**实验8文本分类预测实验**

**1. 实验数据**

(1) **训练集**

训练集包含1267条新闻，存放于文件'news\_train.csv'中。每一行代表一条新闻，第一列(ID)是新闻的标识，第二列(title)和第三列(text)是新闻标题和新闻内容，第四列的标签(label)表示该条新闻是真实(REAL)的还是虚假(FAKE)的。训练集使用pandas读入后，部分截图如下：

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

(2) **测试集**

另有254条未知新闻类别标签的测试数据，保存在'news\_test.csv'文件中。每一行仅给定新闻ID、title和text，而每条新闻的类别 (REAL或FAKE)未知，待建模识别。

表格

中度可信度描述已自动生成

**2. 实验目的**

(1) 设计文本的特征向量，具备一定的特征工程能力。

(2) 利用机器学习分类算法，基于训练集构建分类器模型；

(3) 进而将构建好的分类器模型应用于测试集，给出全体未知标签样本的分类结果，即预测文本数据所属的类别(REAL或者 FAKE)。

**3. 设计思路**

(1) 可使用pandas.read\_csv(FileName)读取文件，并返回DataFrame进行后续处理

(2) 需从每条文本中提取特征，表示为固定长度的特征向量。特征向量设计思路不限，例如，可尝试使用tf-idf等常见文本特征提取算法。

*Tf-idf参考网址：https://scikit-learn.org/stable/modules/generated/sklearn.feature\_extraction.text.TfidfVectorizer.html*

(3) 具体的机器学习算法不限，以预测效果最佳为目标，追求F1 score越高越好；可以尝试多种学习模型的集成。(在计算F1 score时，指定pos\_label='REAL')

**4. 实验要求**

(1) 将预测结果保存在名为“pred.txt”的文本文件中，内容为254行， 每一行只有FAKE或者REAL，代表你的算法对测试数据的预测结果。预测数据顺序须与测试集“news\_test.csv”中的样本顺序保持一致。

(2) 将结果文件“pred.txt” 以附件形式提交至学习通。另请注意：只需提交结果文件“pred.txt”，无需提交本次实验报告文件。

(3) 本次实验成绩评定采用竞赛机制，即计算每位同学预测结果的F1 score，然后由高到低进行排序评定相应的实验成绩。