Web 信息处理与应用：课后作业 2
网页查询 + 排序 + 结果评估部分
请于 2020 年 11 月 19 日前将作业电子版发送至课程邮箱：ustcweb2019@163.com
作业文件与邮件标题命名：PBXXXXX_XXX（姓名）_HW2

1 计算题

1.1 给定以下词项的 idf 值，以及在三篇文档中的 tf，已知总文档数为 811,400，请完成如下计算任务：

	df	tf@Doc1	tf@Doc2	tf@Doc3
Car	18,871	34	8	32
Auto	3,597	3	24	0
Insurance	19,167	0	51	6
Best	40,014	18	0	13

1) 计算所有词项的 tf-idf 值。
2) 试采用欧式归一化方法（即向量各元素平方和为 1），得到处理后的各文档向量化表示，其中每个向量为 4 维，每一维对应 1 个词项。
3) 基于 2)中得到的向量化表示，对于查询“car insurance”，计算 3 篇文档的得分并进行排序。其中，查询中出现的词项权重为 1，否则为 0。

1.2 考虑如下邻接矩阵表示的图（1 代表出边，0 表示无连接）

	节点 0	节点 1	节点 2	节点 3	节点 4	节点 5	节点 6
节点 0	0	0	1	0	0	0	0
节点 1	0	1	1	0	0	0	0
节点 2	1	0	1	1	0	0	0
节点 3	0	0	0	1	1	0	0
节点 4	0	0	0	0	0	0	1
节点 5	0	0	0	0	0	1	1
节点 6	0	0	0	1	1	0	1

1) 当 Restart 部分的随机跳转概率为 0.15 时，写出 PageRank 的（随机）转移概率矩阵。
2) 计算该矩阵所对应的 PageRank 向量（每一维表示一个节点的 PageRank 值）。
3) 将邻接矩阵中“节点 2 指向节点 3”和“节点 6 指向节点 3”的两条边权重设为 0，其它不变。请计算该矩阵所对应的 Hub 值和 Authority 值向量。

1.3 在由 10,000 篇文档构成的文档集中，某个查询的相关文档总数为 10，下面给出了针对该查询的前 20 个有序结果，其中 R 表示相关，N 表示不相关。

R R N N R　N N N R N　R N N N R　N N N N R

请计算:

a) 该查询的 P@10 和 P@20 分别是多少?
b) 该查询前 10 篇文档和前 20 篇文档的 F1 值分别是多少?
c) 当该算法只返回前 20 个结果是，其简易 AP 值为多少?
假定该算法返回了全部 10,000 篇文档，上述 20 篇文档只是最开始的 20 个结果，那么
d) 该算法可能的最大 AP 值是多少?
e) 该算法可能的最小 AP 值是多少?

2 问答题（言之有理即可）

2.1 请简述解决以下问题的思路:

a) 如何从多源情境信息（如手机的多种传感器信息）中，抽象出用户当前所处的状态或行为模式?
b) 在上述过程中，如何既体现用户的个性化因素，又减少用户个人记录稀疏的负面影响?

2.2 用户在浏览网页时，可能通过点击“后退”按钮回到上一次浏览的页面。用户的这种回退行为（包括连续回退行为）能否用马尔科夫链进行建模？为什么?

2.3 如何在网页排序的同时提升结果的多样化水平？如何在实现这一目的的同时保障算法的效率?