کدام آمار؟

تعدادی چکیده‌ی مقاله از دو شاخه‌ی مختلف «آمار در یادگیری ماشین» و «آمار کاربردی» به همراه برچسب شاخه‌ی آن‌ها در اختیار شما قرار گرفته است. شما باید یک دسته‌بند ساده آموزش دهید که بتواند با ورودی گرفتن چکیده، موضوع مقاله را پیش‌بینی کند.

شما می‌توانید از هر کتابخانه پایتونی برای حل این سوال استفاده کنید. دقت کنید که کد نفرات برتر مورد بررسی قرار خواهد گرفت.

مجموعه‌ی‌داده

می‌توانید مجموعه‌داده‌ی مربوط به این مسئله را از این لینک دانلود کنید.

هنگامی که این فایل را از حالت فشرده خارج کنید، دو فایل `train.csv` و `test.csv` در اختیار شما قرار می‌گیرد. فایل آموزش شامل دو ستون به شرح زیر است:

نام ستون	توضیحات ستون
`abstract`	چکیده
`category`	برچسب شاخه‌ی مقاله که شامل یکی از دو مقدار `ML` یا `Applied` می‌باشد

فایل آزمون (داده‌های آزمایش) تنها شامل ستون `abstract` است.

صورت مسئله

از فایل `train.csv` برای پیش‌بینی موضوع مقاله با استفاده از چکیده آن و آموزش مدل استفاده کنید و از فایل `test.csv` برای آزمایش مدل شما در سیستم داوری استفاده می‌شود.

ارزیابی

ارزیابی عملکرد بر اساس دقت (accuracy) بر روی داده‌های آزمایش خواهد بود؛ یعنی تعداد نمونه‌های درست دسته‌بندی شده تقسیم بر تعداد کل نمونه‌ها می‌شود. در نهایت امتیاز شما از این سوال طبق رابطه‌ی زیر محاسبه می‌شود:

$$score = \begin{cases} 0 & accuracy < 0.6 \\ accuracy \times 100 & accuracy \geq 0.6 \end{cases}.$$

داوری این سوال قبل از پایان مسابقه، تنها بر اساس ۳۰ درصد از مجموعه داده آزمایش (test) خواهد بود. پس از اتمام مسابقه، برای به‌روزرسانی نهایی جدول امتیازات، از ۱۰۰ درصد مجموعه داده آزمایش استفاده خواهد شد؛ این کار برای جلوگیری از بیش‌برازش (overfit) روی مجموعه داده آزمایش انجام می‌شود.

خروجی

پیش‌بینی‌های مدل خود بر روی دادگان آزمایش (test.csv) را در فایلی با نام output.csv قرار دهید.

این فایل باید شامل یک ستون prediction باشد. در سطر i ام از این ستون باید پیش‌بینی مدل شما روی داده‌ی با اندیس i باشد. بعد از آماده‌سازی فایل output.csv ، آن را برای ما بارگذاری کنید.

نمونه خروجی فایل output.csv (فقط پنج خط اول به همراه نام ستون)

prediction
ML
Applied
Applied
ML
ML

▼ توجه

با توجه به تعداد بسیار کم دادگان آموزش، پیشنهاد می‌شود که از مدل‌های عمیق استفاده نکنید.

استفاده از وزن مدل‌های از پیش آموزش دیده (pretrained) برای تسهیل آموزش مدل خود، در سوالات مانعی ندارد.

▼ هشدار 😱

فراموش نکنید که **قبل از پایان زمان مسابقه، بایستی** تمامی کدهای این مسابقه را از قسمت **بارگذاری کُد** برای ما ارسال کنید. در غیر این صورت، شما از این مسابقه، امتیازی کسب نمی کنید.

توجه داشته باشید که اگر از `jupter notebook` استفاده می کنید بایستی همانند توضیحات قسمت **بارگذاری کُد**، خروجی `.py` را دریافت و برای ارسال در نظر بگیرید. ارسال فایل‌های `jupyter` همانند `.ipynb` مورد قبول واقع نخواهند شد.