第二次作业（3.25 周一上课交）

1、γ编码为什么是通用性编码?

2、γ编码对倒排索引进行压缩能达到多高的压缩比?

习题 5-5 [*] 写出倒排记录表（777，17743，294068，31251336）的可变字节编码及γ编码。在可能的情况下对间距而不是文档 ID 编码。写出 8 位块的二元码。

习题 5-8 [*] 对于下列采用γ编码的间距编码结果，请还原原始的间距序列及倒排记录表。

1110001110101011111101101111011

习题 6-10 考虑图 6-9 中的 3 篇文档 Doc1、Doc2、Doc3 中几个词项的 tf 情况，采用图 6-8 中的 idf 值来计算所有词项 car、auto、insurance 及 best 的 tf-idf 值。

	Doc1	Doc2	Doc3
car	27	4	24
auto	3	33	0
insurance	0	33	29
best	14	0	17

图 6-9 习题 6-10 中所使用的 tf 值

词项	df_t	idf_t
car	18 165	1.65
auto	6 723	2.08
insurance	19 241	1.62
best	25 235	1.5

图 6-8 idf 值的例子，本图中给出了 Reuters-RCV1 文档集中不同频率的词项的 idf 值

习题 6-15 回到习题 6-10 中的 tf-idf 权重计算，试计算采用欧氏归一化方式处理后的文档向量，其中每个向量有 4 维，每维对应一个词项。

习题 6-17 基于习题 6-15 的词项权重计算结果，对于查询 car insurance 计算 3 篇文档的得分并进行排序。计算时，查询词项的权重计算分别采用如下方法：

(1) 查询中出现的词项权重为 1，否则为 0；

(2) 采用欧氏方式对 idf 进行归一化。

习题 6-19 计算查询 digital cameras 及文档 digital cameras and video cameras 的向量空间相似度并将结果填入表 6-1 的空列中。假定 N=10 000 000，对查询及文档中的词项权重（wf对应的列）采用对数方法计算，查询的权重计算采用 idf，而文档归一化采用余弦相似度计算。将 and 看成是停用词。请在 tf 列中给出词项的出现频率，并计算出最后的相似度结果。

12

表6-1 习题6-19中的余弦相似度计算

词	查询					文档			$q_i \cdot d_i$
	tf	wf	df	idf	q_i=wf-idf	tf	wf	d_i=归一化的wf	
digital			10 000						
video			100 000						
cameras			50 000						