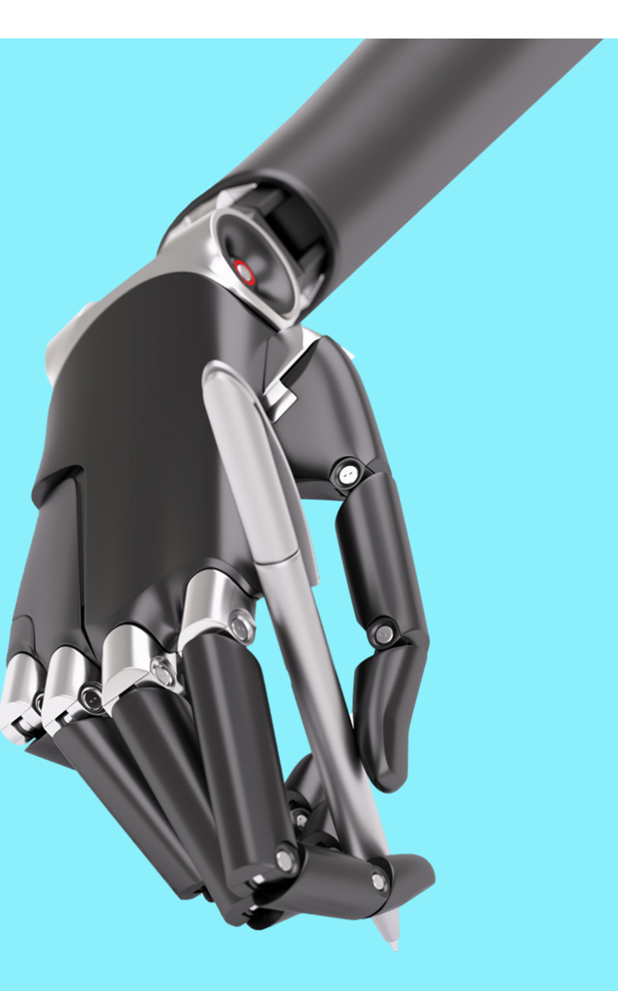
چالش امتیازی ورود به بوت کمپ
ماشین لرنینگ

MACHINE LEARNING BOOTCAMP



علاوه بر سؤال اصلی که پاسخ به آن برای ورود به بوتکمپ ضروری است، در این فایل چند سؤال امتیازی هم برایتان آماده کرده‌ایم.

می‌توانید یکی از این سؤال‌ها را انتخاب کنید و پاسخ‌تان را در همان فایل چالش اصلی برای ما بنویسید.

جواب دادن به این سؤال‌ها اجباری نیست، اما می‌تواند شانس شما را برای حضور در بوتکمپ بیشتر کند .

چالش امتیازی دوره:

یکی از سوالات زیر را به انتخاب خود انجام داده، برای سوال خود یک مجموعه داده مرتبط پیدا و مورد استفاده قرار دهید. اهمیت در نوع نگاه و توانمندی حل مساله و مسیری که برای حل مساله پیش می‌برید است. تمرکز خود را به دقت مدل هایی که می‌سازید یا مواردی از این دست محدود نکنید. هر سوال شامل چند بخش خواهد بود :

- بخش مفهومی (دانش نظری و تعاریف)
- بخش پیاده‌سازی (کدنویسی پایتون با مستندات و Comments)
- تسک تحلیلی ساده (تمرین دستی یا توضیح تحلیلی)
- مصورسازی (ماتریس، بارپلات، نمودار خطی)

سؤال اول

حوزه: NLP – تحلیل احساسات (Sentiment Analysis)

سناریو:

شما به‌عنوان کارشناس یادگیری ماشین در یک فروشگاه آنلاین بزرگ استخدام شده‌اید. هدف، طبقه‌بندی خودکار نظرات متنی کاربران به سه دسته‌ی «مثبت»، «منفی» و «خنثی» است تا مدیریت بتواند نقاط قوت و ضعف محصولات را شناسایی و بهبود دهد.

بخش ۱: پرسش‌های مفهومی

۱) پیش‌پردازش متن:

- چرا حذف علائم نگارشی و «استمینگ» یا «لِماتایزینگ» مهم است؟
- چه تأثیری دارد اگر کلمات «خوب» و «خوبی» را به صورت جداگانه نگه داریم؟

۲) معیارهای ارزیابی:

- دقت (Accuracy) در چه شرایطی گمراه‌کننده است؟ مثال بزنید.
- Precision, Recall و F۱– score را تعریف کرده و کاربرد هرکدام را در این مسأله توضیح دهید.

۳) تحلیل خطا:

- چگونه می‌توان خطاهای مدل را تحلیل کرد تا الگوهای اشتباه (مثلاً تمایز بین «خنثی» و «مثبت») مشخص شود؟
- ایده‌ای برای بهبود مدل بر مبنای نتایج تحلیل خطا بدهید.

بخش ۲: پیاده‌سازی کدنویسی

۱) پاک‌سازی و پیش‌پردازش:

- حذف کاراکترهای غیر حرفی، تبدیل حروف به کوچک، حذف توقف‌کلمات (stop words).

۲) بردارسازی (Vectorization):

- دو روش مختلف (TF–IDF و Word Embedding) را پیاده‌سازی کرده و مقایسه کنید.

۳) مدلسازی:

- الف) یک کلاسیک (مثلاً Logistic Regression یا SVM)
- ب) یک شبکه ساده‌ی LSTM

۴) ارزیابی و مصورسازی:

- رسم ماتریس درهم‌ریختگی برای هر مدل
- نمودار توزیع برچسب‌ها (مثبت/منفی/خنثی)
  مقایسه دقت مدل‌ها در یک نمودار میله‌ای

تسک تحلیلی ساده:

- با استفاده از چند مثال کوتاه (حداقل ۵ نظر)، مدل کلاسیک و LSTM را مقایسه و یک تحلیل دستی (به‌صورت جدولی در مستندات) از تفاوت پیش‌بینی‌شان ارائه کنید.

سؤال دوم

حوزه: Search – رتبه‌بندی نتایج بر اساس ارتباط معنایی

سناریو:

شما محقق داده در تیم توسعهٔ موتور جستجوی یک شرکت دانش‌بنیان هستید. دیتاست شامل پرسش‌های واقعی کاربران و فهرستی از پاسخ‌های پیشنهادی است که برچسب «مرتبط» یا «نامرتبط» دارند.

بخش ۱: پرسش‌های مفهومی

۱) بردارسازی سؤال و پاسخ:

- تفاوت TF–IDF با بردارهای مبتنی بر BERT چیست؟ در چه شرایطی BERT برتری دارد؟

۲) روش امتیازدهی (Scoring):

- تعریف معیار Cosine Similarity و کاربرد آن در سنجش ارتباط معنایی
- مثال عددی ساده (۲ سؤال و ۲ پاسخ) و محاسبهٔ دستی cosine similarity

۳) رتبه‌بندی:

- چگونه می‌توان چند پاسخ «مرتبط» و «نامرتبط» را براساس امتیاز مرتب کرد؟
- مفهوم Precision@k و Recall@k را توضیح دهید و نحوهٔ محاسبهٔ آن‌ها را برای k=۵ شرح دهید.

۴) بهینه‌سازی:

- اگر مدل شما «پاسخ‌های کم‌اهمیت ولی طولانی» را بیش‌ازحد مرتبط تشخیص می‌دهد، چه راهکاری پیشنهاد می‌کنید؟

بخش ۲: پیاده‌سازی کدنویسی

۱) بارگذاری و پیش‌پردازش داده‌ها:

- حذف HTML، توکن‌سازی، لِمتایزینگ

۲) بردارسازی:

- پیاده‌سازی TF–IDF
- استخراج embedding با یک مدل پیش‌آموزش‌دیده BERT

۳) محاسبه امتیاز و رتبه‌بندی:

- محاسبه Cosine Similarity
- رتبه‌بندی پاسخ‌ها برای هر سوال

۴) مصورسازی:

- نمایش یک مثال واقعی: پرسش + پاسخ‌ها + امتیازها
- ترسیم barplot امتیازها یا heatmap ماتریس امتیاز

تسک تحلیلی ساده:

- برای یک سوال نمونه، سه پاسخ اول و سوم مدل را با هم مقایسه کنید و با دلایل تحلیلی توضیح دهید چرا مدل آن‌ها را در رتبه‌های مذکور قرار داده است.

سؤال سوم

حوزه: سامانه‌های توصیه‌گر (Recommender Systems)

سناریو:

یک سرویس ویدیویی آنلاین می‌خواهد به کاربران خود توصیه‌هایی شخصی‌سازی‌شده ارائه دهد (فیلم، سریال، مستند). شما عضو تیم داده و مسئول طراحی و ارزیابی الگوریتم‌های توصیه‌گر هستید.

بخش ۱: پرسش‌های مفهومی

۱) انواع روش‌ها:

- تفاوت Collaborative Filtering (افقی/عمودی)، Content–Based و روش‌های ترکیبی (Hybrid) را توضیح دهید.

۲) معیارهای ارزیابی:

- Precision@K, Recall@K, MAP (Mean Average Precision)، NDCG (Normalized Discounted Cumulative Gain) را تعریف و کاربرد هرکدام را شرح دهید.

۳) مشکلات رایج:

- مشکل Cold–Start (محتوا و کاربر جدید) چیست و چگونه می‌توان آن را حل کرد؟
- اثر پدیده Popularity Bias در توصیه‌گرها چیست و چه راه‌هایی برای کاهش آن پیشنهاد می‌شود؟

بخش ۲: پیاده‌سازی کدنویسی

۱) آماده‌سازی داده:

- ماتریس کاربر–آیتم با مقادیر امتیاز یا تعداد بازدید

۲) مدل‌های ساده:

- الف) User–Item Collaborative Filtering با ماتریس فاکتورگیری (Matrix Factorization)
- ب) Content–Based با نمایش ویژگی‌های فیلم (ژانر، سال تولید، کارگردان)

۳) تولید توصیه:

- کدنویسی تابعی که برای یک کاربر مشخص لیست top–K فیلم‌ها را تولید کند.

۴) ارزیابی و مصورسازی:

- محاسبهٔ Precision@۱۰ و Recall@۱۰ روی دادهٔ آزمایشی
- ترسیم نمودار خطی (Precision و Recall بر حسب K)

تسک تحلیلی ساده:

- برای یک کاربر سرد (با فقط ۲ فیلم امتیازدهی‌شده)، پیشنهاد دهید از چه روشی (Collaborative یا Content–Based) استفاده شود و چرا؟

رهنما کالج
با حرفه‌ای‌ها بگرد!