Web 信息处理与应用：课后作业 2
请于 2023 年 10 月 30 日 23:59 前将作业电子版发送至课程邮箱：ustcweb2022@163.com
作业文件与邮件标题命名：PBXXXXX_XXX（姓名）_HW2

1 计算题

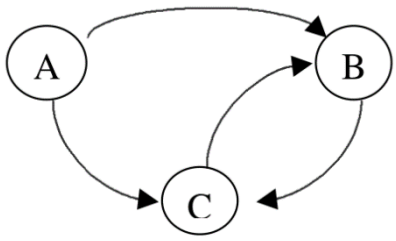
1.1 给定以下词项的 idf 值，以及在三篇文档中的 tf，已知总文档数为 811,400，请完成如下计算任务：

	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，对于查询向量(1, 0, 1, 0)，计算 3 篇文档的得分并进行排序。

1.2 考虑右图的网络结构图

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) 计算各个节点所对应的 PageRank 值、Hub 值和 Authority 值。



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。

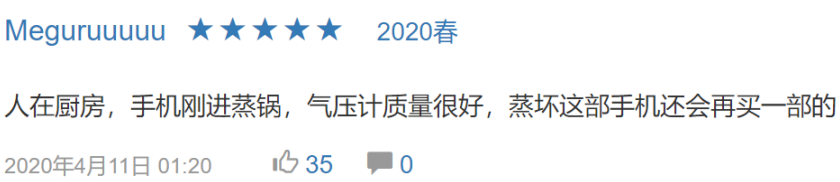
R R N N R　N N N R N　R N N N R　N N N N R

请计算：

a) 该查询的 P@10 和 P@20 分别是多少？
b) 为什么 P、R@N 可能会出现理论上限小于 1 的情况？
c) 该查询前 10 篇文档和前 20 篇文档的 F1 值分别是多少？
d) 当该算法只返回前 20 个结果时，其简化 AP 值为多少？

2　问答题（言之有理即可）

2.1 文本评论是一种较为复杂的用户反馈。通常情况下我们会借助用户评论的倾向性来辅助理解用户的满意程度。然而，部分用户在反馈时可能存在“反话正说”的情况，从而对于我们判断用户满意度造成干扰。以 USTC 评课社区中“大学物理实验”的评价为例：



a) 你认为用户留下的哪些信息（不限该页面）有助于我们判断“反话正说”的现象？

b) 请设计一种机制，在不借助复杂自然语言处理技术进行语义表征的前提下，实现“反话正说”现象的有效判别？

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 在用户意图尚不明确的情况下，使搜索结局具有一定多样性，以确保可能具有不同意图的用户都能够获得相应反馈，是一种常见的排序策略。请简要回答下列问题：

a) 如何在网页排序的同时提升结果的多样化水平？如何在此同时保障算法的效率？

b) 在用户通过点击行为等反馈方式表达了更为具体的意图之后，是否还需要保持结果的多样化？为什么？

2.4 Pairwise 类排序学习方法由于需要对排序样本进行两两成对，导致样本数大为提升，从而造成了计算资源开支的严重增加。试讨论，是否可以减少需要比较的样本对以提升排序效率？如何进行样本对的筛选，并抑制对于排序精度的干扰？