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

答：准确率即为正确预测实例数/实例总数

计算公式为：（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）

答：

Precision 精确率指在预测为正例的结果中，正确的个数所占的比例，也就是预测正确的概率；recall 召回率指在所有的正样本中，预测正确的个数所占的比例，也就是正确查找正例的概率。

显然，垃圾邮件分类要求较高的正确分类概率，precision 指标更重要；流感模型预测要求对感染个体进行准确提取，要求更高的查找正例概率，recall 指标更重要哦。

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）

----------请在下方附上图片----------

```
def fit(self, x_train, y_train, x_valid=None, y_valid=None, epochs=5, batch_size=128, **kwargs):
# 训练
    self.build_model()
    x_train = self.preprocessor.transform(x_train)
    if x_valid is not None and y_valid is not None:
        x_valid = self.preprocessor.transform(x_valid)
    self.model.fit(
        x=x_train,
        y=y_train,
        validation_data= (x_valid, y_valid) if x_valid is not None and y_valid is not None else None,
        batch_size=batch_size,
        epochs=epochs,
        **kwargs
    )
    # from keras.models import load_model
    #模型的保存
    self.model.save('D://textCNN_0715.h5')
```

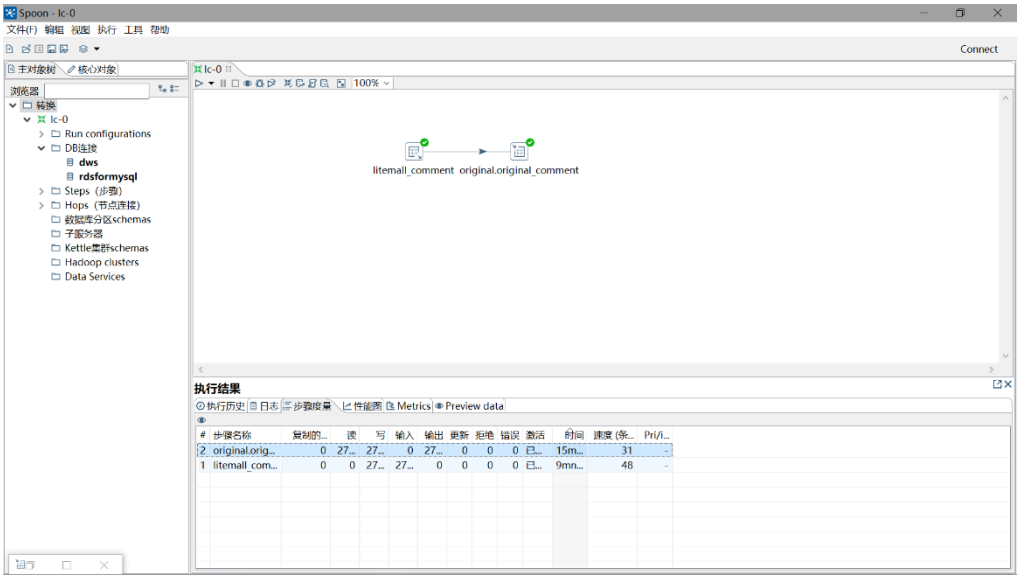
----------请在上方附上图片----------

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。

----------请在下方附上图片及代码----------




```
import psycopg2
import pandas as pd
import numpy as np
sql = '''SELECT user_id, contents
      FROM original.original_connect'''
db = psycopg2.connect(host="114.116.200.18", port="8000", database='zwq_demo',
user="dbadmin", password="Dws@123456")
db.set_client_encoding('utf-8') # 把编码格式换成 utf8，以防止出现乱码
cursor = db.cursor()
cursor.execute(sql)
rows = cursor.fetchall()
cursor.close()
db.close()
data = pd.DataFrame(list(rows), index = range(1, len(list(rows))+1))
data.columns = ['user_id', 'comment']
data.head()
```

----------请在上方附上图片及代码----------