2022夏_第二次作业

2.1. 写出倒排记录表（694，16642，307645，4784824）的可变字节编码及γ 编码。（对间距而不是文档ID编码）

**2.2. 假设对某倒排记录表的间距进行 γ 编码的结构是：
11110101011100011101110111011。请还原原始间距序列及倒排记录表。**

2.3. 考虑下表中的3篇文档doc1、doc2、doc3中几个词项的tf情况，并且词项car、auto、insurance及best的idf值分别是2.32、1.45、4.21、5.21

	doc1	doc2	doc3
car	43.2	0	31.0
auto	6.3	63.2	0
insurance	26.3	64.2	5.37
best	0	27.9	15.4

a). 计算对应的所有 **tf-idf** 值

b). 分别计算三个文档的文档向量（其中每个向量有4维，每维对应一个词项）

c). 对于查询 **auto insurance**，计算三篇文档的余弦相似度得分并排序（计算查询向量时，查询中出现的词权重记为1，反之记为0）

2.4. 设题目2.3中doc1、doc2和doc3的静态质量得分分别为0.3、0.1和0.6，画出文档按静态得分排序的词项倒排索引。（即 g(d) + tf-idf，tf-idf 使用欧几里得归一化后的结果）