《信息检索》作业4

计83李天勤2018080106

题目1：Shown below is portion of a positional index ...

angels: 2: ⟨36,174,252,651⟩; 4: ⟨12,22,102,432⟩; 7: ⟨17⟩;
fools: 2: ⟨1,17,74,222⟩; 4: ⟨8,78,108,458⟩; 7: ⟨3,13,23,193⟩;
fear: 2: ⟨87,704,722,901⟩; 4: ⟨13,43,113,433⟩; 7: ⟨18,328,528⟩;
in: 2: ⟨3,37,76,444,851⟩; 4: ⟨10,20,110,470,500⟩; 7: ⟨5,15,25,195⟩;
rush: 2: ⟨2,66,194,321,702⟩; 4: ⟨9,69,149,429,569⟩; 7: ⟨4,14,404⟩;
to: 2: ⟨47,86,234,999⟩; 4: ⟨14,24,774,944⟩; 7: ⟨199,319,599,709⟩;
tread: 2: ⟨57,94,333⟩; 4: ⟨15,35,155⟩; 7: ⟨20,320⟩;
where: 2: ⟨67,124,393,1001⟩; 4: ⟨11,41,101,421,431⟩; 7: ⟨16,36,736⟩;

对输入查询

1. "fools rush in"
2. "fools rush in" AND "angels fear dread"

分别给出相应返回的文档集。（注：双引号表示其内含查询词串应视作短语）

1. "fools rush in" : document 2, position 1; document 4, position 8; document 7, position 3, position 13
2. "fools rush in" AND "angels fear to tread"; document 4

题目2：参考课堂PPT中Heaps' Law相关内容。我们根据RCV1语科库统计数据，已经得到了Heaps' Law如下两个参数 $k = 44$ 和 $b = 0.49$

Heaps Law is defined as

$$V = Kn^{\beta}$$

where V is the size of the vocabulary and n is the length of the corpus in words (number of tokens in the collection)

那么，当同类语料库规模分别为1亿词，100亿词以及5000亿词时，对应的词典规模分别时多大呢?

$V = 44 * (100,000,000)^{(0.49)} = 365976.059$

$V = 44 * (10,000,000,000)^{(0.49)} = 3495044.23$

$V = 44 * (500,000,000,000)^{(0.49)} = 23765556$