HW2

1. 计算题

1.1

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. 公式:$W_{t,d} = (1 + log(tf_{t,d})) \times log\frac{N}{df_t}$(当$tf_{t,d} \leq 0$时,$1 + log(tf_{t,d})$为0)

   $W_{car,doc1} = (1 + log(34)) \times log\frac{81140}{18871} = 4.1350$

   其余计算同理,得到下表(保留4位小数)

词项\tf-idf值	Doc1	Doc2	Doc3
Car	4.1350	3.1086	4.0920
Auto	3.4761	5.6013	0
Insurance	0	4.4044	2.8925
Best	2.9477	0	2.7630

2. 即将上一问中计算出来各文档的$W_{t,d}$归一化,使其模长为1

   $vec_{doc1} = \frac{\vec{W_{t,d}}}{|W_{t,d}|} = (\frac{4.1350}{\sqrt{37.8704}}, \frac{3.4761}{\sqrt{37.8704}}, 0, \frac{2.9477}{\sqrt{37.8704}}) = (0.6720, 0.5649, 0, 0.4790)$

   $vec_{doc2} = (0.3999, 0.7205, 0.5665, 0)$

   $vecdoc3 = (0.7151, 0, 0.5055, 0.4828)$

3. 根据余弦相似度计算查询向量和文档向量的相似程度

$$\cos\left(\vec{q}, \vec{d}\right) = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2} \sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

其中查询"car insurance"对应的查询向量为$\overrightarrow{query} = (1, 0, 1, 0)$

用已经归一化的第二问中的结果进行计算

$score_{doc1} = \overrightarrow{query} \cdot \overrightarrow{vec}_{doc1} = 0.6720$

$score_{doc2} = 0.3999 + 0.5665 = 0.9664$

$score_{doc3} = 0.7151 + 0.5055 = 1.2206$

排序:$score_{doc3} > score_{doc2} > score_{doc1}$

故返回的文档顺序为doc3 doc2 doc1

1.2

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. $A = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \end{pmatrix}$

$R = \begin{pmatrix} 0 & 0 & 1/3 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1/2 & 1/3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 1/2 & 0 & 0 & 1/3 \\ 0 & 0 & 0 & 1/2 & 0 & 0 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1/2 & 1/3 \end{pmatrix}$

随机转移概率矩阵 $M = dR + \frac{1-d}{N}E$,其中$d = 1 - 0.15 = 0.85$, $N = 7$

计算可得矩阵M为

$$M=\begin{pmatrix}0.0214 & 0.0214 & 0.3048 & 0.0214 & 0.0214 & 0.0214 & 0.0214\\ 0.0214 & 0.4464 & 0.0214 & 0.0214 & 0.0214 & 0.0214 & 0.0214\\ 0.8714 & 0.4464 & 0.3048 & 0.0214 & 0.0214 & 0.0214 & 0.0214\\ 0.0214 & 0.0214 & 0.3048 & 0.4464 & 0.0214 & 0.0214 & 0.3048\\ 0.0214 & 0.0214 & 0.0214 & 0.4464 & 0.0214 & 0.0214 & 0.3048\\ 0.0214 & 0.0214 & 0.0214 & 0.0214 & 0.0214 & 0.4464 & 0.0214\\ 0.0214 & 0.0214 & 0.0214 & 0.0214 & 0.8714 & 0.4464 & 0.3048\end{pmatrix}$$

2. $P_{n+1}=MP_n$

   通过python程序,经过17次迭代,得到最终的 PR 向量

   $(0.05453, 0.37267, 0.11675, 0.24313, 0.21003, 0.03727, 0.30103)^T$

3. 按照HITS算法的python程序迭代,得到最终结果

   $Authority=[0.00058, 0.00058, 0.00157, 0.16846, 0.49801, 0.27257, 0, 80580]^T$

   $Hub=[0.00079, 0.00108, 0.00108, 0.33507, 0.40512, 0.54215, 0.65550]^T$

1.3

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?

1. $P@10=4/10=0.4$

   $P@10=7/20=0.35$

2. $R@10=4/10=0.4$,故$F1@10=0.373$

   $R@10=7/10=0.7$,故$F1@20=0.467$

3. AP=(1/1+2/2+3/5+4/9+5/11+6/15+7/20)/7=0.607

4. APmax=(1/1+2/2+3/5+4/9+5/11+6/15+7/20+8/21+9/22+10/23)/10=0.547

5. APmin=(1/1+2/2+3/5+4/9+5/11+6/15+7/20+8/9998+9/9999+10/10000)/10=0.425

2. 问答题

2.1

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？

1.
   - 结合手机中的GPS定位判断用户位置的变化,可以根据位置变化的快慢来判断用户是在步行或骑车或在某些交通工具上;如果位置长时间不变化,可再根据当前时间判断用户在上班或上学或睡觉. 以此可得到用户一天的行为轨迹
   - 结合手机中的声音传感器,判断周围环境是否嘈杂. 如果环境较为嘈杂,则用户可能是在室外, 如果环境较为安静,则用户大概率在室内
   - 结合手机中的加速度传感器可以判断用户是否在乘坐电梯以及是否在跑步等等
   - 结合手机中的感光设备,可以辅助判断用户的位置. 如当前时间为白天但感光设备感到周围很暗,则用户大概率在密闭的室内休息;反之如果当前时间为夜晚但感光设备感到周围很亮,则用户大概率在亮度较高的室内
2.
   - 为了不侵犯隐私,可以向用户申请是否允许使用这些信息,或找有关志愿者,收集这些目标人员的有关数据, 人工对这些数据进行标注作为训练集, 对模型进行训练、调参，这样可以减少用户个人记录稀疏的影响
   - 将每个用户独有的个性化信息放入对应数据库中，用这些数据对模型进行微调，从而生成针对每个用户的个性化模型，体现用户的个性化因素

2.2

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？

1 马尔可夫链的概念及转移概率

[定义] 设有随机过程 $\{X_n, n \in T\}$，若对于任意的整数 $n \in T$ 和任意的 $i_0, i_1, \ldots, i_{n+1} \in I$，条件概率满足

$$P\{X_{n+1} = i_{n+1} \mid X_0 = i_0, X_1 = i_1, \cdots, X_n = i_n\}$$
$$= P\{X_{n+1} = i_{n+1} \mid X_n = i_n\}$$

则称 $\{X_n, n \in T\}$ 为**马尔可夫链**，简称**马氏链**。

不能。 回退行为（或连续的回退行为）会回到上一个（或多个）页面，这和上一个（或多个）页面相关，而不仅仅与回退前所在的页面有关，故回退行为不具有马尔科夫性，不能用马尔科夫进行建模

2.3

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？

- 将每个网页打上其主题相关的若干个标记，在返回网页排序时兼顾这些标记，在保证网页质量、用户需求的同时尽量将具有不同标记的网页放在前面。（同时可根据用户的历史行为等作相关性的推荐来提升用户对网页主题的满意度）具体实现时可以主题重合度大的页面给予一定的惩罚，以此不断迭代从而获得差异度较大的一组Top N文档
- 效率：将比较对象由网页转换成主题标签，在每个主题标签对应的网页集合中找到少数质量最高、最符合用户需求的网页，并将这些网页组合起来作为Top N文档，这样可以避免大量网页之间的相互比较。

  当然前提是网页的标签能很好地体现网页内容所涉及的主题