر در ارتباط باشید.

محمدجواد ظهرابی – رضا زادکمالی nlp.aut.1401@gmail.com

بخش اول: تعریف مسئله و معرفی دادگان

در این تمرین قصد داریم برای مجموعه داده‌ی اشعار فارسی یک مدل زبانی ایجاد کنیم. در این دادگان هر سطر شامل یک مصراع از شعر است. این مجموعه‌داده به قسمت‌های train.txt, test.txt, validation.txt, test_incomplete.txt, test_incomplete_gold.txt تقسیم شده است که هر کدام به ترتیب شامل، دادگان آموزش، دادگان آزمون، مصراع‌های ناقص و حالت کامل شده‌ی این مصراع‌ها هستند. این مجموعه‌های داده را می‌توانید از طریق لینک زیر دانلود کنید:

https://drive.google.com/file/d/16C0_9i0io43VfABV3-uukUjJYlM6k-2U/view?usp=sharing

بخش دوم: مدل‌های زبانی آماری

در این قسمت میخواهیم مدل‌های زبانی یونیگرم[1] و بایگرم[2] را بر روی دادگان آموزش ایجاد کنیم. مدل‌های زبانی ایجادشده باید به روش Absolute Discounting هموارسازی شوند. توجه داشته باشید که مقدار بهینه پارامتر هموارسازی، با استفاده از دادگان اعتبارسنجی[3] محاسبه شود. مقدار perplexity را بر روی دادگان آزمون برای مدل‌های یونیگرم و بایگرم گزارش کنید.

بخش سوم: تکمیل جملات ناقص با استفاده از مدل‌های زبانی آماری

فایل test_incomplete.txt شامل ده مصراع است که تعدادی از کلمات انتهای هر یک از آن‌ها حذف شده است. در جدول۲ مثال‌هایی از این فایل آمده است. تعداد کلمات حذف شده قبل از مصراع نوشته شده و با استفاده از ### جدا شده است. به کمک مدل‌های زبانی یونیگرم و بایگرمی که در بخش دوم ساخته شد، کلمات حذف شده را پیش‌بینی کنید. در خروجی، جمله‌ای که توسط مدل تکمیل شده است را کنار جمله‌ی درست که در فایل test_incomplete_gold.txt قرار گرفته، چاپ کنید. نتایج به دست آمده از عملکرد هر مدل را در فایل گزارش تحلیل کنید.

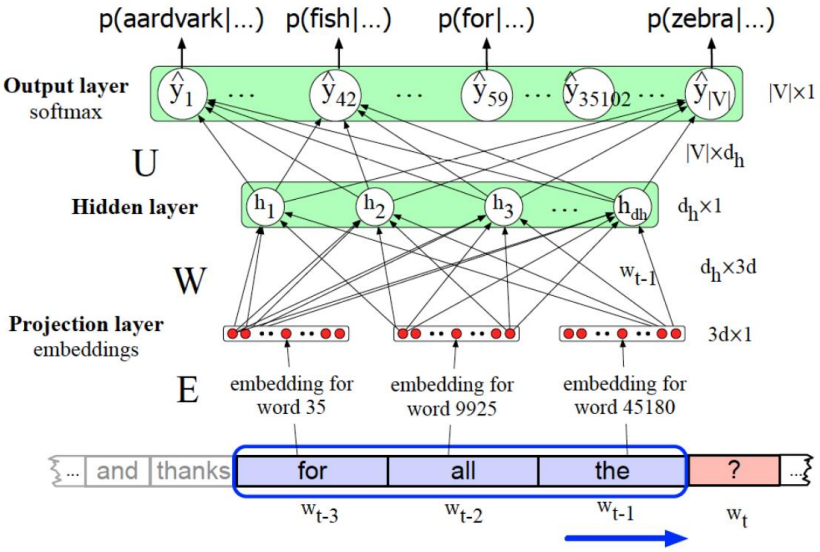
[1] Unigram
[2] Bigram
[3] Validation

جدول۱- نمونه‌هایی از مصراع‌های ناقص و حالت کامل آن‌ها

مصراع ناقص	مصراع کامل
این سخن حقست اگر نزد سخن گستر	این سخن حقست اگر نزد سخن گستر برند
گفت این از خدای	گفت این از خدای باید خواست

بخش چهارم: ایجاد مدل زبانی با استفاده از شبکه عصبی

در اسلایدهای درس با نحوه‌ی ایجاد مدل زبانی از طریق شبکه‌ی جلورو[4] آشنا شدید. در این قسمت می‌خواهیم، یک شبکه عصبی بایگرم و ترایگرم[5] را آموزش دهیم. به این شکل که همانند تصویر زیر، در حالت‌های بایگرم و ترایگرم به ترتیب دو و سه کلمه به عنوان ورودی به شبکه داده شده و کلمه بعدی پیش‌بینی می‌شود. برای جلوگیری از مشکلات مربوط به حافظه، می‌توانید از ۳۰،۰۰۰ سطر اول از دادگان آموزش استفاده کنید.



شکل ۱ شبکه عصبی جلورو

پس از آموزش شبکه از طریق دو روش ذکر شده، مقدار perplexity را بر روی دادگان آزمون برای هر یک مدل‌های ایجاد شده گزارش کنید.

[4] Feedforward
[5] Trigram

بخش چهارم: تکمیل جملات ناقص با استفاده مدل‌های زبانی شبکه عصبی

مراحل توضیح داده‌شده در بخش دوم را برای هر دو مدل زبانی ایجاد شده در بخش قبل، تکرار کنید. نتایج به دست آمده و عملکرد هر مدل را در فایل گزارش تحلیل کنید.

بخش پنجم: تحلیل نتایج

مدل‌های زبانی آماری و شبکه عصبی ایجاد را شده با یکدیگر مقایسه کنید. تحلیل کنید که به نظرتان کدام یک در پیش‌بینی کلمات موفق‌تر عمل کرده است.

موفق باشید