# 新闻分类挑战赛规则

## 任务说明

针对软硬新闻分类任务，利用给定的MIND和Google新闻数据集进行文本分类器设计和训练，并在给定的测试数据集（由教师保存）上进行评测。

* 分类依据：仅可使用原始新闻标题，不包含新闻内容；
* 分类结果：硬新闻（标签0）、软新闻（标签1）、其他（标签2）；
* 测试数据集格式：文本文件（文件名：test.txt），每行一个新闻标题的原始文本；
* 测试结果输出格式：文本文件（文件名：result\_姓名.txt），每行一个分类结果标签（0，1或2）对应输入行的新闻分类结果；

软硬新闻的标准，由两个训练数据集说明给出。一般的，时政、军事、财经类的严肃内容为硬新闻；娱乐、体育类的内容为软新闻。

## Microsoft MIND数据集说明

针对本次分类任务，仅需使用其中的news.tsv文件中的新闻相关数据，具体文件格式参考网页<https://github.com/msnews/msnews.github.io/blob/master/assets/doc/introduction.md>。

该数据集包括约10万新闻，提供了18个主分类和308个子分类，具体分类和软硬新闻对照关系参看分类系统的Excel文件。

注意，该数据集中，某些分类包含的新闻数量很少，可以忽略不记。

## Google News数据集说明

Google News数据与MIND基本格式相同（只包含id，分类和标题信息）。

该数据集包括约11万新闻，提供了8个主分类和42个子分类，具体分类和软硬新闻对照关系参看分类系统的Excel文件。

## 评价标准

本次比赛的成绩由以下4部分构成：

1. 基础分：独立完成本次挑战赛，可获得50%的分数；
2. 演讲分：讲解算法设计，至少提供一张算法结构图和相关程序的训练、使用说明，占比20%；
3. 排名分：在给定测试数据集上，算法表现按F1-Score排名，排名越高分数越高，第一名满分，最后一名0分，分数按幂律递减，占比25%；
4. 时间分：4次答辩机会，4月24日和5月8、15、22日，越靠前分数越高，占比5%。