به نام خدا

تمرین اول درس پردازش زبان طبیعی

«آشنایی با مدل‌های زبانی و روش‌های بازنمایی برداری کلمات»

استاد درس: دکتر ممتازی

دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر

زمستان ۱۴۰۲

برای ارسال تمرین به نکات زیر توجه کنید.

۱- برای ارسال پاسخ تمرین‌های این درس، **مجموعا ۱۰ روز** زمان تاخیر مجاز در نظر گرفته شده‌است و در صورتی که مجموع زمان تاخیرها از این مقدار بیشتر باشد، پاسخ ارسال‌شده مورد بررسی قرار نخواهد گرفت.

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

۳- آخرین مهلت ارسال تمرین، **ساعت ۲۳:۵۵ روز شنبه ۱۸ فروردین** می‌باشد.

۴- فایل‌های ارسالی جهت نمره‌دهی باید شامل پیاده‌سازی و گزارش تمرین باشد، فایل‌های خود را فشرده نمایید و به‌صورت «**شماره دانشجویی_HW1**» مانند HW1_400131022 نام‌گذاری کنید.

۵- زبان برنامه‌نویسی برای انجام این تمرین، تنها می‌تواند **پایتون** باشد.

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

۷- در این تمرین شما باید **موارد خواسته‌شده را پیاده‌سازی نمایید** و استفاده از کتابخانه‌های آماده مجاز نمی‌باشد.

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

محمد نعیمی mohammad.naeimi+nlp@aut.ac.ir

بخش اول: تعریف مسئله و معرفی مجموعه‌داده

مجموعه‌داده در این تمرین، شامل ۳ فایل train.csv، val.csv و test.csv که به ترتیب برای آموزش[1]، ارزیابی[2] و آزمون[3] می‌باشد، در اختیار شما قرار گرفته‌است. این مجموعه‌داده مربوط به دسته‌بندی متن می‌باشد که مجموعه‌ای از مقالات خبری فارسی است. هدف ما در این تمرین این است که با استفاده از این مجموعه‌داده مدل‌زبانی احتمالاتی تولید کنیم. همچنین بازنمایی برداری هر کلمه را تولید نموده، برای تعدادی از کلمات، کلمات هم‌معنی را پیدا کنیم. مجموعه‌داده ارائه‌شده شامل دوستون زیر می‌باشد:

ویژگی	توضیحات
content	متن داده
label	برچسب داده

همچنین تعداد محتوای هریک از فایل‌ها به‌شرح زیر می‌باشد:

نام فایل	تعداد ورودی‌ها
train	13,314
val	1,480
test	1,644

بخش دوم: مدل‌های زبانی احتمالاتی (۴۰ امتیاز)

در این قسمت سه مدل‌زبانی احتمالاتی یونیگرم[4]، بایگرم[5] و ترایگرم[6] را بر روی مجموعه‌داده آموزش ایجاد کنید. هریک از مدل‌های زبانی ایجادشده را به صورت مجزا به دو روش Back-off Smoothing و Absolute Discounting هموارسازی[7] نموده، مقدار بهینه پارامترهای هموارسازی را با استفاده از مجموعه‌داده ارزیابی محاسبه کنید. برای بررسی عملکرد مدل‌های زبانی ایجادشده معیار ارزیابی perplexity را بر روی مجموعه‌داده آزمون محاسبه کنید.

الف) معیار ارزیابی را برای تمامی متن‌های مجموعه‌داده آزمون گزارش کنید.

ب) معیار ارزیابی را به صورت مجزا برای هر برچسب مجموعه‌داده آزمون گزارش کنید.

ج) نتایج مدل‌های مختلف را با یکدیگر مقایسه نموده، تاثیر هموارسازی‌های متفاوت را بررسی و تحلیل کنید.

---

[1] Train
[2] Validation
[3] Test
[4] Unigram
[5] Bigram
[6] Trigram
[7] Smoothing

بخش سوم: روش‌های بازنمایی برداری کلمات (۴۰ امتیاز)

می‌خواهیم با استفاده از مجموعه‌داده ارائه‌شده، بازنمایی برداری متن‌ها را تولید نموده و آن‌ها را دسته‌بندی کنیم. بعد از تولید بردار متناظر با هر متن، یک مدل دسته‌بندی را با استفاده از مجموعه‌داده آموزش، آموزش داده و با استفاده از مجموعه‌داده آزمون، ارزیابی کنید. می‌توانید از هر روش دلخواه (مانند KNN یا شبکه‌های عصبی) برای دسته‌بندی متن‌ها استفاده کنید. توجه کنید که خروجی بردار متن حداقل ۲۰۰ در نظر گرفته‌شود. برای تولید بردارهای متن‌ها، هر یک از روش‌های زیر را به صورت جداگانه استفاده کنید، نتایج معیارهای ارزیابی دسته‌بندی را گزارش نموده، مقایسه و تحلیل نمایید. (برای استفاده از مدل word2vec، کتابخانه genism پیشنهاد می‌شود.)

ب) میانگین حسابی بردارهای word2vec (در حالت skip-gram) کلمات متن.

ج) میانگین وزنی بردارهای word2vec (در حالت skip-gram) کلمات متن، با مقدار TF-IDF هر کلمه به عنوان وزن.

بخش چهارم: استفاده از روش‌های ارزیابی و تحلیل نتایج (۲۰ امتیاز)

الف) همانطور که توضیح داده‌شد، برای بررسی عملکرد مدل‌های زبانی ایجادشده، معیار ارزیابی perplexity را بر روی مجموعه‌داده آزمون محاسبه می‌کنیم. این معیار ارزیابی را پیاده‌سازی نمایید. توضیح دهید perplexity چیست و چگونه برای ارزیابی مدل‌های زبانی استفاده و تفسیر می‌شود؟

ب) معیارهای ارزیابی Accuracy و F1-score را برای دسته‌بندی چندکلاسه پیاده‌سازی و استفاده نمایید.

موفق باشید

محمد نعیمی