一、题目名称

疫情期间网民情绪识别

二、题目背景

2019 新型冠状病毒（COVID-19）感染的肺炎疫情发生对人们生活生产的方方面面产生了重要影响，并引发国内舆论的广泛关注，众多网民参与疫情相关话题的讨论。为了帮助政府掌握真实社会舆论情况，科学高效地做好防控宣传和舆情引导工作，本题目针对疫情相关话题开展网民情绪识别的任务。

三、数据集

数据集依据与“新冠肺炎”相关的 230 个主题关键词进行数据采集，抓取了 2020 年 1 月 1 日—2020 年 2 月 20 日期间微博数据，并对其进行人工标注，标注分为三类，分别为：1（积极），0（中性）和-1（消极）。

训练数据以 csv 格式存储在 train.csv 文件中，其中包含 45000 条微博数据，具体格式如下：

[微博中文内容，情感倾向]

1.微博中文内容，格式为字符串

2.情感倾向，取值为{1,0,-1}

四、任务描述

根据 train.csv 文件中的微博数据，设计算法对 test.csv 文件中的 4500 条微博内容进行情绪识别，判断微博内容是积极的（1）、消极的（-1）还是中性的（0）。

将结果存储在 csv 文件中，编码采用 UTF-8 编码，格式如下：

微博中文内容　　情感倾向

新冠肺炎……　　1

五、评测标准

基于以下混淆矩阵(confusion matrix)，采用 Precision，Recall，F1-score 三个指标评价算法结果，要对比 3 种以上算法的结果，可进一步自由发挥，做算法参数敏感性的实验及对比分析等。

Confusion matrix		真实值	
		positive	negative
预测值	positive	TP	FP
	negative	FN	TN

其中，TP 是真阳例，TN 是真阴例，FP 是假阳例，FN 是假阴例。

1. Precision

精确率(查准率)，即为在预测为 1 的样本中，预测正确（实际为 1）的人占比，,用混淆矩阵中的字母可表示为：

$$Precision = TP/(TP+FP)$$

2. Recall

召回率（查全率），即为在实际为 1 的样本中，预测为 1 的样本占比，用混淆矩阵中的字母可表示为：

$$Recall = TP/(TP+FN)$$

3. F1-score

F1 分数（F1 Score），是统计学中用来衡量二分类模型精确度的一种指标。它同时兼顾了分类模型的准确率和召回率。F1 分数可以看作是模型准确率和召回率的一种加权平均，它的最大值是 1，最小值是 0。

$$F1=2*Precision * Recall /(Precision + Recall)$$