به نام خدا

«پروژه دوم»

بوت‌کمپ هوش مصنوعی کوئرا

تابستان ۱۴۰۴

مهلت ارسال پاسخ: تا ساعت ۲۳:۵۹ روز سه‌شنبه ۲۹ مهرماه

زمان ارائه‌ی گروهی: شنبه و یکشنبه ۲ و ۳ آبان ماه

مسئله‌ی ۱: تشخیص تصادف

تشخیص وقوع تصادف از روی ویدئو مسئله‌ای کاربردی در بینایی ماشین است که می‌تواند زمان اطلاع‌رسانی و اعزام امداد را کاهش دهد. در این پروژه قصد داریم با تکیه بر یک مجموعه‌داده ویدئویی کوتاه‌مدت (کلیپ‌های چند ده ثانیه‌ای)، مدلی بسازیم که با مشاهدهٔ هر ویدئو، وقوع تصادف را پیش‌بینی کند.

مجموعه‌داده شامل ویدئوها و چند ستون متادیتا است؛ مهم‌ترین‌شان:

- time_of_event: زمان وقوع تصادف در همان ویدئو (در صورت نبودِ تصادف، مقدار آن تهی/NaN است).
- time_to_accident، scene، weather، light_conditions، time_of_alert: اطلاعات تکمیلی که می‌توانند در تحلیل و مدل‌سازی مورد استفاده قرار گیرند.

برای برچسب‌گذاری ساده، کافی است وجود/عدم‌وجود مقدار، در time_of_event را به عنوان نشانگر قرار دهیم.

حال باید یک سیستم باینری کلسیفیکیشن طراحی کنیم که با ورودی گرفتن یک ویدئو، تشخیص دهد تصادف رخ داده است یا خیر. به طور کلی، در طول دوره کار با ویدیو (هر چه کوتاه) رو نداشتیم. در این بخش هم این رو نداریم، ویدیو هایی که داریم در واقع تشکیل شده از تعداد بالایی فریم هستند که طبیعت سیکونشال دارن. حال باید از روی این ویدیو ها، با فواصل ثابت (در حد چند میلی ثانیه) تعدادی عکس (فریم) استخراج کنیم و مدل کلسیفایر رو براساس این ورودی ها بسازیم.

در این بخش احتمال خیلی زیادی وجود دارد که به مشکلاتی از جمله سیو/لود کردن دیتا، مدل و مشکلات سخت افزاری برخورد کنید. نتیجه ای که ما نیاز داریم با پایین ترین سرعت اینترنت و گوگل کولب عادی قابل انجامه. در واقع این چالش ها راه حل دارن که درآوردن این راه حل هم بخشی از پروژه هستش.

مسئله‌ی ۲: تحلیل احساس نظرات

جهت دریافت مجموعه‌داده‌ی **آموزش** این بخش **اینجا** کلیک کنید.
جهت دریافت مجموعه‌داده‌ی **آزمون** این بخش **اینجا** کلیک کنید.
جهت دریافت **جدول نگاشت** شناسه‌ی محصولات به عنوان و برند آن‌ها **اینجا** کلیک کنید.

مقدمه

تجزیه و تحلیل احساس (Sentiment Analysis) شاخه‌ای از پردازش زبان طبیعی (NLP) است که سعی دارد با استفاده از الگوریتم‌های یادگیری ماشین به شناسایی و استخراج خودکار اطلاعات ذهنی از متن بپردازد. هدف از تجزیه و تحلیل احساسات، تعیین احساسات یا عواطف پشت یک متن است، خواه مثبت، منفی یا خنثی باشد. تحلیل احساس در صنعت کاربرد بسیاری دارد و می‌توان آن را برای طیف گسترده‌ای از داده‌های مبتنی بر متن، از جمله پست‌های رسانه‌های اجتماعی، بررسی محصول، بازخورد مشتریان، مقالات خبری و موارد دیگر اعمال کرد. در این مسئله نیز مجموعه‌داده‌ای از نظرات ثبت‌شده برای کالاهای الکترونیکی در فروشگاه آمازون در اختیار شما قرار گرفته تا بتوانید به استخراج بینش‌هایی از این داده‌ها و همچنین ساخت یک مدل تحلیل احساس بپردازید.

توضیحات مجموعه‌داده

جزئیات ستون‌های این مجموعه‌داده به شرح زیر است:

- **overall**: امتیاز محصول (توسط فرد نظر دهنده) از ۱ تا ۵
- **vote**: تعداد رای‌های دیدگاه از نظر مفید بودن (helpful)
- **verified**: آیا تایید و منتشر شده است یا خیر
- **reviewTime**: تاریخ ثبت نظر
- **reviewerID**: شناسه‌ی شخص نظر دهنده
- **Asin**: شناسه‌ی محصول (برای دسترسی به لینک محصول می‌توانید شناسه را بعد از https://www.amazon.com/dp/ قرار دهید)
- **style**: دیکشنری برخی توضیحات محصول مثل رنگ و سایز و غیره
- **reviewerName**: نام شخص نظر دهنده
- **reviewText**: متن نظر
- **summary**: خلاصه‌ی نظر

- **unixReviewTime**: زمان ثبت نظر با فرمت unix time

بخش ۱) تجزیه و تحلیل اولیه از داده‌ها

در ابتدا از شما می‌خواهیم به سوالات زیر پاسخ داده تا بینش بهتری از داده‌های موجود پیدا کنید:

۱. توزیع ستون overall را رسم کنید. آیا مجموعه‌داده متوازن است؟ اگر خیر، آیا نیاز است برای مدل‌سازی خود آن را متوازن کنید؟ چه راه‌حلی برای این کار پیشنهاد می‌کنید؟

۲. فرض کنید نظراتی که مقدار ستون overall آن‌ها ۴ یا ۵ است را همراه با حس مثبت، نظراتی که مقدارشان ۳ است را خنثی و نظراتی که مقدارشان ۱ یا ۲ است را حس منفی بدانیم. به‌ازای هر کدام از این سه دسته یک ابر کلمات (Word Cloud) رسم کنید تا بتوان کلمات پرتکرار هر دسته را مشاهده کرد. تا حد ممکن سعی کنید ابر کلمات به‌دست‌آمده شامل اطلاعات مفیدی باشد و کلمات زائد (Stop words) بین آن‌ها وجود نداشته باشد. آیا اشتراکی بین کلمات دسته‌ی مثبت و منفی وجود داشته است؟ چگونه آن‌ها را تفسیر می‌کنید؟

۳. از بین نظردهندگان، ۱۰ نفری که در مجموع نظرات‌شان بیشتر مفید واقع شده (مجموع vote بیشتری داشته‌اند) را پیدا کنید. به‌عنوان مثال اگر شخص «الف» مجموعاً ۲۰ نظر ثبت کرده باشد، باید مجموع مقدار vote تمام ۲۰ نظر وی را محاسبه کنید. این کار را برای تمام افراد انجام داده و ۱۰ نفر برتر را پیدا کنید. نام هر فرد و مجموع vote آن را به‌ترتیب نمایش دهید.

۴. هیستوگرام طول متن (تعداد کاراکتر) ستون reviewText را رسم کنید. یک‌بار با حالت اصلی رسم کنید و یک‌بار به‌صورت فیلترشده (آن دسته‌هایی که تعداد نمونه‌های کم و پرتی دارند را در نظر نگیرید) ترسیم کنید. انتخاب تعداد دسته‌ها (bins) برعهده‌ی خودتان است و نمودار خروجی شما باید مناسب و خوانا باشد. آیا نیاز است در هنگام مدل‌سازی محدودیتی روی تعداد کاراکترها بگذاریم؟ اگر بله، بازه‌ی پیشنهادی شما چه عددهایی است؟

۵. کدام محصولات بیشترین امتیاز ۵ را کسب کرده‌اند؟ ۱۰ مورد برتر را به‌ترتیب به‌صورت یک جدول شامل نام برند، عنوان محصول و تعداد نظرات با امتیاز ۵ نمایش دهید.

۶. ابتدا ۱۰ برندی که بیشترین تعداد نظر را داشته‌اند پیدا کنید. سپس میانگین امتیاز هر کدام را محاسبه کرده و یک جدول شامل نام برند و میانگین امتیاز آن به‌ترتیب میانگین امتیاز نمایش دهید.

بخش ۲) میزان رضایت از یک جنبه‌ی مشخص

فرض کنید می‌خواهیم نظراتی که در آن‌ها درباره‌ی ضمانت کالا (گارانتی، وارانتی و غیره) صحبت شده را برای هر محصول پیدا کرده و میانگین امتیاز (overall) کاربران را پیدا کنیم. این بدین معنی‌ست که قصد داریم تقریبی از میزان رضایت کاربران را درباره‌ی ضمانت کالای مربوطه به دست آوریم. یک راه ساده این است که به‌ازای هر نظر ثبت‌شده برای یک محصول دقیقاً به دنبال کلماتی مثل warranty یا guarantee بگردیم و اگر چنین کلمه‌ای وجود داشت در نتیجه در آن نظر درباره‌ی این جنبه از کالا بحث شده است. اما چنین روشی نمی‌تواند واقعا تمام داده‌های مورد نظر را پیدا کند زیرا که ممکن است در متن کاربر، کلمات مشابه یا مترادف دیگری به‌جای این کلمه استفاده شده باشد، یا حتی ممکن است فرد در نوشتار این کلمه غلط تایپی داشته باشد.

یک راه پیشنهادی برای حل این مسئله این است که ابتدا به کمک بردارهای تعبیه (به‌عنوان مثال بردار word2vec یا بردارهای از پیش‌آموخته‌ی مدل‌های زبانی عظیم مثل GPT یا Cohere)، کلمات مشابه warranty یا guarantee را نیز پیدا کرده و سپس علاوه بر دو کلمه‌ی اصلی، به دنبال چنین کلماتی نیز بگردید. فراموش نکنید که غلط‌های املایی ممکن و رایج را نیز در نظر بگیرید.

بنابراین در این بخش نیاز است ابتدا به‌ازای هر دو کلمه، کلمات مشابه آن‌ها را توسط این روش پیدا کرده، سپس نظراتی که در آن‌ها حداقل یکی از این کلمات ظاهر شده بود را جدا کرده و در نهایت طبق این داده‌ی فیلترشده، میانگین امتیاز هر کالا را محاسبه و گزارش کنید.

نکته: راه‌حل شرح‌داده‌شده صرفاً یک راه‌حل ساده‌ی پیشنهادی بوده و اگر علاقه دارید از روش خلاقانه‌ی دیگری بهره ببرید با تایید منتور بلامانع است و در صورت بهتر بودن رویکرد شما شامل نمره‌ی اضافه نیز خواهد شد.

بخش ۳) مدل تحلیل احساس

در این قسمت به حل مسئله‌ی خوش‌تعریف تحلیل احساس خواهید پرداخت. نیاز است مدلی طراحی کنید که با دریافت متن نظر کاربر، احساس/رضایت وی نسبت به کالا را بین عددی از ۱ تا ۵ تعیین کند. بنابراین متغیر هدف شما همان ستون overall خواهد بود. ورودی مدل شما می‌تواند علاوه بر متن نظر (reviewText) شامل خلاصه‌ی نظر یا اطلاعات دلخواه دیگری نیز باشد اما مبنای اصلی و الزامی کار همان متن نظر خواهد بود.

برای این قسمت مجاز هستید از هر مدل دلخواهی استفاده کنید، اما به نکات زیر توجه کنید:

- اگر از یک مدل پیش‌آموخته (pre-trained) استفاده می‌کنید، حتماً نیاز است آن را ویژه‌ی دامنه‌ی مسئله‌ی خود آموزش دهید (fine-tune کردن یا اضافه کردن لایه‌های دیگر).

- تمام اعضای فعال گروه شما باید تسلط کافی نسبت به الگوریتم و پیاده‌سازی آن را داشته باشند. بنابراین اگر قصد استفاده از مدلی همچون ترنسفورمرها دارید سعی کنید معماری مورد استفاده را در گروه خود مطالعه و بررسی کنید.

نکات کلی

- لزومی به استفاده از تمامی داده‌های موجود در داده‌های آموزشی وجود ندارد. می‌توانید برای کاهش منابع سخت‌افزاری مورد نیاز برای فرآیند آموزش تنها از بخشی از مجموعه‌داده استفاده کنید.
    - فراموش نکنید که بخشی از داده‌های آموزشی را برای اعتبارسنجی (validation) جدا کنید.
- استفاده از هر نوع پیش‌پردازش، کتابخانه و مدلی، آزاد است. تنها شرط لازم برای استفاده از موارد ذکر شده، تسلط تمامی اعضای فعال در گروه بر آن‌ها است.
- بخش مهمی از این مسئله، نحوه‌ی پیش‌پردازش داده‌های متنی است. بنابراین سعی کنید از تکنیک‌های مختلفی جهت پیش‌پردازش هر چه بهتر متن‌ها بهره ببرید و نیاز است انتخاب‌های شما برای این مرحله همراه با دقت کافی و قابل استدلال باشد. به‌عنوان مثال اگر قصد حذف کلمات زائد (Stop words) را دارید دقت کنید که کلمات مهم برای این نوع مسئله‌ی خاص حذف نشوند.

ارزیابی نهایی

مجموعه‌ی آزمونی که در اختیار شما قرار گرفته شامل برچسب حقیقی نیست. نیاز است پس از تکمیل کار خود، از مدل نهایی برای پیش‌بینی برچسب این نمونه‌ها استفاده کرده و یک فایل csv به شکل جدول زیر آماده کنید. پس از اتمام مهلت ارسال پروژه و آپلود فایل‌های شما، به مدت چند ساعت بخش جدیدی در سامانه باز خواهد شد تا بتوانید این فایل را آپلود کرده و نتیجه‌ی مدل خود را مشاهده کنید.

معیار ارزیابی: f1 score با روش میانگین‌گیری میکرو (micro)

ساختار فایل: نام فایل شما باید q2_submission.csv باشد و شامل یک ستون از احساس پیش‌بینی‌شده (predicted) باشد. ردیف اول باید مربوط به نمونه‌ی اول داده‌های آزمون، ردیف دوم مربوط به نمونه‌ی دوم و الی آخر باشد. لطفاً نمایه‌ها (index) را نیز ذخیره نکنید. به نمونه‌ی زیر دقت کنید:

predicted
5
0

نکته‌های اصلی

- به دلیل سنگین بودن داده‌ها سعی کنید این پروژه را بر بستر گوگل کولب پیش ببرید.
- کدهای خود را خوانا و تمیز بنویسید. خروجی هر قسمت باید نمایش داده شده باشد.
- به انتخاب‌های خود در هر مرحله از کار دقت کنید، زیرا باید بتوانید برای آن‌ها دلیل موجهی بیاورید.
- در هنگام پیاده‌سازی نظرات سایر اعضای تیم را جویا شوید و سعی کنید زودتر یک نسخه‌ی اولیه از کار خود را آماده کنید تا زمان کافی برای بررسی و کشف باگ‌های آن توسط اعضای تیم و منتور وجود داشته باشد.
- به نکات ذکر شده در ارتباط با نحوه‌ی ارسال فایل در صفحه‌ی پروژه در کلاس توجه فرمایید.

بخش امتیازی (بیشینه: ۴۰ نمره)

- مستندسازی غنی و مناسب در نت‌بوک‌ها (۴ نمره)
- استفاده از گیت و مشارکت فعال در آن (۳ نمره)
- تحلیل‌های بیشتری که بینش‌های مفیدی را به ارمغان آورند (هر تحلیل مفید ۲ نمره و حداکثر ۶ نمره)
- طراحی داشبورد برای قسمت‌های تحلیلی (داشبورد پایه حداکثر ۳ نمره و داشبورد تعاملی حداکثر ۶ نمره)
- طرح مسئله‌ی جدید و تلاش برای حل آن با تایید منتور (حداکثر ۲۱ نمره. استفاده از متا دیتا در مسئله اول هم در این بخش قرار میگیرد)

موفق باشید 🤩