HOMEWORK #4（10分）

一、分类：（5分）

下表给出了两组人的数据，每组数据分别包含4个和5个样本。

1）使用任意一种决策树方法建立该数据集的二分类器，使它能正确区分这两组人，写出建立过程（2分），并用所建分类器说明给定样例（矮，金色，成年）是属于第几组（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），并用所建分类器对给定样例（矮，金色，成年）分类（0.5分）。

组次	id	身材	发色	年龄
第一组	1	矮	金色	老人
	2	高	红色	老人
	3	高	金色	老人
	4	矮	金色	成年
第二组	1	高	黑色	儿童
	2	矮	黑色	老人
	3	高	黑色	老人
	4	高	黑色	成年
	5	矮	金色	儿童

二、聚类：（5分）

1）给定下列 13 个数据点：

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

使用 K-means 算法对它们进行聚类。令 k=2，初始中心点为(0,4)和(6,5)，写出聚类过程（2 分）。

2）我们提供了 HuffPost 的部分新闻语料，在文件 news.txt 中，每一行表示一篇新闻的标题。请使用任意一种编程语言，对新闻标题进行 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，将每个新闻标题表示成 tf-idf 向量，将 tf-idf 向量作为新闻标题的表示进行聚类。

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。