大数据实验四

191098159 毛钰成

GitHub 仓库地址：https://github.com/Budian-mao/ex4

第一题

读取 industry 列的 string 数据

```
String line=value.toString();
String industry=line.split( regex: ",")[10];
if(!industry.equals("industry")) {
    word.set(line.split( regex: ",")[10]);
    context.write(word, one);
```

将出现的次数进行叠加

```
int sum = 0;
for (IntWritable val : values) {
    sum += val.get();
}
result.set(sum);
context.write(key, result);
```

运行结果：

```
[root@myc191098159-master workspace]# hadoop jar /workspace/-4/wordcount-1.0.jar /p1/dataset2/train_data.csv  /output
2021-12-18 08:06:50,317 INFO client.RMProxy: Connecting to ResourceManager at myc191098159-master/192.168.219.153:8032
Exception in thread "main" org.apache.hadoop.mapred.FileAlreadyExistsException: Output directory hdfs://myc191098159-mast
er:9000/output already exists
        at org.apache.hadoop.mapreduce.lib.output.FileOutputFormat.checkOutputSpecs(FileOutputFormat.java:164)
        at org.apache.hadoop.mapreduce.JobSubmitter.checkSpecs(JobSubmitter.java:277)
        at org.apache.hadoop.mapreduce.JobSubmitter.submitJobInternal(JobSubmitter.java:143)
        at org.apache.hadoop.mapreduce.Job$11.run(Job.java:1570)
        at org.apache.hadoop.mapreduce.Job$11.run(Job.java:1567)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:422)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1729)
        at org.apache.hadoop.mapreduce.Job.submit(Job.java:1567)
        at org.apache.hadoop.mapreduce.Job.waitForCompletion(Job.java:1588)
```

金融业 48216
电力、热力生产供应业 36048
公共服务、社会组织 30262
住宿和餐饮业 26954
文化和体育业 24211
信息传输、软件和信息技术服务业 24078
建筑业 20788
房地产业 17990
交通运输、仓储和邮政业 15028
采矿业 14793
农、林、牧、渔业 14758
国际组织 9118
批发和零售业 8892
制造业 8864

第二题

运行结果：

```
[root@myc191098159-master code]# spark-submit 2.py
21/12/18 16:21:00 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where appl
icable
21/12/18 16:21:01 INFO SparkContext: Running Spark version 3.0.2
21/12/18 16:21:01 INFO ResourceUtils: ==============================================================
21/12/18 16:21:01 INFO ResourceUtils: Resources for spark.driver:

21/12/18 16:21:01 INFO ResourceUtils: ==============================================================
21/12/18 16:21:01 INFO SparkContext: Submitted application: 2.py
21/12/18 16:21:01 INFO SecurityManager: Changing view acls to: root
21/12/18 16:21:01 INFO SecurityManager: Changing modify acls to: root
21/12/18 16:21:01 INFO SecurityManager: Changing view acls groups to:
21/12/18 16:21:01 INFO SecurityManager: Changing modify acls groups to:
21/12/18 16:21:01 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(ro
ot); groups with view permissions: Set(); users  with modify permissions: Set(root); groups with modify permissions: Set()
21/12/18 16:21:02 INFO Utils: Successfully started service 'sparkDriver' on port 42660.
21/12/18 16:21:02 INFO SparkEnv: Registering MapOutputTracker
21/12/18 16:21:02 INFO SparkEnv: Registering BlockManagerMaster
```

```
PROBLEMS   OUTPUT   TERMINAL   DEBUG CONSOLE                                   2: Python

21/12/19 00:05:00 INFO DAGScheduler: Job 0 finished: collect at /workspace/code/2.py:11, took 12.506608 s
[('(12000,13000)', 20513), ('(6000,7000)', 15961), ('(9000,10000)', 10458), ('(21000,22000)', 5507), ('(22000,23000)', 3544), ('(17000,1
8000)', 4388), ('(5000,6000)', 16514), ('(11000,12000)', 7472), ('(13000,14000)', 5928), ('(24000,25000)', 8660), ('(3000,4000)', 9317),
 ('(25000,26000)', 8813), ('(31000,32000)', 752), ('(26000,27000)', 1604), ('(32000,33000)', 1887), ('(30000,31000)', 6864), ('(19000,20
000)', 4077), ('(1000,2000)', 4043), ('(33000,34000)', 865), ('(34000,35000)', 587), ('(29000,30000)', 1144), ('(37000,38000)', 59), ('(
38000,39000)', 85), ('(8000,9000)', 16384), ('(20000,21000)', 17612), ('(10000,11000)', 27170), ('(28000,29000)', 5203), ('(4000,5000)',
 10071), ('(16000,17000)', 11277), ('(14000,15000)', 8888), ('(35000,36000)', 11427), ('(15000,16000)', 18612), ('(2000,3000)', 6341), (
'(7000,8000)', 12789), ('(18000,19000)', 9342), ('(23000,24000)', 2308), ('(27000,28000)', 1645), ('(40000,41000)', 1493), ('(39000,4000
0)', 30), ('(36000,37000)', 364), ('(0,1000)', 2)]
21/12/19 00:05:00 INFO AbstractConnector: Stopped Spark@667a21d4{HTTP/1.1, (http://1.1)}{0.0.0.0:4040}
21/12/19 00:05:00 INFO SparkUI: Stopped Spark web UI at http://myc191098159-master:4040
```

第二题出现问题：一开始没有过滤掉第一列，导致数据处理时遇到字符串 string 类型出错，后来发现读取 csv 文件的时候，没有过滤掉第一列，导致读入了 csv 第一列的 string 从而报错。

```python
sc = SparkContext()
lines = sc.textFile("train_data.csv").map(lambda x:map_func(x)).cache()
result = lines.reduceByKey(lambda x,y:x+y).collect()
print(result)
sc.stop()
```

然后用 filter 过滤掉第一列然后再运行的即可

```
sc=SparkContext()
lines = sc.textFile("train_data.csv").filter(lambda line: not line.startswith("loan_id")).ma
result = lines.reduceByKey(lambda x,y:x+y).collect()
print(result)
sc.stop()
```

第三题

```
#第三题第一问
total_num=df.count()
df1 = df.groupBy('employer_type').count().toPandas()
df1["count"]=df1["count"]/total_num
df1.to_csv("3-1.csv",index=0,header=0,float_format="%.4f")

#第三题第二问
df1=df.withColumn("total_money",df.year_of_loan*df.monthly_payment*12-df.total_loan).select("user_id","total_
```

```
#第三题第三问
from pyspark.sql.functions import udf
from pyspark.sql.types import IntegerType
#1. 将dataframe里面的work_year转化为int型
def cal_work_year(work_year):
    if work_year == None:
        return 0
    elif '<' in work_year:
        return 1
    else:
        year = work_year.split(' ')[0]
        year = year.split('+')[0]
        return int(year)


# 自定义函数
udf_cal_work_year = udf(cal_work_year, IntegerType())
# 完成任务三
df3 = df.withColumn('new_work_year', udf_cal_work_year(df.work_year))
df3 = df3.select(df3.user_id, df3.censor_status, df3.new_work_year).filter(df3.new_work_year > 5).toPandas().to_csv("3-3.csv",inde
```

出现问题：主要是 bdkit 使用不习惯，对应的包去安装出现了很多卡顿和报错问题，之后安装好就可以了。

p2.ipynb 3.py 2.py

code > 3.py > ...

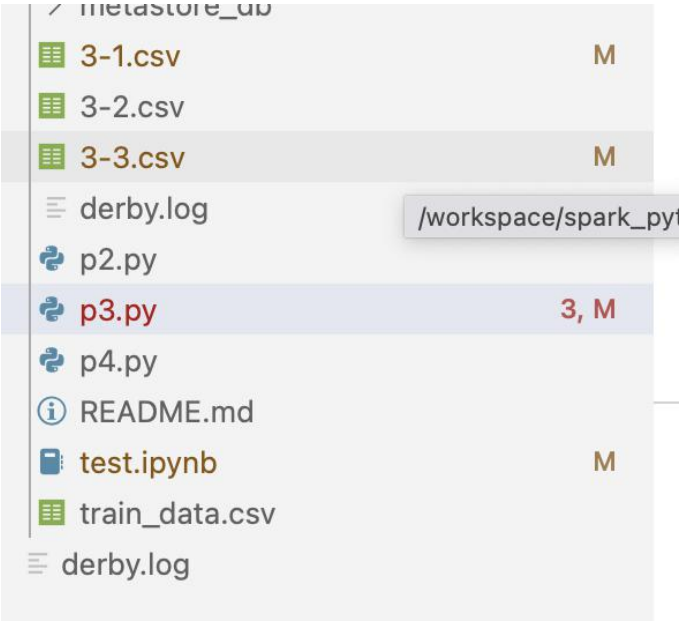
```
#第三题
from pyspark.sql import SparkSession
```

3.py 1 of 4 problems

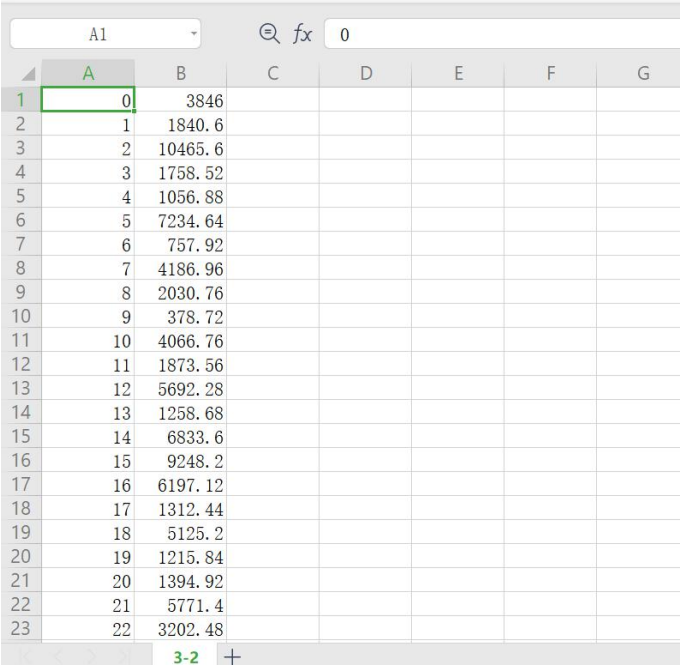
Unable to import 'pyspark.sql' pylint(import-error)

```

spark = SparkSession.builder \
    .enableHiveSupport().getOrCreate()

df=spark.read.options(header='True',inferSchema='True').csv("train_data.csv")
```

运行结果：



	A	B	C	D
1	幼教与中小	0.1		
2	上市企业	0.1001		
3	政府机构	0.2582		
4	世界五百强	0.0537		
5	高等教育机	0.0337		
6	普通企业	0.4543		
7				
8				



	A	B	C	D	E	F	G
1	0	3846					
2	1	1840.6					
3	2	10465.6					
4	3	1758.52					
5	4	1056.88					
6	5	7234.64					
7	6	757.92					
8	7	4186.96					
9	8	2030.76					
10	9	378.72					
11	10	4066.76					
12	11	1873.56					
13	12	5692.28					
14	13	1258.68					
15	14	6833.6					
16	15	9248.2					
17	16	6197.12					
18	17	1312.44					
19	18	5125.2					
20	19	1215.84					
21	20	1394.92					
22	21	5771.4					
23	22	3202.48					

3-2

A1 fx 1

	A	B	C	D	E	F
1	1	2	10			
2	2	1	10			
3	5	2	10			
4	6	0	8			
5	7	2	10			
6	9	0	10			
7	10	2	10			
8	15	1	7			
9	16	2	10			
10	17	0	10			
11	18	1	10			
12	20	1	7			
13	21	2	10			
14	25	2	10			
15	26	0	10			
16	30	0	10			
17	31	0	6			
18	33	1	10			
19	38	0	10			
20	39	1	10			
21	40	1	6			
22	45	1	6			
23	46	0	8			

第四题：

遇见问题：

一开始试图运行 scala 的代码，用 sbt 打包后运行，但是因为本地改过数据集，然后传新数据集到 bdkit 又太大了，就失败了。

```
}
val conf = new SparkConf().setAppName("DefaultForecast")
val sc = new SparkContext(conf)
val spark=  SparkSession.builder().getOrCreate()
import spark.implicits._
val rdd = sc.textFile(args(0)).map(x=>x.split(",")).repartition(1)
var data = rdd.map(x=>Data(x(0),x(1),x(2).toInt,x(3).toInt,x(4).toInt,x(5).toInt,x(6).toInt,x(7).toInt,x(8).toInt,x(9).toInt,x(10).toInt)).toDF()
val assemble = new VectorAssembler().setInputCols(Array("work_type","employer_type","industry","house_exist",
"house_loan_status","censor_status","marriage", "offsprings")).setOutputCol("features")
data = assemble.transform(data)
val Array(trainData,testData) = data.randomSplit(Array(0.8,0.2))
val classifier: DecisionTreeClassifier = new DecisionTreeClassifier().setLabelCol("label").setFeaturesCol("features").setMaxBins(16).setImpurity("gini").setS
(10)
val dtcModel: DecisionTreeClassificationModel = classifier.fit(trainData)
val treetrainPre = dtcModel.transform(trainData)
```

后来运行 py 在 bdkit 上运行发现会输出不了结果，就在 pycharm 本地上重新配置 spark 运行。

```
# 逻辑回归
log_reg = cl.LogisticRegression(labelCol='is_default').fit(train_df)
res = log_reg.transform(test_df)
log_reg_auc = ev.BinaryClassificationEvaluator(labelCol="is_default").evaluate(re
print("逻辑回归：%f" % log_reg_auc)

# 决策树
DTC = cl.DecisionTreeClassifier(labelCol='is_default').fit(train_df)
res = DTC.transform(test_df)
DTC_auc = ev.BinaryClassificationEvaluator(labelCol="is_default").evaluate(res)
print("决策树：%f" % DTC_auc)
# 支持向量机
SVM = cl.LinearSVC(labelCol='is_default').fit(train_df)
res = SVM.transform(test_df)
SVM_auc = ev.BinaryClassificationEvaluator(labelCol="is_default").evaluate(res)
print("支持向量机：%f" % SVM_auc)
```

运行结果：训练集和测试集 8：2

```
逻辑回归: 0.814323
21/12/19 16:25:23 WARN
决策树: 0.616604
支持向量机: 0.796813
```

训练集和测试集 7：3

```
逻辑回归: 0.814973
21/12/19 16:27:26 WARN
决策树: 0.616595
支持向量机: 0.802035
```

训练集和测试集 9：1

```
逻辑回归: 0.813057
21/12/19 16:29:32 WARN
决策树: 0.613447
支持向量机: 0.796562
```