实验二：聚类与分类

1. 实验目的

掌握对数据进行聚类分析和分类的办法，并理解其在大数据环境下的实现方式。

2. 实验环境

操作系统：Windows、MacOS、Linux(建议)
框架：伪分布式Hadoop环境
编程语言：Java

3. 实验内容

3.1 聚类分析

1. 基于Hadoop环境，将聚类分析数据集中的数据导入HDFS中；
2. 基于MapReduce框架实现K-Means聚类算法，对聚类分析数据集进行聚类分析（K值请同学们自行设定），并将聚类结果导出至HDFS中；
3. 自行选择K-Means以外的聚类算法，基于MapReduce框架进行实现，对聚类分析数据集进行聚类分析，并将聚类结果导出至HDFS中；
4. 聚类结果的导出格式为每行一个标签，标签为数字格式，标签的顺序与原始数据集中的数据顺序相同，以下为示例：

```
1
2
1
2
...
0
```

3.2 数据分类

1. 基于Hadoop环境，将分类数据集中的训练数据、验证数据和测试数据均导入HDFS中；
2. 基于MapReduce框架实现朴素贝叶斯算法，使用训练数据对模型进行训练，然后对测试数据进行测试，并将测试结果导出至HDFS中；
3. 自行选择朴素贝叶斯以外的分类算法，基于MapReduce框架进行实现，使用训练数据对模型进行训练，然后对测试数据进行测试，并将测试结果导出至HDFS中；
4. 分类结果的导出格式与聚类部分相同，每行一个标签，标签为数字格式，标签的顺序与原始数据集中的测试数据的顺序相同；

4. 分数说明

1. 完成实验3.1中的前两步，可以获得本实验总分的35%；
2. 完成实验3.2中的前两步，可以获得本实验总分的35%；
3. 我们会测试实验3.1中K-Means算法导出的聚类结果并进行评价，根据评价结果与代码可以获得本实验总分的0%~5%；

4. 我们会测试实验3.2中聚类分析算法导出的分类结果并进行评价，根据评价结果与代码可以获得本实验总分的0%~5%；
5. 完成实验3.1中的第三步，并根据其导出的聚类结果的评价，可以获得本实验总分的8%~10%；
6. 完成实验3.2中的第三步，并根据其导出的分类结果的评价，可以获得本实验总分的8%~10%；

5. FAQ

Q：分类数据集中的验证数据的作用？

A：可以帮助同学们查看下自己的算法的大致准确率，避免算法过拟合等；

Q：聚类与分类的浮动分数部分，算法结果越好（如准确率较高）分数越高吗？

A：我们对不同的算法的标准不同，如朴素贝叶斯算法通常情况下准确率并不会太高，但即使如此，如果准确率只有50%显然也是不合理的，会影响分数；此外结果不是分数判定的唯一标准，我们也会根据同学们的代码水平、风格等进行综合判断；