2.1. 写出倒排记录表（694，16642，307645，4784824）的可变字节编码及γ编码。

可变字节码：

- 694：00000101 10110110
- 15948：01111100 11001100
- 291003：00010001 01100001 10111011
- 4477179：00000010 00010001 00100001 11111011

γ编码：

- 694：1 11111111 0,0 10110110
- 15948：11111 11111111 0,11110 01001100
- 291003：11 11111111 11111111 0,00 01110000 10111011
- 4477179：111111 11111111 11111111 0,000100 01010000 11111011

2.2. 假设对某倒排记录表的间距进行γ编码的结构是：1111010101110001 11011 101 11011。请还原原始间距序列及倒排记录表。

γ编码:

- 111101010: 26
- 1110001: 9
- 11011: 7
- 101: 3
- 11011: 7

倒排记录表: (26, 35, 42, 45, 52)

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

考虑到有0存在，用$w_{ij} = tf_{ij} * idf_i$来计算权重：

	doc1	doc2	doc3
car	100.224	0	71.92
auto	9.135	91.64	0
insurance	110.723	270.282	22.6077
best	0	145.359	80.234

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）

- doc1：(100.224, 9.135, 110.723, 0) 采用L2归一化后为：(0.6699, 0.0611, 0.7400, 0)
- doc2：(0, 91.64, 270.282, 145.359) 采用L2归一化后为：(0, 0.2861, 0.8439, 0.4538)
- doc3：(71.92, 0, 22.6077, 80.234) 采用L2归一化后为：(0.6532, 0, 0.2053, 0.7288)

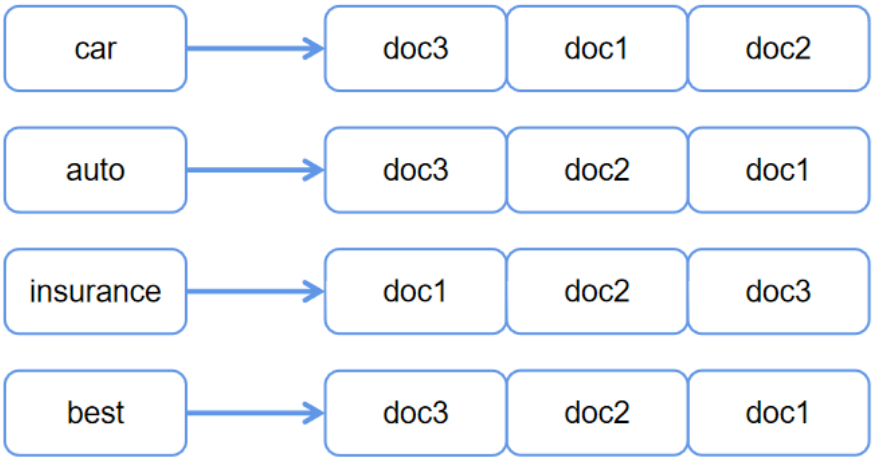
c). 对于查询 auto insurance，计算三篇文档的余弦相似度得分并排序（计算查询向量时，查询中出现的词权重记为1，反之记为0）

采用归一化之后的向量：

- Q: (0, 1, 1, 0)
- cos(Q, doc1) = 0.5664
- cos(Q, doc2) = 0.7990
- cos(Q, doc3) = 0.1451
- cos(Q, doc2) > cos(Q, doc1) > cos(Q, doc3)

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）

	doc1	doc2	doc3
car	0.9699	0.1	1.2532
auto	0.3611	0.3861	0.6
insurance	1.0400	0.9439	0.8053
best	0.3	0.5538	1.3288


car
doc3
doc1
doc2
auto
doc3
doc2
doc1
insurance
doc1
doc2
doc3
best
doc3
doc2
doc1