دانشگاه صنعتی شریف

دانشکده مهندسی کامپیوتر

درس یادگیری ماشین

دکتر عباس حسینی

فاز دوم پروژه

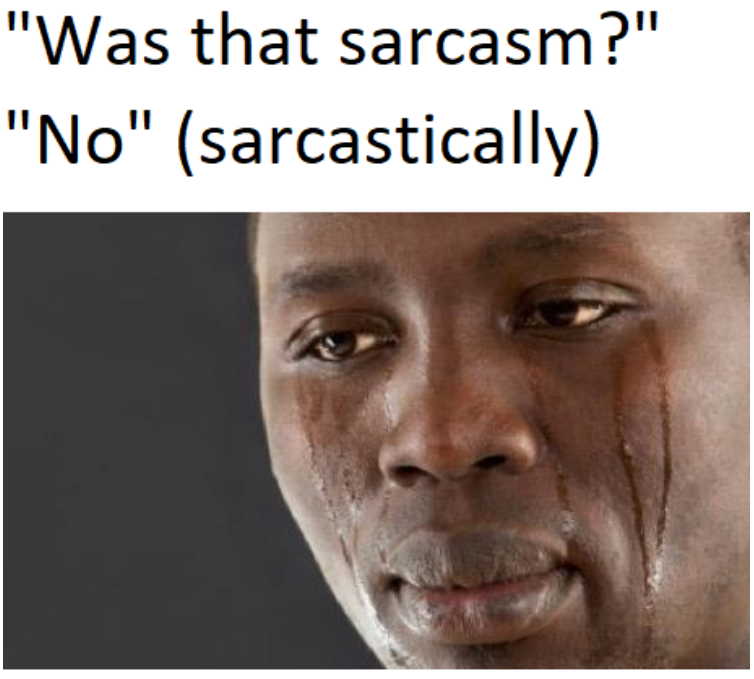
زمان انتشار: ۲۸ خرداد ۱۴۰۰

زمان تحویل: ۲۳ تیر ۱۴۰۰

۱- مقدمه: کنایه چیست؟

یکی از پراستفاده‌ترین کاربردهای مدل‌های یادگیری ماشین، در بحث تشخیص چارچوب، منظور و یا احساسات جملات است. درواقع ماشین به منظور برقراری ارتباط با انسان لازم است منظور و معنی نظرات بیان شده توسط انسان‌ها را متوجه شود. حتی در خیلی مواقع این ارتباط ممکن است میان دو انسان نیز باشد، مثلا حالتی را در نظر بگیرید که ماشین قرار است پیامی را از زبانی به زبان دیگر ترجمه کند.

یکی از اصلی‌ترین چالش‌های مساله‌ی مطرح شده، هنگامی است که پیام موردنظر همراه با کنایه باشد و در نتیجه تشخیص کنایه آمیز بودن یا نبودن پیام و در گام‌های بعدی، تشخیص منظور اصلی پیام به یکی از مسائل مهم و جذاب تبدیل شده است. تشخیص کنایه در خیلی مواقع حتی برای انسان هم آسان نیست، پس احتمالا برای ماشین چند برابر دشوارتر نیز خواهد بود. تمرکز اصلی این پروژه تشخیص کنایه آمیز بودن یا نبودن یک پیام با استفاده از ویژگی‌های آن است. (۱)



۲- شرح مساله: ماشین می‌تواند کنایه را تشخیص دهد؟

همانطور که در بخش قبل گفتیم، هدف اصلی این پروژه تشخیص کنایه آمیز بودن یا نبودن یک پیام متنی (در قالب یک مساله‌ی دسته‌بندی باینری) است.

اصلی‌ترین و مهم‌ترین بخش داده‌های این مساله، داده‌های متنی هستند و در نتیجه باید این پیام‌های متنی را به فرمتی تبدیل کنید که یک مدل، با استفاده از آن‌ها بتواند آموزش داده شود. پیدا کردن این تبدیل یکی از مهم‌ترین بخش‌های این پروژه است. در ساده‌ترین رویکرد ممکن، می‌توان برای هر پیام یک لیست به اندازه‌ی تعداد کلمات کل این پیکره[1] درنظر گفت که تمام درایه‌های آن به جز کلمات آن پیام صفر است و درایه‌ی متناظر با کلمات این پیام، تعداد تکرار آن‌ها باشد. بنابراین امبدینگ (embedding) انتخاب شده برای این مساله، تاثیر مستقیمی روی عملکرد و کارایی مدل خواهد داشت و باید با دقت انتخاب شود. در فایل گزارش باید فرایند پیدا کردن امبدینگ و خود امبدینگ استفاده شده را به طور کامل شرح دهید. (۵)

همچنین چالش‌های دیگری نیز وجود دارد که در روند حل مساله با آن‌ها روبرو خواهید شد و باید با دقت نسبت به آن‌ها برخورد کنید. سعی کنید ایده‌هایی که در فرآیند حل مسئله استفاده می‌کنید را حتما در فایل گزارش ثبت کنید، به طور مثال ممکن است کلماتی را پیدا کنید که خیلی informative نیستند و با حذف آن‌ها حجم داده‌ی مساله را کاهش دهید، یا روشی پیاده‌سازی کرده‌اید که بتواند کلمات هم‌معنی را پیدا کند و از این طریق باعث بهبود عملکرد مدل شود یا هر ایده‌ی دیگری که ممکن است عملکرد مدل را بهبود دهد. مرحله‌ی پیش‌پردازش در این مساله اهمیت بسیار زیادی دارد. همچنین محدودیتی نسبت به استفاده از مدل‌های عمیق و مدل‌های پردازش زبان‌های طبیعی در این پروژه -به شرط توضیح کافی در فایل گزارش- ندارید.

۳- توضیحات داده‌گان

داده‌گانی که برای این پروژه استفاده می‌کنیم از کامنت‌های شبکه اجتماعی ردیت جمع‌آوری شده است. اگر قبلا تجربه‌ی کار با این سایت را نداشتید، پیشنهاد می‌کنیم سری به آن بزنید و ساختار کامنت‌ها را در آن مشاهده کنید. (۲)

[1] corpus

کل داده‌گان مورد نظر شامل حدود یک میلیون نظر (comment) از این سایت است. هر نظر همراه با ۱۰ ویژگی متناظر با آن آورده شده است. یکی از این ویژگی‌ها برچسب کنایه آمیز بودن یا نبودن نظر متناظر با آن پیام است. باقی ویژگی‌ها نیز می‌توانند در احتمال کنایه آمیز بودن یک پیام نقش داشته باشند. به طور مثال ممکن است در برخی روزهای سال به دلیل پیش آمدن مسائل مختلفی، احتمال کنایه آمیز بودن پیام بالاتر از روزهای دیگر باشد، یا برخی اشخاص به طور کلی از کنایه بیشتر استفاده کنند. استفاده‌ی مناسب از این ویژگی‌ها، از جمله وظایفی است که در این پروژه به عهده دارید.

این یک میلیون نظر به نسبت ۴ به ۱ به ترتیب به عنوان داده‌ی تمرین و آزمون جداسازی شده‌اند و شما می‌بایست از هر کدام در مراحل آموزش و آزمودن مدل‌هایتان استفاده کنید.

۴- نحوه‌ی ارزشیابی

فاز قبلی پروژه تاکید زیادی روی اکتشافات روی داده به منظور مدل‌سازی و ساختن یک داده‌گان نهایی مناسب برای اجرای یک مدل پیش‌بینی بود. در این فاز از پروژه‌ی درس تمرکز را از مراحل اولیه‌ی انجام پروژه به مراحل بعدی، یعنی گرفتن نتایج مناسب و تست کردن مدل‌های یادگیری مختلف می‌بریم.

۴.۱- تحلیل اکتشافی داده[2] - (۱۰ + ۵ درصد)

همچون فاز قبل، در قسمت نخست از پروژه شما انتظار داریم تا جایی که ممکن است بتوانید شواهد مفیدی از داده برای مراحل بعدی ایجاد کنید تا بتوانید ویژگی‌های مهم را استخراج کنید و مدل بهتری را طراحی کنید.

در این قسمت همچنین مهم است که علاوه بر فراهم کردن شواهد داده‌ای بتوانید برداشت خود را از شواهد بیان کنید و بگویید که این شواهد چه کمکی به شما در اخذ نتایج بهتر و شناخت داده کرده است.

[2] Exploratory Data Analysis (EDA)

۴.۲- مهندسی ویژگی‌ها[3] - (۱۵ + ۱۰ درصد)

پس از انجام مرحله‌ی قبل و به دست آوردن شواهد کافی کیفی و کمی از وضعیت داده‌گان، نوبت این است که آن را برای وارد شدن به مرحله‌ی یادگیری آماده کنید. یکی از کارهای مهمی که در این بخش (و در این فاز پروژه) می‌بایست انجام دهید تبدیل کردن متن‌های کامنت‌های موجود در داده‌گان به حالتی است که بتواند برای ماشین مفهوم نزدیک‌تری به واقعیت را نمایش داده و در نتیجه باعث بهبود عملکرد ماشین در انجام وظیفه[4]ی نهایی بشود. فضای ابعاد ورودی متن‌های موجود در داده‌گان ابعاد بسیار بزرگی دارد و همچنین معنای کلمات آن برای ماشین قابل فهم نیست. متدهایی وجود دارند تا بتوان به وسیله‌ی آنها فضایی ایجاد کرد که در آن فضا ابعاد متون کوچک‌تر باشد و/یا معنای جملات یا لغات در قابل فهم باشد.

تست و مقایسه کردن متدهای مختلف embedding (چه ساده و چه پیشرفته) در این قسمت از ارزش بالایی برخوردار است و در نتیجه‌ی نهایی مدل هم تاثیر زیادی خواهد داشت. دقت کنید که شما باید حداقل دو امبدینگ مختلف را بررسی و مقایسه کنید. دوتا از پراستفاده‌ترین امبدینگ‌ها word2vec و tf-idf هستند و لازم است این دو امبدینگ را پیاده‌سازی، تست و مقایسه کنید. استفاده از دیگر امبدینگ‌ها یا ایجاد تغییرات خلاقانه در این امبدینگ‌ها باعث کسب نمره‌ی امتیازی خواهد شد.

۴.۳- تست مدل‌های مختلف - (۲۰ + ۱۵ درصد)

برای حل کردن این مساله با کیفیت بالا می‌بایست ابتدا ابزارهای مختلف پاسخگویی به سوال را بررسی کنید و عملکرد هرکدام را بسنجید. از شما می‌خواهیم تا مسیری را که پس از انجام استخراج ویژگی‌ها[5] تا انتخاب مدل مناسب نهایی طی می‌کنید گزارش کرده و نتایج به دست آمده را با یکدیگر مقایسه کنید. برای انجام این مرحله می‌توانید مدل‌های ماشین لرنینگ ساده‌ی مانند Logistic Regression تا مدل‌های شبکه عصبی پیچیده Recurrent را امتحان کنید. همچنین با جستجو کردن ببینید مسائل مشابه بیشتر با چه راه‌حل‌هایی به نتایج مختلف رسیده‌اند. دقت کنید که لازم است حداقل ۳ مدل مختلف را بررسی و مقایسه کنید. سعی کنید همانطور که اشاره شد، مدل اولی که پیاده‌سازی می‌کنید از نظر پیچیدگی ساده تقلی شود (به عنوان مثال Logistic Regression). مدل دومی، از نظر پیچیدگی بین دو حالت گفته شده قرار گیرد (به عنوان مثال SVM) و و مدل سوم مدلی باشد که از پیچیدگی به نسبت بیشتری برخورد باشد (به عنوان مثال مدل‌های مبتنی بر یادگیری عمیق). تست‌کردن دو مدل اول نمره‌ی اصلی محسوب شده و

---

[3] Feature Engineering
[4] task
[5] Feature extraction

خلاقیت در استفاده از مدل‌ها و همچنین قسمت سوم استفاده از مدل‌های پیچیده‌تر (عمیق) بعنوان نمره‌ی امتیازی در نظر گرفته خواهند شد.

بعلاوه **نتیجه‌ی هر مدل را به اندازه‌ی تفسیرپذیری آن تفسیر کنید**. تفسیر کردن به این معناست که بتوانید گزارش دهید که یک مدل (فارغ از کیفیت عملکرد) چگونه دارد تصمیم‌گیری می‌کند و به چه ویژگی‌هایی در داده توجه بیشتری می‌کند! انتظار داریم تا مدل‌ها (به اندازه‌ای که تفسیرپذیر هستند) در هنگام تست مورد تفسیر هم قرار بگیرند؛ مثلا شاید بتوان در بعضی از مدل‌ها به این سوال پاسخ داد: «وجود کدام یک از کلمات در جملات ورودی بیشترین تاثیر را در کنایی بودن جمله خواهد داشت؟». (۶)

نکته: توجه کنید که ممکن است پس از انجام تست روی مدلی به این نتیجه برسید که برای آن مدل خاص یک مهندسی متفاوت از ویژگی‌ها می‌تواند نتیجه‌ی بهتری بدهد و لازم است تا داده‌گان نهایی را با توجه به مدل انتخابی تغییر بدهید! بنابراین مراحل دوم و سوم پروژه می‌توانند به صورت رفت و برگشتی انجام شوند و لزوماً در طول هم نیستند.

۴.۴- نتایج مدل نهایی - (۳۰ درصد)

نیمی از نمره‌ی این بخش مربوط به بررسی کردن و توضیحاتی است که شما از دلایل انتخاب مدل نهایی می‌دهید. برای این موضوع باید به سوالات زیر پاسخ بدهید:

۱- با چه معیارهایی این مدل را نسبت به باقی مدل‌های تست شده مقایسه شده؟

۲- در این معیارها وضعیت این مدل نهایی چگونه بوده و چگونه این مدل در نهایت انتخاب شده؟

۳- مدل نهایی در چه حالاتی از داده ضعیف عمل می‌کند؟ علت این عملکرد ضعیف چیست؟

۴- نقاط قوت و ضعف مدل به طور کلی چیست؟

نصف باقی (۱۵/۳۰) نمره‌ی شما در این بخش از رابطه‌ی زیر به دست می‌آید.

F1: امتیاز اف[6] (!) به دست آمده از مدل نهایی شما (بین صفر تا صد) (بیشتر بهتر)

VAR: واریانس امتیازات به دست آمده در امتیازات f چارک دوم و سوم دانشجویان (کمتر بهتر)

[6] F-score

$$F_1 \times \frac{(1 + \frac{1}{VAR})}{5}$$

این امتیازات بایستی توسط مدل نهایی شما هنگام آزموده شدن با داده‌ی آزمون[7] که در اختیارتان قرار گرفته به دست بیاید. می‌دانید که باید تا قبل از شروع فاز تست، مدل هیچ برخوردی با داده‌ی آزمون نداشته باشد و فقط با داده‌ی تمرین[8] آموزش داده شود.

۴.۵- گزارش - (۲۰ درصد)

جامعیت، دقت و صحت گزارش شما سه ویژگی‌ای است که به آن کیفیت می‌بخشد. لذا سعی کنید تمامی مراحلی را که تا به خروجی رسیدن طی می‌کنید به خوبی مستند کنید و در قالب گزارش مناسبی تحویل دهید. توجه کنید که این بخش نمره‌ی امتیازی‌ای ندارد. این امر به این معنی است که تحویل دادن یک گزارش خیلی باکیفیت از پروژه یک بخش واجب از پروژه است! لطفاً وقت کافی برای انجام آن بگذارید و به منظور از دست نرفتن جزئیات این بخش را موازی کارهای دیگر پیش ببرید.

دقت کنید که **پروژه‌ی بدون گزارش مورد بررسی قرار نمی‌گیرد و نمره‌ای هم در بر ندارد.**

[7] Test Data
[8] Training Data

۵- چند نکته

- این پروژه به منظور تقویت یادگیری خارج از مباحث درس شما کمی از مباحث اصلی درس خارج است. این مورد به عمد و به قصد «ایجاد چالش یادگیری به منظور کاربرد آنی» وجود دارد و انتظار نداریم که کیفیت و سادگی انجام آن توسط شما مانند مباحث تمارین مرتبط درس باشد.
- بعضی توضیحات جامع در راجع به موارد ارزشیابی که در فاز قبل آمده بود در این مستند تکرار نشده.توصیه می‌شود قسمت ارزشیابی فاز قبلی را مروری بکنید.
- یکی از اهداف این پروژه تقویت مهارت یادگیری شما بسته به نیاز مسئله است؛ بنابراین جستجو در منابع مختلف و انجام مطالعات و دیدن مثال‌های مشابه بسیار مورد استقبال قرار خواهد گرفت. از شما می‌خواهیم تا **منابعی که مورد مطالعه قرار می‌دهید را در گزارش خود بیاورید**.
- برای هر مساله در مورد زبان برنامه‌نویسی و ابزارهای مورد استفاده **هیچ محدودیتی** وجود ندارد. اگر چه استفاده از زبان پایتون توصیه می‌شود.
- فاز دوم پروژه به صورت گروهی و در قالب **گروه‌های سه نفره** است.
- مهلت تحویل پروژه تمدید نخواهد شد.
- خروجی‌های مورد نیاز پروژه: ۱- کد ۲- مستند توضیح
  - استفاده از خروجی jupyter notebook به دلیل اینکه خروجی‌های مورد نیاز را به صورت یک‌پارچه قابل ارائه می‌کند توصیه می‌شود.
  - اگر از ژوپیتر استفاده می‌کنید می‌توانید مستندات و توضیحات را در قالب همان یک فایل تحویل دهید.
- آپلود شدن پروژه توسط یکی از اعضا کافی است. نام و شماره‌دانشجویی افراد را در گزارش ذکر کنید.
- همگی در ارائه‌ی پروژه حضور داشته باشید و سعی کنید در انجام پروژه سهم یکسانی را بر عهده بگیرید.
- دقت کنید که تمامی خروجی‌های پروژه به عنوان دارایی معنوی[9] هر عضو گروه تلقی خواهد شد. تیم تدریس نتایج ارائه شده توسط شما را صحت‌سنجی خواهند کرد. لذا هرگونه شباهت نامتعارف کد شما با دیگر گروه‌ها کپی صرف از منابع یا برداشت بدون ذکر منبع پس از بررسی به عنوان تخلف آموزشی مورد پیگرد قرار خواهد گرفت.
- مشکلات و سوالاتتان از هر جنس را در **پیاتزای درس** بپرسید.

[9] Intellectual Property

۶- لینک‌های مفید

پیشنهاد می‌کنیم، پیش از شروع پروژه به لینک‌های زیر توجه کنید.

۱- کنایه چیست؟

۲- توضیحات بیشتر درمورد داده‌گان و بررسی تعداد روش برای حل مساله‌ی تشخیص کنایه

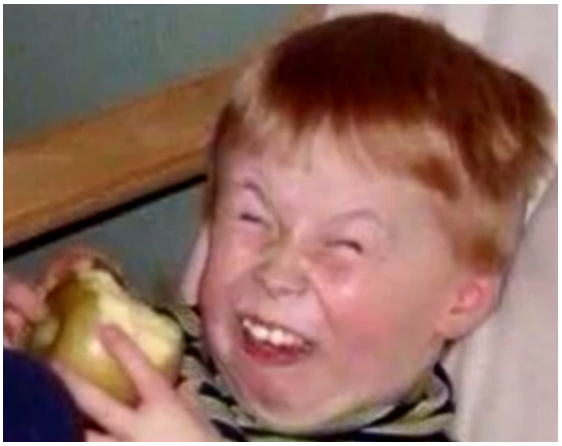
۳- پست مدیوم در موضوع تشخیص کنایه

۴- مقالات به همراه کد در رابطه با موضوع تشخیص کنایه در سایت paperswithcodes

۵- در مورد word-embedding (سوالات موجود قسمت نظرات لینک را هم مطالعه کنید!)

۶- در مورد چگونگی و چرایی تفسیرپذیری

۷- یک منبع طولانی در مورد NLP Basics - اگر علاقه‌مند بودید



امیدواریم از این پروژه‌ی کوتاه و ساده لذت ببرید.

موفق باشید.