به نام آنکه آموخت انسان را آنچه نمی دانست

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

درس پردازش زبان‌های طبیعی

تمرین ۱

اسفند ۱۳۹۹

مقدمه

جهت حل مسائل و شبیه‌سازی، به مدل‌سازی‌های اولیه نیاز می‌باشد. جهت شبیه‌سازی زبان انسان در زمینه هوش‌مصنوعی (پردازش متن و صدا) نیز می‌توان از مدل‌های زبانی که پس‌زمینه ریاضیاتی دارند استفاده کرد. از جمله مدل‌های پایه‌ای در این زمینه که پس‌زمینه آماری و احتمالی دارد می‌توان به مدل زبانی ان‌گرام[۱] اشاره کرد.

اهداف این تمرین عبارتند از:

- آشنایی با دادگان[۲] متنی بر پایه شبکه‌اجتماعی توئیتر[۳]
- پیش پردازش دادگان متنی
- کاربرد ان‌گرام‌ها
- کاربرد معیارهای نهادین[۴] (سرگشتگی[۵]) و بیرونی[۶] (تشخیص تنفر/توهین آمیز بودن پیام)
- کاربرد معیارهای ارزیابی در سطح micro و macro

هنگام انجام تمرین لطفا به نکات زیر توجه کنید:

- در تمامی تمرین‌ها، نمره اصلی به تحلیل نتایج با توجه خروجی‌ها تعلق می‌گیرد. (تحلیل اجباری است.)
- از تعریف مجدد تمامی معیارها و مدل‌های بیان شده ضمن در کلاس، ضمن تحلیل نتایج پرهیز کنید.
- استفاده از نمودارها و کشف پیام‌های مرتبط از دادگان در صورتی که موجب افزایش کیفیت تفسیرها گردد، تاثیر مثبت در نمره شما خواهد داشت. به علاوه، در صورتی که بتوانید با توجه به مشاهدات و تحلیل‌هایتان، نتایج ان‌گرام‌ها را بهبود دهید، نمره اضافی دریافت خواهید کرد. دقت کنید، ایده‌هایی قابل قبول هستند که شهود تجربی یا تئوری داشته باشند.
- بدیهی است که حجم گزارش معیار نمره‌ی شما نیست، به تفسیرهایی که بدون آزمایش و صرفا به صورت فرضی بیان گردند نمره‌ای تعلق نمی‌گیرد

---

۱ N-gram Language Models
۲ Dataset
۳ Twitter
۴ Intrinsic
۵ Perplexity
۶ Extrinsic

دادگان

سالانه یک ورکشاپ[۱] در زمینه تسک‌های کاربردی پردازش متن به نام SemEval برگزار می‌گردد. با توجه به گسترش متون توهین‌آمیز در شبکه‌های اجتماعی در سال ۲۰۱۹ دو تسک (شماره ۵ و ۶) در زمینه متون تنفرآمیز[۲] و توهین‌آمیز[۳] معرفی گردید. دادگان هر دو تسک از شبکه اجتماعی توئیتر جمع‌آوری گردیده و کلاس‌های دادگان به صورت دوتایی می‌باشد (هر پیام، تنفر/توهین آمیز می‌باشد یا نه).

اطلاعات دادگان در پوشه Data به شرح زیر می‌باشد.

دادگان	داده‌های آموزشی[۶]		داده‌های اعتبارسنجی[۵]		داده‌های تست[۴]	
	تنفر/توهین آمیز	غیر تنفر/توهین آمیز	تنفر/توهین آمیز	غیر تنفر/توهین آمیز	تنفر/توهین آمیز	غیر تنفر/توهین آمیز
HateEval	۹۰۰۰		۱۰۰۰		۳۰۰۰	
	۳۷۸۳	۵۲۱۷	۴۲۷	۵۷۳	۱۲۶۰	۱۷۴۰
OffenseEval	۱۳۲۴۰		۳۲۰		۸۶۰	
	۴۴۰۰	۸۸۴۰	۷۷	۲۴۳	۲۴۰	۶۲۰

۱ Workshop
۲ Hate Speech
۳ Offensive Language
۴ Test Data
۵ Validation Data
۶ Train Data

سوال ۱ – پیش‌پردازش (۱۰ نمره)

پیش‌پردازش یکی از مهم‌ترین مراحل پروژه‌های پردازش زبان طبیعی است که کیفیت آن بر روی نتایج، تاثیر مستقیم دارد.

هر توئیت (پیام در شبکه‌اجتماعی توئیتر) می‌تواند شامل متن پیام، شکلک[۱]، هشتگ[۲]، ارجاع[۳] و آدرس[۴] باشد. با استفاده از کتابخانه‌هایی که در فایل main.ipynb قرار گرفته است. نسبت به پیش‌پردازش موارد اشاره شده و نشانه‌گذاری متون اقدام نمایید.

راهنمایی:

- **منظور از نشانه‌گذاری مشخص‌نمودن ابتدا و انتهای تمامی جملات است.**

Twitter is a social network by @JackDorsey. Every tweet max length is 140 characters #twitter #introduction.

<s>Twitter is a social network by @JackDorsey</s> <s>Every tweet max length is 140 characters #twitter #introduction </s>

- معرفی کتابخانه‌های دیگر که باعث بهبود کیفیت پیش‌پردازش همانند تصحیح کلماتی که به صورت عامیانه نوشته شده‌اند، نمره اضافه تعلق می‌گردد.
- در این قسمت تنها بهترین پیش‌پردازش با توجه به نتایج قسمت نهایی اعلام گردد.

---

۱ Emoji (♡)
۲ Hashtag (#)
۳ Mention (@)
۴ URL

سوال ۲- ایجاد مدل‌های زبانی (۳۰ نمره)

الف)

بر روی دادگان آموزشی HateEval، مدل‌های زبانی یکتا[1] و دوتایی[2] در سطح ان‌گرام‌های کلمه و حرف ایجاد کنید.

راهنمایی:

- **مدل یکتایی کلمه: یک کلمه در متن را در نظر بگیرید.**
- **مدل دوتایی کلمه: دوکلمه کنار هم در متن را در نظر بگیرید.**
- **مدل یکتایی حرف: هر حرف در متن را در نظر بگیرید.**
- **مدل دوتایی حرف: دوحرف کنار هم در متن را در نظر بگیرید.**
- **جهت ایجاد مدل‌های زبانی می‌توانید از کتابخانه‌های موجود کمک بگیرید.**
- **زمانی می‌خواهید مدل‌های زبانی برای مثال یکتایی کلمه ایجاد کنید. برای کلاس تنفرآمیز (تمامی پیام‌های تنفرآمیز) یک مدل‌زبانی یکتایی کلمه و سپس برای کلاس غیرتنفرآمیز (تمامی پیام‌های غیرتنفرآمیز) نیز یک مدل زبانی یکتایی کلمه ایجاد کنید. با توجه به سوال ۳، زمانیکه می‌خواهید داده‌های اعتبارسنجی را در دو کلاس تنفرآمیز و غیرتنفرآمیز طبقه‌بندی کنید. به ازای هر پیام ورودی، مقدار سرگشتگی را هم به ازای مدل یکتایی تنفرآمیز و غیرتنفرآمیز محاسبه کنید. مدلی که مقدار سرگشتگی کمتری داشت (برای مثال تنفرآمیز)، در اینصورت پیام را تنفرآمیز علامت‌گذاری کنید.**

[1] Unigram
[2] Bigram

سوال ۳ – معیار سرگشتگی (۵۰ نمره)

الف)

در سوال ۲، برای هر مدل زبانی، ۲ ان‌گرام (تعداد کلاس‌ها) ایجاد گردید. حال برای هریک از پیام‌ها در دادگان اعتبارسنجی HateEval کلاس موردنظر را با توجه به معیار سرگشتگی تعیین کنید.

راهنمایی:

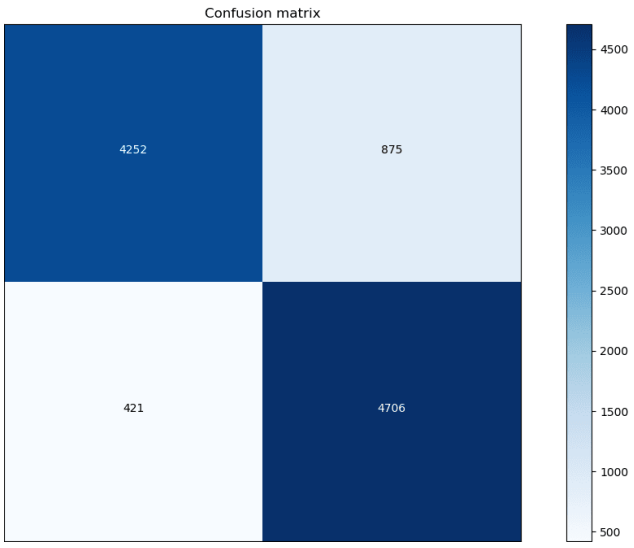
- **برای بدست آوردن سرگشتگی بر روی مدل‌های زبانی می‌توانید از کتابخانه‌های موجود کمک بگیرید.**
- **جهت هموارسازی از مدل add-one موجود در کتابخانه بالا می‌توانید استفاده کنید.**

ب)

برای هریک از مدل‌های زبانی (۴ مدل زبانی)، یک ماتریس در هم‌ریختگی رسم‌کنید.

راهنمایی:

نمونه‌ای از ماتریس درهم ریختگی با قدرت تشخیص بالا



ج)

به ازای هر مدل زبانی مقدار macro F1 و micro F1 را گزارش کنید. علت اختلاف چیست؟

د)

اندازه کلاس‌ها در دادگان HateEval به اصطلاح نامتوازن[1] می‌باشند (مقدار داده‌های تنفرآمیز کمتر از داده غیرتنفرآمیز می‌باشد). یکی از راه‌های ایجاد توازن افزایش داده می‌باشد. جهت توازن، تنها داده‌های آموزشی در OffenseEval را به HateEval به نحوی که توازن ایجاد گردد، اضافه کنید. آیا قدرت تشخیص مدل‌ها افزایش می‌یابد؟

ه)

با توجه به نتایج، مدل ان‌گرام حرف نسبت ان‌گرام چه امتیازاتی (به خصوص در مرحله‌ی پیش‌پردازش) می‌تواند داشته باشد.

سوال ۴- دادگان تست (۱۰ نمره)

جهت ارزیابی نتایج شما، دادگان تست در فولدر Data/HateEval/test.csv که فاقد کلاس (label) می‌باشند در اختیار شما قرار گرفته. با توجه به بهترین مدل‌زبانی که در سوال ۳ بدست آورید. لطفا برچسب هر متن به فرمت زیر در فایلی تحت عنوان Result.csv ارسال کنید.

text , label

This is outrageous! #StopIllegalImmigration #MeritImmigration , 1

[1] Unbalance

ملاحظات (حتما مطالعه شود)

تمامی نتایج شما باید در یک فایل فشرده با عنوان NLP_CA1_StudendID تحویل داده شود.

- خوانایی در گزارش نهایی از اهمیت ویژه‌ای برخوردار است. به تمرین‌هایی که به صورت کاغذی تحویل داده شوند یا به صورت عکس در سایت بارگذاری شوند، ترتیب اثر داده نخواهد شد. گزارش نهایی خود را حتما به صورت PDF در قالب فایل Solution.docx در سایت درس بارگذاری نمایید.
- گزارش به صورت تایپ شده در قالب PDF شامل شرح آزمایش‌های انجام شده، پارامترهای آزمایش، نتایج و تحلیل‌ها باشد تحت عنوان فایلی به عنوان Report قرار گیرد.
- جهت آسایش با استفاده از محیط jupyter notebook تمامی کدهای ارسالی بایستی در فایل main.ipynb با توجه به راهنمایی‌های لازم زده شود و کنار فایل گزارش ارسال گردد.
- مهلت تحویل تمرین به هیچ عنوان تمدید نخواهد شد. تمرین تا یک هفته بعد از مهلت تعیین شده با جریمه تحویل گرفته می‌شود که جریمه تاخیر تحویل تمرین تا یک هفته ۳۰درصد است.
- توجه کنید این تمرین باید به صورت تک نفره انجام شود و پاسخ‌های ارئه شده باید نتیجه فعالیت فرد نویسنده باشد (همفکری و به اتفاق هم نوشتن تمرین نیز ممنوع است). در صورت مشاهده تقلب به همه افراد مشارکت کننده، نمره تمرین صفر و به استاد نیز گزارش می‌گردد.
- در صورت بروز هرگونه مشکل با این ایمیل در ارتباط باشید:

  mailto:m.ghoroobi@gmail.com

مهلت تحویل بدون جریمه: ۲۴ام اسفند ۱۳۹۹

مهلت تحویل با تاخیر، با جریمه ۳۰ درصد: ۱ام فروردین ۱۴۰۰