HW2

PB18111764 钟溯飏

1.计算题

1.1

1)$W_{t,d}=(1+\log tf_{t,d})\log\frac{N}{df_t}$

e.g.: $W_{car,doc1}=(1+\log 34)\times\log\frac{811400}{18871}=4.135$

	tf-idf@doc1	Tf-idf@doc2	tf-idf@doc3
Car	4.135	3.109	4.092
Auto	3.476	5.601	0
Insurance	0	4.404	2.892
Best	2.947	0	2.763

2)

$$\mathbf{doc1}=\frac{\overrightarrow{W}_{t,d}}{|W_{t,d}|}=(\frac{4.135}{\sqrt{37.86561}},\ \frac{3.476}{\sqrt{37.86561}},\ 0,\ \frac{2.947}{\sqrt{37.86561}})=(0.672,\ 0.565,\ 0,\ 0.478)$$

$$\mathbf{doc2}=(0.400,\ 0.720,\ 0.567,\ 0)$$

$$\mathbf{doc3}=(0.715,\ 0,\ 0.505,\ 0.483)$$

3) 已经归一化：使用余弦相似度进行相似度计算：$\cos(\vec{q},\vec{d})=\sum_{i=1}^{|V|}q_i d_i$

$$\mathbf{doc1_{score}}=0.672$$

$$\mathbf{doc2_{score}}=0.400+0.567=0.967$$

$$\mathbf{doc3_{score}}=0.715+0.505=1.220$$

1.2

1)

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

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

跳转矩阵$P=d\mathbf{R}+[(1-d)/N]\mathbf{I}$

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

2) $M'=PM$

$$\begin{aligned}M&=PM\\&=\begin{pmatrix}0.0214&0.0214&0.3048&0.0214&0.0214&0.0214&0.0214\\0.0214&0.4464&0.0214&0.0214&0.0214&0.0214&0.0214\\0.8714&0.4464&0.3048&0.0214&0.0214&0.0214&0.0214\\0.0214&0.0214&0.3048&0.4464&0.0214&0.0214&0.3048\\0.0214&0.0214&0.0214&0.4464&0.0214&0.0214&0.3048\\0.0214&0.0214&0.0214&0.0214&0.0214&0.4464&0.0214\\0.0214&0.0214&0.0214&0.0214&0.8714&0.4464&0.3048\end{pmatrix}\begin{pmatrix}1/7\\1/7\\1/7\\1/7\\1/7\\1/7\\1/7\end{pmatrix}=\begin{pmatrix}0.0619\\0.0821\\0.2440\\0.1631\\0.1226\\0.0821\\0.2440\end{pmatrix}\end{aligned}$$

不断迭代，收敛为

$M=[0.0545,\ 0.0373,\ 0.1166,\ 0.2431,\ 0.2101,\ 0.0373,\ 0.3012]^T$

3)更新后的邻接矩阵为M

$$M=\begin{pmatrix}0&0&1&0&0&0&0\\0&1&1&0&0&0&0\\1&0&1&0&0&0&0\\0&0&0&1&1&0&0\\0&0&0&0&0&0&1\\0&0&0&0&0&1&1\\0&0&0&0&1&0&1\end{pmatrix}$$

采用迭代式$a_{k+1}=M^Th_k$，$h_{k+1}=Ma_{k+1}$，每次迭代后进行欧式归一化。

使用python(不知道可不可以)，迭代了100次，最终得到

$$a = \begin{pmatrix} 0.001 \\ 0.001 \\ 0.002 \\ 0.168 \\ 0.498 \\ 0.272 \\ 0.805 \end{pmatrix} \qquad h = \begin{pmatrix} 0.001 \\ 0.002 \\ 0.002 \\ 0.335 \\ 0.405 \\ 0.542 \\ 0.655 \end{pmatrix}$$

1.3

a. P@10=4/10=0.2　P@20=7/20=0.14

b. R@10=3/10=0.3　R@20=7/10=0.7

F1@10=2PR/(P+R)=0.24　F1@20=0.2333

c. AP=(1/1+2/2+3/5+4/9+5/11+6/15+7/20)/5=0.7499

d. 最大的时候也就是剩下的文档全部在前面，即
AP=(1/1+2/2+3/5+4/9+5/11+6/15+7/20+8/21+9/22+10/23)/10=0.6474

e. 最小的时候也就是剩下的文档全部在最后，即
AP=(1/1+2/2+3/5+4/9+5/11+6/15+7/20+8/9998+9/9999+10/10000)/10=0.5252

2.问答题

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?

1. 结合陀螺仪和手机定位信息可以获取用户所在位置。比如陀螺仪检测人不在运动而定位在运动，那么可以推断出用户在乘坐交通工具；如果这两者信息很久都不更新，且定位信息在晚上经常停留在某个位置，可以推断出用户在家里，或是在办公室（白天）。以此类推。通过对此两者的建模，可以构建出用户一天的行为轨迹，以及用户所处的当前行为状态。
2. 根据手机的噪声监测功能（不是直接请求用户麦克风）判断用户所处状态，可结合前者推断出用户所乘坐交通工具（安静是出租车或自驾，吵是公共交通），办公还是学习，等。
3. 根据手机的光线检测功能（不是调用摄像头）及时间判断用户所处位置，如早上但是光线暗，说明是室内，晚上光线强，也是在室内。
4. 不直接调用麦克风或是摄像头是因为这些涉及用户隐私，可能更多用户会选择关闭。

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. 首先收集志愿者数据，对数据进行人工标注，比如对用户行走的时候的某段时间的陀螺仪和定位信息的相对改变进行标注（<行走，室外，寝室-教室路上>，<公交车，室内，家-办公室路上>），并以此建立训练集，在前期先对模型进行预训练，产生一个囊括基本信息的模型。
2. 将基本信息矩阵建立某种映射形成向量。通过对各个用户不同的个性化行为模式，在数据库中建立其独有模型，并通过数据采集对其进行微调。

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. Markov性质：对任何一列状态$i_0, i_1, \cdots, i_{n-1}, i, j$，及对任何$n \geq 0$，随机过程$\{X_n, n \geq 0\}$满足Markov性质：$P\{X_{n+1} = j \mid X_0 = i_0, \cdots, X_{n-1}, X_n = i\} = P\{X_{n+1} = j \mid X_n = i\}$，则$X_n$为Markov链；即，之前的“后退”与现状态的“后退”无关，是独立的。
2. 据此，我认为不可以用马尔科夫链进行建模。
    - 用户点击后退按钮是想要查看自己记忆中的某个特定时间点的内容，而点击多少次后退与自己在此之后看了多少个网页有关，并不是独立的，所以不是马尔科夫的。
3. 但是，如果不包括连续后退，在某些场景下可以勉强看作马尔科夫链：
    - 用户选择单次后退，往往是在本页面上完成了某次行为或者在子页面上看完了某个内容。在这点上，是否后退只取决于本页面。
    - 假设用户的记忆很好，不会突然想起自己是不是漏了什么结果。那么他在使用搜索引擎时，可看成一个马尔科夫链：点进搜索结果<->返回搜索界面，即，本页面上没找到自己想看的东西，后退；找到了，结束。

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?

1. 在排序算法中引入考虑多样性的指标。在获得初始的相关性排序结果(如pagerank)基础上，直接惩罚相似度高的结果，使排序结果的前面保持一定的差异性，从而得到具有多样性的排序结果。最后的评价指标即为文档与用户查询的**相关度**与文档间的**差异度**。以此作为迭代方式，不断迭代，并放入已选文档集合，即可获得一个既有相关性也有差异性的检索集合。
2. 这样的算法会导致需要对查询文档间一一进行计算，如果查询文档规模为n，将是一个n的全连接图，效率不高。
    - 可以考虑引入主题敏感排序，在判断时，不必与所有的网页判断相似度，而是与各个主题相比判断相似度，最终能保证获得的结果中包含有不同的主题。
    - 使用一些模型：Relational Learning To Rank, R-LTR、α-NDCG、ERR-IA等。