关卡 5：商品评论情感分析

华为技术有限公司

课堂思考题

课堂思考题（20 分）

1. 混淆矩阵如下所示，在二分类问题中，准确率（accuracy）如何计算？请列出计算公式。（10 分）

真实情况	预测结果		合计
	正例	反例	
正例	TP（真正例）	FN（假反例）	P
反例	FP（假正例）	TN（真反例）	N

分类准确率（accuracy）=划分正确的样本数 / 所有的样本数

$$accuracy = \frac{TP + TN}{P + N}$$

2. （简答）在垃圾邮件分类问题和流感模型预测这两种二分类的场景情况下，分场景说明 precision 和 recall 这两个指标哪一个更重要？简要说明（10 分）

$precision = \frac{TP}{TP + FP}$ 精确率表示的是在所有预测结果为真中的准确率，所以可以看出，precision 关注的是模型在预测结果为真中的准确率，表示的是模型在预测结果为真时的可信度，而不关心模型能从真正的正例中识别出多少正例。

对于垃圾邮件分类问题，需要在所有邮件中对垃圾邮件进行分类和标注，而一封邮件被误判为垃圾邮件可能导致错过很多重要的通知，因此，相对于从大量邮件中识别出垃圾邮件的数量也就是收回度而言，我们更加关注判断垃圾邮件为真时的可信度，所以查准率 precision 更重要。

$recall = \frac{TP}{TP + FP}$ 召回率表示的是在 Label 中的所有正类中被预测为正类的比例。衡量的是模型对实际正类的提取能力，Recall 关注的是模型在所有真正正类中的准确率，反应了一个模型能识别出正类的能力。

对于流感模型预测，需要从现有的数据和模型中检测出流感的感染人群，更加关注模型预测的识别能力，强调不能放过任何一个可能得流感的人，相对于错误判断的时间和金钱等成本来说，放过一个可能的患者带来的损失是更加巨大的，因此查全率 recall 比查准率 precision 更重要。

实验

模型保存（20 分）

截图内容：调用模型保存的接口，并将对应的保存的模型文件截图，模型文件名规范：textCNN_(日期信息)。（20 分）

```
#模型的保存
self.model.save('textCNN_20200716.h5')
```

创新题（40 分）

1. 实现从 DWS 当中获取评论数据来完成本关卡实验数据收集。提示：当前 DWS 中并未创建评论相关表格，也未导入评论数据，需要实现从 RDS for MySQL 到 DWS 的数据迁移。（40 分）

提交内容包含但不限于：1.Kettle 当中的“步骤度量”信息；2.python 读取 DWS 数据的相关代码。

执行历史 | 日志 | 步骤度量 | 性能图 | Metrics | Preview data

#	步骤名称	复制的记录行数	读	写	输入	输出	更新	拒绝	错误	激活	时间	速度 (条记录/秒)	Pri/in/out
1	litemall_comment	0	0	27613	27613	0	0	0	0	已完成	14.5s	1,906	-
2	original_comment	0	27613	27613	0	27613	0	0	0	已完成	23.1s	1,193	-

```
import psycopg2
import pymysql
import numpy as np
import pandas as pd
import re
import datetime

def insert_table(sql,ip,port,database,user, password):
    '''
    连接 DWS，操作相应的 sql 语句来进行查询
    '''
```

```
        connection = psycopg2.connect(host=ip, port=port, database=database, user=user,
password=password)
        connection.set_client_encoding('utf-8') #  把编码格式换成  utf8，以防止出现乱码

        #'''创建游标操作：'''
        cursor = connection.cursor()
        cursor.execute(sql) #执行  sql  命令

        connection.commit()
        cursor.close() #关闭游标对象
        connection.close() #关闭数据库连接

sql = '''DROP TABLE IF EXISTS original.original_comment;
CREATE TABLE original_comment (
    id int NOT NULL,
    value_id int NOT NULL,
    type tinyint NOT NULL,
    content varchar(1023) NOT NULL,
    admin_content varchar(511) NOT NULL,
    user_id int NOT NULL,
    has_picture tinyint,
    pic_urls varchar(1023),
    star smallint,
    add_time date,
    update_time date,
    deleted tinyint,
    PRIMARY KEY (id)
    );
'''
insert_table(sql,'114.116.200.18','8000','tsz_demo','dbadmin','Dws@123456')
```