大数据实验要求

2020春

分布式计算平台实验：请参考最新的spark文档

■ 第0步，准备工作

- 0.1准备Linux操作系统，ubuntu 16.04
- 0.2安装JDK1.8，配置JAVA_HOME
- 0.3下载spark2.4，解压缩

■ 第1步 配置Spark standalone模式

- 1.1 开启terminal 进入解压缩后的spark文件夹
- 1.2 运行 ./bin/pyspark
- 1.3在线运行
- textFile = spark.read.text("README.md")
- textFile.count()

■ 第2步 配置singlenode cluster模式 **（选做）**

2.1修改conf/slaves 文件，增加localhost
2.2 sbin/start-master.sh查看http://localhost:8080
2.3sbin/start-slaves.sh spark://127.0.0.1:7077 查看 http://localhost:8080
2.4 运行/bin/pyspark --master spark://127.0.0.1:7077
2.5（退出上一步的交互式环境）运行Spark例子：
./bin/spark-submit --master spark://127.0.0.1:7077 ./examples/src/main/python/pi.py 1000
查看http://localhost:8080中的信息

■ 第3步 愉快的开始spark Python编程(运行ppt中的例子)

频繁模式挖掘实验

■内容1.运行代码 (位于code-Apriori.zip)

■内容2.阅读代码与Apriori算法对应，将伪代码描述对应到代码，写清注释；

■内容3.更换较大的数据集income.csv,

✓以最小支持度为0.1，最小置信度为0.5建立Apriori关联规则

✓以最小支持度为0.1，最小置信度为0.6建立Apriori关联规则

✓以最小支持度为0.2，最小置信度为0.5建立Apriori关联规则

- 比较三个关联规则的数目。

分类实验

- 0.以没有训练模型过程的方式（自己写代码）实现KNN方法；（可以使用code_KNN.zip中的代码作为测试用例）
- 1.要求可以选择不同的距离度量准则：L2，L1；
- 2.比较在不同距离度量下分类面的表现；3.比较在不同k值下分类面的表现并分析原因；
- 4.比较与sklearn提供的API在计算时间上的区别，找到原因；
- 原因提示：数据结构、算法、编译

聚类实验

- 内容1：运行代码了解聚类算法的基本作用（code-cluster.zip）。
- 内容2：查看min_samples参数在不同取值下的结果，理解参数含义。
- 内容3：将DBSCAN算法更换为KMeans算法，观察区别。
- 内容4：总结原型聚类方法KMeans和密度聚类方法DBSCAN的区别。

基本推荐方法实验

- 内容1. 对照基于用户的协同过滤算法的伪代码，完善代码注释。(code-recsys-CF.zip)
- 内容2. 在样例代码的数据集上，实现基本的基于物品的协同过滤算法（Item - CF）