实验 5：频繁项集挖掘

实验内容与要求

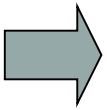
1.请实现课堂上介绍的“Apriori频繁项集挖掘算法”。

2.要求程序利用Spark进行并行计算（动手搭建Spark环境，可采用伪分布式模式），Spark版本号为2.3。

3.在给定最小支持度min_supp下，输出所有极大频繁项集（包括给出极大频繁项集的项数及其支持度）。

4. 输入输出文件的格式和其他具体要求请见FTP上“实验要求”文件夹下对应的详细要求PDF文档。

```
44 46 49 52 55 58 61 64 67 70 73
43 46 49 52 55 58 61 64 67 70 74
44 46 49 52 55 58 61 64 67 70 74
44 46 49 52 55 58 61 64 67 70 74
44 46 49 52 55 58 61 64 67 70 74
44 46 49 52 55 58 61 64 67 70 74
44 46 49 52 55 58 61 64 67 70 74
45 46 49 52 55 58 61 64 67 70 73
```



```
极大频繁项集的项数为6项：
[34,40,48,60,62,66]: 0.8153942
[3,29,36,52,58,60]: 0.84574467
[25,29,40,48,58,60]: 0.8426158
[3,29,34,36,52,58]: 0.8153942
[5,34,36,40,52,58]: 0.86451817
```

实验5：频繁项集挖掘

实验内容与要求

5.实验结果提交：要求书写一个实验报告，其中包括：

- 实验设计说明，包括主要设计思路、算法设计、程序和各个类设计说明
- 本地Spark环境的搭建说明及其截图
- 基于Spark 的 Apriori 并行算法设计思路
- 算法的伪代码(或者带注释的实际代码最终统计出的极大频繁项集及其支持度，输出结果文件的截图，如果使用实际代码，请做好排版)
- 程序运行性能分析
- 性能扩展性等方面可能存在的不足和可能的改进之处
- 源代码、可执行程序 JAR 包(.py 文件)、JAR 包(.py 文件) 运行方式说明(助教可能会重新执行 JAR 包或python文件)
- WebUI 执行报告（WebUI 默认为 http://localhost:4040）
- 实验报告文件命名规则：MPLab5-组号-组长姓名.doc
- 实验完成时间： 1 月 10 日前完成并提交报告