**实验8文本分类预测实验**

**1. 实验数据**

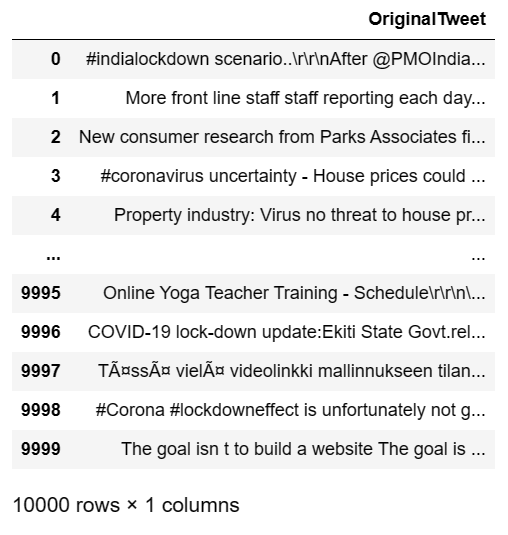
(1) **训练集**

训练集包含20000条关于新冠疫情的推文，存放于文件'sentiment\_train.csv'中。每一行代表一条推文，第一列是具体的推文（OriginalTweet），第二列是推文的情感标注，表示该条推文代表对当前内容非常正面（Extremely Positive），正面（Positive），中性（Neutral），负面（Negative）还是非常负面（Extremely Negative）的评价。训练数据集使用pandas读入后(pd.read\_csv('sentiment\_train.csv'))，截图如下：

****

1. **测试集**

另有10000条未知情感标签的测试数据，保存在'sentiment\_test.csv'文件中。每一行仅给定推文的文本（pd.read\_csv('sentiment\_test.csv')，而每条推文的情感标注(Extremely Positive， Positive，Neutral, Extremely Negative或Negative)未知，待建模识别。

****

**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(Extremely Positive， Positive，Neutral, Extremely Negative或Negative)。

**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（sklearn.metrics.accuracy\_score）；体会多分类模型的评价和2分类评价的不同， 思考如何处理不平衡的数据集（imbalanced dataset）。可以尝试多种学习模型的集成。

**注意：在调试模型的过程中，可以分出训练集中的80%用作训练，用剩下的20% 验证模型的好坏。不要使用现成的一些情感分类模型如nltk 中的 SentimentIntensityAnalyzer， 效果会非常差，本实验中的数据集是针对特定主题的情感状况进行判断，不同于一般的文本情感分析。**

**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10000行， 每一行只有Extremely Positive， Positive，Neutral, Extremely Negative或Negative，代表你的算法对测试数据的预测结果（**注意大小写**）。预测数据顺序须与测试集“sentiment\_test.csv”中的样本顺序保持一致。

(2) **将结果文件“pred.txt”，以及源代码**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”和源代码**，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accuracy的分数高低，然后由高到低进行排序评定相应的实验成绩。