web hw2

PB21051012

1.

1.1

1)

对于词项Car：

TF-IDF@Doc1 = 34 * log(811,400 / 18,871) ≈ 2.5267

TF-IDF@Doc2 = 8 * log(811,400 / 18,871) ≈ 1.0107

TF-IDF@Doc3 = 32 * log(811,400 / 18,871) ≈ 4.0429

对于词项Auto：

TF-IDF@Doc1 = 3 * log(811,400 / 3,597) ≈ 8.6252

TF-IDF@Doc2 = 24 * log(811,400 / 3,597) ≈ 68.7515

TF-IDF@Doc3 = 0

对于词项Insurance：

TF-IDF@Doc1 = 0

TF-IDF@Doc2 = 51 * log(811,400 / 19,167) ≈ 74.9077

TF-IDF@Doc3 = 6 * log(811,400 / 19,167) ≈ 8.9889

对于词项Best：

TF-IDF@Doc1 = 18 * log(811,400 / 40,014) ≈ 6.1035

TF-IDF@Doc2 = 0

TF-IDF@Doc3 = 13 * log(811,400 / 40,014) ≈ 4.4181

2)

Doc1 = (0.233857 0.798300 0 0.564906)

Doc2 = (0.009940 0.676151 0.736696 0)

Doc3 = (0.374303 0 0.832216 0.409040)

3)

余弦相似度 = 查询向量 * 文档向量 / (|查询向量| * |文档向量|)

与Doc1的余弦相似度为:0.1653868

与Doc2的余弦相似度为:0.5280311

与Doc3的余弦相似度为:0.8532666

1.2

1)

(1-d)/3 = 0.15

d = 0.55

转移概率矩阵为 P = (0 1/2 1/2)　　(0 0 1)
　　(0 1 0)

2)

PageRank值:计算一次就收敛了，结果为 (0.15 0.425 0.425)

Hub值,Authority值:

邻接矩阵M　(0 1 1)

　　(0 0 1)

　　(0 1 0)

a0 = (1 1 1) h0 = (1 1 1)

a1 = (0 1/2 1/2) h1 = (1/3 1/3 1/3)

a2 = (0 1/2 1/2) h2 = (1/3 1/3 1/3)

收敛

Hub值为(1/3 1/3 1/3)

Authority值为(0 1/2 1/2)

1.3

a)

P@10 = 5 / 10 = 0.5

P@20 = 7 / 20 = 0.35

b)

如果相关文档数小于N，P@N的理论上限必定小于 1

同理，可得R@N，R@N的理论上限必定小于 1

Precision (P) 是指检索结果中相关文档的比例，所以它的理论上限是文档集中相关文档的比例。如果文档集中只有 10 个相关文档，那么 Precision 的理论上限就是 10 / 10 = 1。如果所有检索结果都是相关文档，那么 P 可以达到 1，但通常情况下，不可能所有结果都是相关的。

Recall (R) 是指在所有相关文档中，被检索出来的比例，所以它的理论上限是在所有相关文档都被检索到的情况下，R 可以达到 1。但在实际情况下，往往有一些相关文档被漏掉，导致 R 小于 1。

c)

F1 = 2 * (P * R) / (P + R)

对于前 10 篇文档：P@10 = 5 / 10 = 0.5R@10 = 5 / 10 = 0.5

F1@10 = 2 * (0.5 * 0.5) / (0.5 + 0.5) = 0.5

对于前 20 篇文档：P@20 = 7 / 20 = 0.35R@20 = 7 / 10 = 0.7

F1@20 = 2 * (0.35 * 0.7) / (0.35 + 0.7) ≈ 0.4667

d)

AP = (1/1 + 2/2 + 3/5 + 4/9 + 5/11 + 6/15 + 7/20) /7=0.606999

2

2.1

a) 用户留下的以下信息可能有助于我们判断“反话正说”的现象：

1. 上下文信息：查看用户评论的上下文，包括产品或服务的性质、前一条评论、用户的历史行为等，可以帮助判断用户的真实意图。

2. 表情符号和情感标识：用户在评论中使用的表情符号和情感标识（如笑脸、愤怒符号等）可能揭示了他们真正的情感。如果用户在评论中表达不满，但同时使用了笑脸，这可能是“反话正说”的迹象。

3. 逻辑矛盾：查看评论中是否存在逻辑上的矛盾，例如用户可能会表达满意但在评论中提到了负面的方面。这种矛盾可能是“反话正说”的信号。

4. 情感强度：用户评论中的情感强度可以提供线索。如果评论中出现强烈的负面情感词汇，但整体评价却是正面的，这也可能是“反话正说”。

5. 反讽和幽默：用户有时会使用反讽和幽默来表达不满。如果评论中包含了这些元素，需要谨慎分析其真实意图。

b)

1. 识别情感标识符：建立一个情感标识符的列表，包括常见的表情符号和情感词汇。对用户评论进行文本分析，检查是否存在这些标识符。如果评论中包含相矛盾的情感标识符，可以标记该评论以进行进一步分析。

2. 矛盾检测：开发简单的规则来检测评论中的逻辑矛盾，例如正面评价但包含负面细节。这些规则可以用于自动筛选潜在的“反话正说”评论。

3. 情感强度分析：使用基本的情感分析工具来检测评论中的情感强度。如果评论中存在负面情感词汇，但整体评价是正面的，可以将其标记为潜在的“反话正说”评论。

4. 用户历史数据：考虑用户的历史评论和行为。如果用户过去通常以积极的方式评价产品或服务，但最近的评论却包含了负面情感，这可能表明“反话正说”现象。

2.2

不能，马尔科夫链的下一状态只与当前状态有关，回退行为需要记录之前的状态，两者矛盾。

2.3

a)多重排名策略,使用多样化特征,修改优化目标，直接在将多样化指标写入目标函数，扩大召回率，增加精排、重排模块

b)需要。确保全面性：即使用户通过点击行为提供了具体的意图，他们可能仍然希望看到更多相关内容，以确保他们没有错过其他相关信息。多样性可以确保提供全面的信息，满足用户的需求。避免过度特化，如果仅根据用户的具体意图提供结果，可能导致结果过于特化，而缺乏信息的多样性。这可能不利于用户获取更广泛的知识和视角。新颖性，用户可能对已经看到的内容感到厌倦，他们希望看到新颖的结果。多样性可以帮助引入新的、不同的内容，提供新颖性。

2.4

可以减少需要比较的样本数量。

样本对的筛选可以采用负样本采样策略，减少生成的负样本对的数量。同时，为样本对引入重要性权重，使一些样本对的比较更加重要，而另一些则可以降低权重。这可以通过对每个样本对分配权重来实现，以确保关键样本对的比较不被忽略，以此抑制对于排序精度的干扰。