第三次作业（4.1 周一上课交）

习题 7-3 假定所有查询都仅包含单个词项，请说明为什么采用全局胜者表（$r=K$）已经能够充分保证找到前 K 篇文档。如果所有查询都只由 s 个词项组成（s 是个常数且 $s>1$），如何对上述思路进行修正以保证能够找到前 K 篇文档？

习题 7-5 重新考察习题 6-23 中基于 nnn.atc 权重计算的数据，假定 Doc1 和 Doc2 的静态得分分别是 1 和 2。请确定在公式（7-2）下，如何对 Doc3 的静态得分进行取值，才能分别保证它能够成为查询 best car insurance 的排名第一、第二或第三的结果。

习题 7-7 设定图 6-10 中 Doc1、Doc2 和 Doc3 的静态得分分别是 0.25、0.5 和 1，画出当使用静态得分与欧几里得归一化 tf 值求和结果进行排序的倒排记录表。

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

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

$$\text{net-score}(q,d)=g(d)+\frac{\vec{V}(q)\cdot\vec{V}(d)}{|\vec{V}(q)||\vec{V}(d)|} \tag{7-2}$$

习题 8-7 [**] 两个集合之间的 Dice 系数是度量其交集大小占两个集合大小之和的比率的一个指标，取值在 0 到 1 之间，其计算公式为

$$\text{Dice}(X,Y)=\frac{2|X\cap Y|}{|X|+|Y|}$$

请证明，平衡 F 值等于检索结果文档集合和相关文档集合的 *Dice* 系数。

习题 8-10 [**] 下表中是两个判定人员基于某个信息需求对 12 个文档进行相关性判定的结果（0=不相关，1=相关）。假定我们开发了一个 IR 系统，针对该信息需求返回了文档集{4, 5, 6, 7, 8}。

docID	判断 1	判断 2
1	0	0
2	0	0
3	1	1
4	1	1
5	1	0
6	1	0

7	1	0
8	1	0
9	0	1
10	0	1
11	0	1
12	0	1

a. 计算两个判断之间的 kappa 统计量；

b. 当两个判断均认为是相关文档时才认为该文档相关，此时计算上述系统的正确率、召回率及 F_1 值；

c. 只要有一个判断认为是相关文档则认为该文档相关，此时计算上述系统的正确率、召回率及 F_1 值

习题 9-3 假定用户的初始查询是 cheap CDs cheap DVDs extremely cheap CDs。用户查看了两篇文档 d_1 和 d_2，并对这两篇文档进行了判断：包含内容 CDs cheap software cheap CDs 的文档 d_1 为相关文档，而内容为 cheap thrills DVDs 的文档 d_2 为不相关文档。假设直接使用词项的频率作为权重（不进行归一化也不加上文档频率因子），也不对向量进行长度归一化。采用公式（9-3）进行 Rocchio 相关反馈，请问修改后的查询向量是多少？其中 $\alpha = 1$，$\beta = 0.75$，$\gamma = 0.25$。

习题 9-4 [*] Omar 实现了一个带相关反馈的 Web 搜索系统，并且为了提高效率，系统只基于返回网页的标题文本进行相关反馈。用户对结果进行判定，假定第一个用户 Jinxing 的查询是

banana slug

返回的前三个网页的标题分别是：

banana slug Ariolimax columbianus

Santa Cruz mountains banana slug

Santa Cruz Campus Mascot

Jinxing 认为前两篇文档相关，而第 3 篇文档不相关。假定 Omar 的搜索引擎只基于词项频率（不包括长度归一化因子和 IDF 因子）进行权重计算，并且假定使用 Rocchio 算法对原始查询进行修改，其中 $\alpha = \beta = \gamma = 1$。请给出最终的查询向量（按照字母顺序依次列出每个词项所对应的分量）。

$$\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j \qquad (9\text{-}3)$$

习题 9-7 如果词项-文档矩阵 A 是一个布尔共现矩阵，那么矩阵 C 中的元素是什么？