**任务三：IMDB数据集电影评测分类（二分类问题）**

**（1）数据集讲解**：

该数据集是IMDB电影数据集的一个子集，已经划分好了测试集和训练集，训练集包括25000条电影评论，测试集也有25000条，该数据集已经经过预处理，将每条评论的具体单词序列转化为词库里的整数序列，其中每个整数代表该单词在词库里的位置。例如，整数104代表该单词是词库的第104个单词。为实验简单，词库仅仅保留了10000个最常出现的单词，低频词汇被舍弃。每条评论都具有一个标签，0表示为负面评论，1表示为正面评论。

训练数据在train\_data.txt文件下，每一行为一条评论，训练集标签在train\_labels.txt文件下，每一行为一条评论的标签；测试数据在test\_data.txt文件下，测试数据标签未给出。

**（2）思路：**

这里提供一个最简单的思路，还有其它思路同学们可以自行思考。

首先，需要将每条评论转换为特征向量，这里可以采用one-hot编码，举个例子：这里采用的词库大小为10000，因此转换的one-hot编码也是10000维的，如某条评论为[3, 5]，则转换得到的one-hot编码的10000维向量，只有索引为3和5的元素为1，其余全部为0。

将每条评论都转换为one-hot编码后，再采用决策树算法进行分类。

**（3）具体要求**：

将测试数据预测结果，与训练数据标签存储方式相同，存储为txt文件，每一行为一条评论的标签。实验报告（和前两个任务的报告写在一起，上传到决策树作业栏）中需要写明具体实验流程，思路等。

将测试集预测结果的txt文件单独上传到对应编程作业栏（这里不上传报告，只上传预测结果txt文件）。如下。

