达观数据技术新兵训练营

欢迎你加入光荣达观数据技术战队！在成为正式的达观特战队的成员之前，我们安排了为期一周的新兵训练营，让新同学能快速掌握战场技能，成为骁勇善战的达观战士！

你需要在**规定时间**依次完成三道培训题，请牢记培训要点：

1 每位新同学会被指派一位导师（Mentor），导师会帮助大家开通开发帐号、设置开发环境、并检查程序代码

2 请认真阅读培训题，依次完成任务，每题完成后请导师对编码进行review，review通过后可以完成下一题，review不通过则被打回修改，直至通过

3 请严格按照任务时间点完成训练题，如果在时间点内无法完成需要向导师申请延期，每次申请，多延期一天

4 训练题请大家独立完成，可以上网搜索资料，实在遇到无法解决的困难可以找导师求助，导师会给出提示，但不会代劳

5 确保所编写的程序完成功能，又能规范严谨。完成的又快又好来证明你的实力！

6 开发环境Linux，开发语言要求用Python2.7，如果需要，可以使用外部算法库作为补充

## 培训题一寻找两个文件中相同的句子

已有两个文件a.dat和b.dat，每个文件都包含若干行句子，现在需要你编写python程序，来找到两个文件中相同的句子，完成：

1 将相同的句子打印到common.dat文件中，每行一句

2 在控制台打印出程序执行的时长，和相同句子的数量

【数据路径】 123.56.231.83:/home/common/onboard/train1/data

【时间点】一天内完成

## 培训题二查找最相似的英文句子

已有文件sentence.dat是由很多行英文句子构成，请你对这个文件进行处理，构造相应的数据结构，开发出一个快速查找最相似句子的系统（相似度计算方法见附注），该系统能实现的功能为：

1 在控制台（console）中可以随时输入一个英文句子，单词长度小于等于8，运算后立刻在console中打印出在sentence.dat中与输入的英文句子**最相似**的前十个英文句子，和对应的相似度。

2 每执行依次相似句子查询后，在控制台打印出本次执行的总时长，并要求从每次从console获得输入到执行完毕，总时长**小于100毫秒**。

附注：两个英文句子A和B的相似性计算公式为：

分子表示两个句子中**相同单词的数量**，分母子表示两个句子中**总共的单词数量**，例如句子A=“we are the world”，句子B=“they are the students”，则为2（注：相同单词为 are、the），为6（注：总共单词为we, are, the, world, they, students），则A和B的相似度为2/6 = 0.3333，两个完全相同的句子的相似度为1，完全不同的句子的相似度为0

【数据路径】 123.56.231.83:/home/common/onboard/train2/data

【时间点】二天内

## 培训题三猜猜谁拿的年薪能超过5万美金

在salarytrain.dat文件中有超过三万行用户数据，每行中的用户个人数据（按逗号分隔）依次包括如下字段：

1. Age: continuous

2. Workclass: 8 values

3. Fnlwgt: continuous.

4. Education: 16 values

5. Education-num: continuous.

6. Marital-status: 7 values

7. Occupation: 14 values

8. Relationship: 6 values

9. Race: 5 values

10. Sex: Male, Female

11. Capital-gain: continuous.

12. Capital-loss: continuous.

13. Hours-per-week: continuous.

14. Native-country: 41 values

15. >50K Income: Yes, No

其中第15个字段（Income）表示该用户的年薪是否超过了5万美元，现在请你根据salarytrain.dat中各个字段的数据，设计算法来预测salarytest.dat中哪些用户的年薪超过了5万美元。最终结果要求为：

1 按salarytest.dat中的数据，生成预测结果文件 salaryresults.dat，文件中每行为一个结果（<=50K或>50K），和salarytest.dat中每行一一对应。（格式可参考sampleresults.dat）

2 运行python /home/common/onboard/train3/evaluate.pyc ${predict\_file}来获得算法准确率得分（注：准确率=预测正确的结果数/总结果数），要求得分至少要**高于0.8**，并且越高越好

【数据路径】 123.56.231.83:/home/common/onboard/train3/data

【时间点】二天内