اخبار ایران و جهان

در این سوال ما یک مجموعه داده‌ی خبری داریم که دارای ۲۳ زیر مجموعه خبری است. شما باید مدلی بسازید که از این مجموعه داده استفاده کند و برای تخمین زیر مجموعه خبر مورد استفاده قرار گیرد.

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

مجموعه داده

مجموعه داده سوال را می‌توانید از این یا این لینک دانلود کنید.

هنگامی که این فایل را از حالت فشرده خارج کنید فایل آموزش (`train.csv`) و آزمایش (`test.csv`) را مشاهده می‌کنید. فایل آموزش، دارای ساختار زیر است:

نام ستون	توضیحات ستون
title	عنوان خبر
subgroup	زیرگروه خبر
abstract	خلاصه خبر
body	مشروح خبر

تنها تفاوت مجموعه داده آموزش با آزمایش در این است که مجموعه داده آزمایش، ستون `subgroup` را ندارند.

صورت مسئله

با استفاده از مجموعه داده آموزش، یک مُدل برای پیشبینی زیرگروه خبر (ستون `subgroup`) هر سطر آموزش دهید.

ارزیابی

برای ارزیابی پاسخ شما از معیار *F1* استفاده خواهد شد. این معیار به صورت زیر تعریف می‌شود:

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall}$$

ولی به دلیل اینکه ستون پیشبینی دارای بیش از ۲ کلاس است ما از معیاری به نام *Weighted F1* استفاده میکنیم که *F1* میانگین وزنی کلاس ها را محاسبه میکند.

برای مطالعه بیشتر در مورد این معیار می‌توانید به این منبع مراجعه کنید.

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 دارای یک ستون به اسم `subgroup` باشد.(بزرگ و کوچک بودن حروف نام ستون رعایت شود) که ردیف i ام هر ستون، پیش‌بینی شما برای نظر ردیف i ام از فایل `test.csv` باشد. بعد از آماده‌سازی فایل `output.csv` ، آن را برای ما بارگذاری کنید.

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

subgroup
اجتماعی
اجتماعی
سیاسی
اقتصادی

▼ **توجه**

حتما فایل `output.csv` باید دارای ۱۰۱۹۵۰ سطر (بدون در نظر گرفتن `header`) و یک ستون باشد.

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.