Homework#3

一、分类（5 分）

银行希望通过分析客户的职业、婚姻状况和教育水平来预测他们是否对特定的投资基金产品感兴趣。下面是客户的数据。

1）使用任意一种决策树方法建立该数据集的二分类器，使它能正确分类客户是否对投资基金产品感兴趣，写出建立过程（2 分）。并用所建分类器说明给定客户（职员，单身，非高等教育）是否对投资基金产品感兴趣（0.5 分）。

2）用朴素贝叶斯建立二分类器，写出建立过程（不用考虑平滑）(2 分)，并用所建分类器说明给定客户（职员，单身，非高等教育）是否对投资基金产品感兴趣（0.5 分）。

客户编号	职业类型	婚姻状况	教育水平	对投资基金感兴趣
1	职员	已婚	非高等教育	否
2	工人	单身	非高等教育	是
3	管理者	已婚	高等教育	是
4	职员	离异	非高等教育	否
5	工人	已婚	高等教育	是
6	管理者	单身	非高等教育	否
7	管理者	离异	高等教育	是

8	职员	单身	非高等教育	否
9	职员	已婚	高等教育	是
10	工人	单身	高等教育	是

二、聚类（5 分）

1）给定下列 12 个数据点：

(2,3); (1,2); (3,1); (3,3); (2,2); (4,2); (5,4); (6,3); (5,6); (6,5); (7,4); (7,6)

使用 K-means 算法对它们聚类。令 k=2，初始中心点为（1，3）和（7，5），写出聚类过程（2 分）。

2）我们提供了 twitter 的语料，在 twitter.txt 文件中。每一行表示一个 twitter 的推文。请使用任意一种编程语言，对该语料进行 K-means 聚类。请在聚类后给出每类的关键词，尝试不同的 k 值（k=2,3,4）进行分析（3 分）。

提示：

a. 对语料进行去除停用词、分词等预处理，将每个推文表示成 tf-idf 向量， 将 tf-idf 向量作为推文的表示进行聚类。

b. tf-idf 和 K-means 算法可以调用直接调用第三方的库。 提交说明:需要提交源代码与报告。报告中简单说明 2)的实现思路，结果与分析。