Harvard CS50 Project ——degree

孙以琳 19211341

指导老师：陈一帅

摘要

根据广度优先算法研讨Harvard CS50 Project 求degree问题。使用python编写代码，实现功能。因为我之前没有学习过python，所以先学习了一下python基础语法。

*关键词—*广度优先算法、python

# 题目介绍

编写一个程序，确定两个参与者之间的“分离程度”是多少。

题目背景是根据[《凯文·培根的六度](https://en.wikipedia.org/wiki/Six_Degrees_of_Kevin_Bacon) 》游戏，好莱坞电影业的任何人都可以在六个步骤内与凯文·培根建立联系，每一步都包括找到一部两位演员都出演的电影。

在这个问题上，是通过选择一系列连接他们的电影来找到任何两个演员之间的最短路径。例如，詹妮弗·劳伦斯和汤姆·汉克斯之间的最短路径是2：詹妮弗·劳伦斯通过出演《X战警：头等舱》与凯文·培根相连，凯文·培根通过两部主演《阿波罗13号》与汤姆·汉克斯的联系。

我们可以把它归结为一个搜索问题：我们的状态是人。我们的行为是电影，它们将我们从一个演员带到另一个演员。我们的初始状态和目标状态由我们尝试连接的两个人定义。通过使用广度优先搜索，我们可以找到从一个演员到另一个演的最短路径。

# 代码解答

首先了解一下给出的代码。每个数据集由三个 CSV 文件组成。分别是people.csv、star.csv、movie.csv,在people.csv每个人都有一个唯一的id，movie.csv中每部电影还有一个唯一id，star.csv中建立了people.csv和movie.csv的关系。每一行都是一对person\_id的值和movie\_id的值。

在degree.py中定义了几个数据结构来存储CSV文件中的信息。并用几个dictionary将姓名，人物，电影id分别映射出来。Name dictionary是一种按姓名查找人员的方法，它将姓名映射到一组相应的id。load\_data函数将数据从 CSV 文件加载到这些数据结构中。此程序中的 main 函数首先将数据加载到内存中。然后，该函数提示用户键入两个名称。person\_id\_for\_name函数检索任何人的 id（并在多个人具有相同姓名的情况下处理提示用户进行澄清）。然后，该函数调用shortest\_path函数来计算两个人之间的最短路径，并打印出该路径。

我写的最短路径算法用的是广度优先算法。广度优先搜索算法将同时跟踪多个⽅向，在每个可能的⽅向上走一步，然后在每个⽅向上走第二步。边界作为队列数据结构进⾏管理。根据 “先进先出”的原则，所有新节点都按顺序相加，并且根据先添加的节点来考虑节点。

首先定义从边界删除一个节点并返回它的函数。把根节点放到队列的末尾。

1. 将根节点加入到Queue中
2. 使用while循环，当Queue为空时，结束循环。
3. 将Queue中的节点依次搜索出来检查。
4. 如果没有找到要找的值，就将搜索出来的节点的子节点再加入到Queue中（如果没有子节点就不加）

# 心得收获

## Python安装与配置

Python没有学过，从安装开始就很艰难，按完之后发现没有安装python编译器，只安装了PyCharm。我学习到的东西比较基础。

我在使用pomegranate时下载这个软件包，通过命令提示符安装软件，但是由于没有安装VS所以一直安装不成功，安装成功后，由于下载安装路径python没有之间按照指定路径安装，编译器环境没有编辑上去，所以还是不可以。这里告诉我安装软件的时候要注意编译器的环境路径是否正确。

## 编译问题

运行的时候发现python对于缩进要求非常严格，缩进不对将不能使用，所以我上网百度了一下，怎么解决缩进问题。在出现错误的那一行，按空格或Tab（但不能混用）键缩进就行。该缩进的地方就要缩进，不进行缩进反而会出错。1. Ctrl + A全选代码。2. Code -> Auto-Indent Lines自动缩进快捷键 Ctrl + Alt + I

## 学习收获

在B站上学习了python的一些基本语法，包括数据类型，结构，函数和类。并联系了一些比较简单的代码。

通过大作业，我首先把老师课件上的代码运行了一下，主要是概率模型那一章。通过学习我发现python语法和C在意义上是相同的，但是python更加简洁明了。比如创建节点为每个节点提供概率，用条件概率表直接类似矩阵形式写出来，非常简洁明了。然后创建模型，引用贝叶斯模型，增加节点到节点形成边。这里model模型的就成功创建了。求某个事件的概率就直接可以用model.probability求出来。然后学习了建立采样模型，python语言就直接用sample={}将随机变量名映射到生成的样本，parents = {}将分布映射到生成的样本，然后假设拓扑顺序，遍历所有状态，使用for循环，嵌套if else模型，在父映射中跟踪采样值。然后学习了for函数的写法，重复采样1000次，发现python语言和C在for写法上python直接写变量 循环次数。课件中还有马尔可夫模型和隐马尔可夫模型，模型建立方法和贝叶斯模型建立方法相似，model.\*\*的形式。

练习了cs50中搜索单元中project求degree问题。文件里给出了一个代码将将数据从 CSV 文件加载到这些数据结构中。需要写最短路径算法，我采用了⼴度优先搜索算法。⼴度优先搜索算法将同时跟踪多个⽅向，在每个可能的⽅向上⾛⼀步，然后在每个⽅向上⾛第⼆步。在这种情况下，边界作为队列数据结构进⾏管理。“先进先出”，所有新节点都按顺序相加，并且根据先添加的节点来考虑节点但是由于我不太会把变量路径带进去，不知道该选取哪个变量，所以不太会。我看了下给出的其他代码，学习了将CSV文件引入数据结构中，了解了dictionary将姓名映射到对应的id的算法编程。

通过大作业学习，我认识到我应该学习python了，对python语言认识不足导致不太会写代码。Havard cs50的作业比我之前做的一些研讨以及作业相比起来应用性更强，对于题干因素的分析难度更高一些。