دسته‌بندی

▼ **برای این سوال امکان استفاده از huggingface را ندارید.**

در این سوال شما باید اقدام به ساخت مدلی کنید که موضوع (برچسب) خبرهارا پیش‌بینی کند.

مجموعه داده

مجموعه داده مورد نیاز برای این سوال را از این لینک دانلود کنید مجموعه داده آموزش دارای سه ستون زیر می‌باشد:

نام ستون	توضیحات
title	عنوان خبر
description	متن خبر
tags	موضوع خبر

هر سطر داخل مجموعه داده مربوط به یک خبر است و موضوع هر خبر نیز در ستون *tags* قرار داده شده است. این ستون دارای موضوعات مختلفی است اما پیش‌بینی مدل شما برای مجموعه داده آزمایش (*test*) باید یکی از موضوعات زیر باشد:

- اجتماعی
- اقتصادی
- ایران_استانها
- بین الملل
- سیاسی
- علمی_فرهنگی_ورزشی

در نتیجه ممکن است ستون هدف (*tags*) در مجموعه داده اولیه نیازمند تغییرات باشد. توجه داشته باشید که در این مرحله از مسابقه مجاز به برچسب گذاری اخبار حداکثر با یک دسته اصلی می‌باشید.

به عنوان مثال، اخبار مربوط به کرونا با توجه به موضوع، می توانند در دسته‌های زیر قرار بگیرند:

- اجتماعی: قرنطینه‌ها در کشورهای مختلف

- اقتصادی: مسائل مربوط به مشاغل کوچک و جبران خسارت برای کارآفرینان
- علمی_فرهنگی_ورزشی: المپیک توکیو به دلیل همه گیری جابه‌جا شد..
- علمی_فرهنگی_ورزشی: اخبار مربوط به علائم، نکاتی درباره سالم ماندن، جستجوی واکسن‌ها و غیره

صورت مسئله

مدل شما باید برای هر خبر (سطر) از مجموعه داده آزمایش، پیش‌بینی کند که محتوای آن خبر در کدام یک از ۶ دسته اصلی مذکور قرار میگیرد. در مرحله بعد باید بر اساس جدول زیر، رشته‌های بدست آمده را encode کرده و جواب نهایی را بر اساس مقادیر عددی ارسال کنید

موضوع	مقدار عددی
اجتماعی	0
اقتصادی	1
ایران_استانها	2
بین الملل	3
سیاسی	4
علمی_فرهنگی_ورزشی	5

ارزیابی

برای ارزیابی مدل شما از معیار `F1 Score` استفاده می‌شود و مدل میانگین‌گیری نیز به صورت `Weighted` است.

▼ توجه

در طول مسابقه امتیازی که مشاهده می‌کنید، فقط نتیجه‌ی `F1 Score` روی ۳۰ درصد از فایلی است که برای کوئرا آپلود می‌کنید. بعد از پایان زمان مسابقه، **امتیاز نهایی** شما روی ۷۰ درصد مابقی محاسبه می‌شود.

این کار به منظور جلوگیری از `overfitting` و حفظ عمومیت مدل انجام می‌شود تا مطمئن شویم مدل‌هایی که دچار بیش‌برازش شده‌اند، در امتیازهی نهایی، افت می‌کنند.

خروجی

پیش‌بینی‌های مدل خود بر روی مجموعه داده آزمایش (`test_data.csv`) را در فایلی با نام `submission.csv` قرار دهید. این فایل باید دارای یک ستون با نام `prediction` باشد که ردیف i ام آن، دسته پیش‌بینی شده برای خبر ردیف i ام از مجموعه داده آزمایش باشد (دقت کنید که ستون باید حتما دارای `header` باشد). بعد از آماده‌سازی فایل `submission.csv` ، آن را برای ما بارگذاری کنید.

prediction
0
1
2

prediction
3
4
5

▼ **نکات مهم در مورد فایل ارسالی**

- **توجه ۱**: توجه کنید که ستون گفته شده حتما دارای `header` باشد.
- **توجه ۲**: مراقب باشید در فایل نهایی اندیس ذخیره نشود و فقط یک ستون `prediction` باشد.

▼ **هشدار**

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

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