**信息检索HW4**

周凯兴 2018080052 精83

题目1：

1. "fools rush in"

答："fools rush in" :

document 2, position 1;

document 4, position 8;

document 7, position 3, position 13

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

答："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 ： V = Knβ

V是Vocabulary的大小，n是语料库的单词长度(集合中标记的数量)

答：

1. 同类语料库规模1亿词：

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

1. 100亿词：

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

1. 5000亿词时：

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