1. 考虑如下几篇文档：

   文档 1　new home sales top forecasts

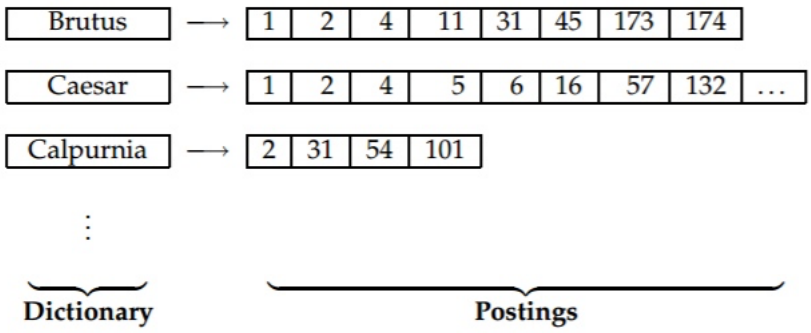
   文档 2　home prices rise in june

   文档 3　increase in home sales in june

   文档 4　july new home sales rise

   a. 画出文档集对应的词项-文档矩阵

   b. 参考下图例子，画出该文档集的倒排索引



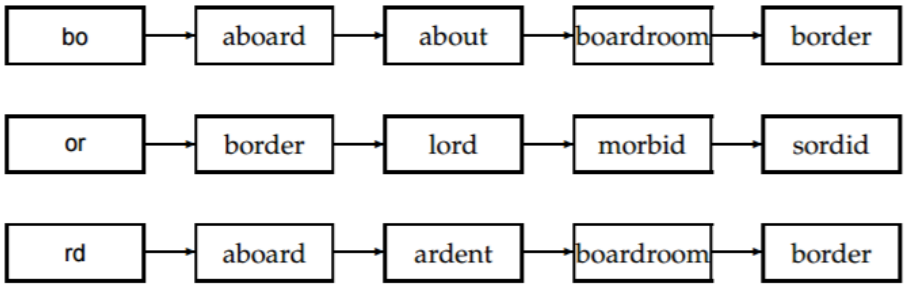
   c. 给定如下查询，返回的结果是什么？

      a) rise AND new

      b) sales AND NOT (forecasts OR july)

2. 写出倒排记录表(777，17743，294068，31251336)的 VB 编码以及 γ 编码。在可能的情况下对间距而不是文档 ID 编码。请以 8 位二进制串的方式写出这些编码。(γ 编码不考虑对 0 编码问题，仅对原始文档 ID 以及间隔进行编码)

3. 给定以下词



   a) 计算查询 “bord” 与图中每个包含 2-gram “or” 的词项之间的 2-gram Jaccard 系数，并写出计算过程。

   b) 思考计算 k-gram（k>2）时，如何添加首尾标志符？（给出思路即可，不需要计算）

4. 考虑表 1 中的 3 篇文档 Doc1、Doc2、Doc3 中几个词项的 tf 情况，表 2 为词项在所有文档中的 idf 值。

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

表 1 tf 值

	df_t	idf_t
car	18165	1.65
auto	6723	2.08
insurance	19241	1.62
best	25235	1.5

表 2 idf 值

a. 计算所有词项 car、auto、insurance、best 的 tf-idf 值。
b. 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。
c. 对于查询 car insurance 计算 3 篇文档的得分并进行排序。查询词项的权重计算采用：查询中出现的词项权重为 1，否则为 0。