« پروژه ۲ »

بوت‌کمپ هوش مصنوعی کوئرا

بهار۱۴۰۴

مهلت ارسال پاسخ: تا ساعت **۲۳:۵۹** روز یک‌شنبه ۲۲ تیرماه

زمان ارائه‌ی گروهی: چهارشنبه و پنج‌شنبه ۲۵ و ۲۶ تیرماه

مسئله‌ی ۱: دسته‌بندی غذا

جهت دریافت مجموعه‌داده‌ی **آموزش** این بخش **اینجا** کلیک کنید.

مقدمه

فرض کنید صاحب یک اپلیکیشن فروش آنلاین غذا هستید، فروشندگان می‌توانند اقلام خوراکی خود را در این اپلیکیشن قرار دهند تا کاربران با بررسی قیمت و عکس خوراکی‌ها و نظرات کاربران دیگر، خوراکی مورد نظر خود را انتخاب کنند. در دنیای پویا و رقابتی فروشگاه‌های اینترنتی، توسعه و بهبود فرآیندهای فروش و خدمات به مشتریان اهمیت بسیاری دارد. از آن‌جا که تعداد محصولات موجود در این فروشگاه‌ها به سرعت افزایش می‌یابد و تنوع آن‌ها نیز به شدت گسترده است، تشخیص دقیق و سریع دسته‌بندی محصولات از تصاویر تبدیل به یک چالش بزرگ در جهت بهبود تجربه‌ی مشتریان و بهره‌وری کسب‌وکارها شده است.

تصویربرداری محصولات و ارائه‌ی تصاویر مرتبط با آن‌ها به مشتریان، ابزار حیاتی در این صنعت است. اما چگونه می‌توان از تصاویر استفاده کرده تا محصولات را به درستی دسته‌بندی کنیم؟ به عبارت دقیق‌تر، چگونه می‌توانیم از تکنولوژی یادگیری عمیق بهره ببریم تا تشخیص دسته‌بندی محصولات از تصاویر را با سرعت و دقت بالا داشته باشیم؟

در قدم اول برای ایجاد یک اپلیکیشن هوشمند باید بتوانیم نوع غذای موجود در یک عکس را شناسایی کنیم و این دقیقا کاری است که ما انتظار داریم شما در این سوال انجام دهید. هدف اصلی این پروژه، توسعه‌ی مدل‌های یادگیری عمیق به منظور دسته‌بندی هرچه دقیق‌تر محصولات یک اپلیکیشن اینترنتی فروش غذا از تصاویر آن‌هاست. توسعه‌ی این مدل‌ها نه تنها به بهبود تجربه‌ی مشتریان و کاهش خطاهای انسانی کمک می‌کند بلکه به بهره‌وری بالاتر در مدیریت موجودی و بهره‌برداری از داده‌های تصویری نیز می‌انجامد.

توضیحات مجموعه‌داده

داده‌های معرفی‌شده برای این مسئله شامل دو مجموعه‌ی آموزشی و آزمون است. مجموعه‌داده‌ی آموزش شامل ۲۲ دسته داده است که از هر نوع داده تعدادی عکس موجود است. تصاویر مربوط به هرکدام از دسته‌ها داخل پوشه‌ای با برچسب آن دسته قرار گرفته است. مجموعه‌داده‌ی آزمون شامل ۲۲۰۰ تصویر است که بدون برچسب در اختیار شما قرار می گیرد. **برای جلوگیری از تقلب ۶ ساعت قبل از ددلاین مجموعه داده آزمون در اختیار شما قرار می‌گیرد**. پس از ددلاین نهایی پروژه، دقت مدل شما با استفاده از نتیجه‌ی به‌دست‌آمده روی این مجموعه‌داده سنجیده می‌شود. توضیحات تکمیلی در بخش ارزیابی نهایی آمده است.

بخش ۱) آماده‌سازی داده‌ها

در این بخش، شما باید ابتدا تصاویر را بارگیری کنید[1]. پس از دریافت فایل‌ها می‌توانید با استفاده از توابع مخصوصی که در کتابخانه‌های حوزه‌ی یادگیری عمیق همچون کراس (Keras) یا پای‌تورچ (PyTorch) فراهم شده آن‌ها را بخوانید. سپس این تصاویر را به دو دسته‌ی آموزشی و اعتبارسنجی تقسیم کنید تا بتوانید مدل‌های خود را با داده‌های آموزشی، آموزش داده و طبق داده‌های اعتبارسنجی ارزیابی کنید. دقت داشته باشید یکی از هایپرپارامترهای مهم که براساس سیستم و توان پردازش شما باید به‌صورت بهینه انتخاب شود، اندازه‌ی دسته‌هایی (batch size) است که به مدل می‌دهید.

بخش ۲) تعیین مدل

در این بخش شما مجاز هستید از هر مدل یا معماری دل‌خواهی استفاده کنید. می‌توانید از مدل‌های یادگیری عمیق آماده مانند MobileNet، Inception، VGG، ResNet، AlexNet و سایر معماری‌های معروف استفاده کنید و از وزن‌های پیش‌آموخته‌ی (pre-trained) آن‌ها نیز بهره ببرید. هم‌چنین، می‌توانید مدل‌های دل‌خواه خود را نیز طراحی و پیاده‌سازی کنید. انتخاب معماری مدل وابسته به خصوصیت‌های داده‌ها و پیچیدگی مسئله است.

به‌طور معمول، برای این مسائل، از مدل‌های یادگیری عمیق پیش‌آموخته بر روی مجموعه‌داده‌های بزرگ مانند ImageNet بهره می‌برند و آن‌ها را به شکل کاملاً فریز شده یا با آموزش برخی از لایه‌های آن به‌عنوان پایه‌ی مدل خود به کار برده و در انتها لایه‌های ویژه‌ی مسئله‌ی خود را قرار می‌دهند. این امر به افزایش سرعت و دقت آموزش کمک می‌کند، زیرا این مدل‌ها ویژگی‌های بسیار خوبی از تصاویر را یاد گرفته‌اند.

همچنین، تنظیم هایپرپارامترها مانند نرخ یادگیری، سایز و عمق شبکه و تعداد دورهای آموزش مدل (epochs) نیز از اهمیت بالایی برخوردار است. برای هر مدل می‌بایست این هایپرپارامترها را تا جای ممکن بهتر انتخاب کرد تا به حداکثر دقت و کارایی برسید.

در این بخش، شما می‌توانید مدل‌های مختلف را امتحان کنید و با ارزیابی دقیق نتایج، مدل مناسب برای مسئله خود را انتخاب کنید.

توجه مهم : فراموش نکنید مدل خود را ذخیره نمایید تا بتوانید دوباره از آن برای پیش‌بینی مجموعه داده آزمون استفاده کنید.

[1] در صورتی‌که از گوگل کولب استفاده می‌کنید می‌توانید به کمک کتابخانه‌ی gdown، فایل‌ها را به‌صورت مستقیم از گوگل درایو دریافت کنید.

نکات کلی

- پیشنهاد می‌شود در صورت احتمال وقوع بیش‌برازش از تکنیک‌های تقویت داده (Data Augmentation) مختلف که برای داده‌های تصویری وجود دارد استفاده کنید.
- در صورتی که قصد دارید از یک مدل معروف با وزن‌های پیش‌آموخته استفاده کنید دقت کنید که نیاز است از تابع پیش‌پردازش ویژه‌ی آن مدل بهره ببرید.
- سعی کنید در ابتدای کار خود برخی از تصاویر موجود در مجموعه‌داده را به همراه برچسب آن‌ها رسم کنید. همچنین بعد از مدل‌سازی نیز برخی از تصاویر اعتبارسنجی را به‌همراه برچسب حقیقی و برچسب پیش‌بینی‌شده نمایش دهید تا عملکرد مدل شما به شکل شهودی‌تری مشخص شود.
- ابعاد عکس‌ها یکسان نیستند.
- برچسب‌های دادگان آموزش، توسط نیروی انسانی انجام شده‌است. به همین دلیل، شاید تعدادی از عکس‌های هر نوع غذا، به اشتباه برچسب خورده باشند. مدیریت این مسئله، جزوی از چالش این سوال و بر عهده شما می‌باشد.

ارزیابی نهایی

مجموعه‌ی آزمونی که در اختیار شما قرار می‌گیرد 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 q1_submission.csv باشد و شامل دو ستون نام فایل تصویر (name) و دسته‌بندی پیش‌بینی‌شده (predicted) باشد. به نمونه‌ی زیر دقت کنید:

name	predicted
a8d5d37ad16a61a000428187a8b7e44ca3a58c33_1609252208.jpg	baked_potato
2c4a86d25d413379ce9b58bcfc91f71607821919_1628724399.jpg	baklava
04aa0a23a9b842b6b15b9e3145555d5489b84483_1630354184.jpg	caesar_salad

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)

موفق باشید 🤩