金庸的江湖——金庸武侠小说中的人物关系挖掘

组长：韩畅，组员：李展烁、王一之、闫旭芃

2020 年 7 月 29 日

1 实验规划与设计

1.1 任务分配

171860551, 韩畅：组长, 算法设计与实验规划，任务六
171860550, 王一之：算法设计与实验规划, 任务四
171860549, 闫旭芃：算法设计与实验规划，任务五
171840565, 李展烁：算法设计与实验规划, 任务三

1.2 任务要求

1.3 设计思路

2 实验实现

2.1 任务一

todo

2.2 任务二

todo

2.3 任务三

todo

2.4 任务四: 基于人物关系图的 PageRank 计算

2.4.1 PageRank 算法介绍

PageRank，又称网页排名，名字源于 google 创始人之一的 Larry Page，是 Google 公司所使用的对与网页重要性排序的算法。
PageRank 通过网页之间的超链接评价网页重要性，它的基本思想是：

1) 如果一个网页被多个网页所指向，则该网页比较重要

2) 如果一个重要的网页指向另一个网页，则另一个网页也比较重要

该算法模拟一个上网者，随机打开一个网页，之后随机点击该网页的链接，统计上网者分布在每个网页的概率。
最初，每个网页的概率均等，每次跳转时，网页 X 将其 PR(PageRank) 均分到所指向的所有页面，记链接数为 L(X), 于是，经过一次跳转后：

$$PR(A) = \frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + ...$$

我们将每个网页抽象成一个节点，超链接抽象为有向边，共同构成一个图。则每次跳转可视为所有页面 PR 构成的特征向量 R 与该图的出度表矩阵相乘，即：

$$R = \begin{bmatrix} PR(p_1) \\ PR(p_2) \\ \vdots \\ PR(p_n) \end{bmatrix} \quad M = \begin{bmatrix} p_1 \to p_1 & p_2 \to p_1 & \cdots & p_n \to p_1 \\ p_1 \to p_2 & p_2 \to p_2 & \cdots & p_n \to p_2 \\ \vdots & \vdots & \ddots & \vdots \\ p_1 \to p_n & p_2 \to p_n & \cdots & p_n \to p_n \end{bmatrix}$$

$$R_1 = MR_0$$

2.4.2 设计思路

任务四的输入为任务三的输出，格式如下：
人物 [名字 $_1$, 影响 $_1$| 名字 $_2$, 影响 $_2$|...| 名字 $_n$, 影响 $_n$]
影响 $_i$ 为名字 $_i$ 与该人物名字归一化后的同现次数，表示名字 $_i$ 对该人物的影响权重。

2.5 任务五

todo

2.6 任务六

todo

3 实验经验总结与改进方向

1) todo

2) todo