

**课 程 实 验 报 告**

**课程名称： 大数据分析**

**专业班级： CS1908班**

**学 号： U201915160**

**姓 名： 陈千鹤**

**指导教师： 杨驰**

**报告日期： 2021.12.10**

**计算机科学与技术学院**

**目录**

[实验二 PageRank算法及其实现 1](#_Toc90027732)

[**2.1实验目的** 1](#_Toc90027733)

[**2.2 实验内容** 1](#_Toc90027734)

[**2.3 实验过程** 1](#_Toc90027735)

[2.3.1 编程思路 1](#_Toc90027736)

[2.3.2 遇到的问题及解决方式 2](#_Toc90027737)

[2.3.3 实验测试与结果分析 2](#_Toc90027738)

[**2.4 实验总结** 3](#_Toc90027739)

# 实验二 PageRank算法及其实现

## **2.1实验目的**

1、学习pagerank算法并熟悉其推导过程；

2、实现pagerank算法，理解阻尼系数的作用；

3、将pagerank算法运用于实际，并对结果进行分析。

## **2.2 实验内容**

提供的数据集包含邮件内容（Emails.csv），人名与id映射（Persons.csv），别名信息（Aliases.csv），Emails文件中只考虑MetadataTo和MetadataFrom两列，分别表示收件人和寄件人的名称，但这些名称包含许多别名，因此需要对邮件中的名称进行统一并映射到唯一id。（提供预处理代码preprocess.py参考）。

完成这些后，由寄件人和收件人为节点构造有向图，不考虑重复边，编写pagerank算法的代码，根据每个节点的入度计算其pagerank值，迭代直到误差小于10-8。

输出人名id及其对应的pagerank值。

加分项：加入teleport β，用以对概率转移矩阵进行修正，解决dead ends和spider trap的问题。

## **2.3 实验过程**

### 2.3.1 编程思路

1. 实现数据处理：根据别名信息和邮件往来名称获得邮件往来用户的唯一id元祖。

2. 数据预处理：根据往来信息清洗重复的边，同时统计节点，并将节点映射为连续数字表示。

3. 生成邻接矩阵EM：根据边生成领接矩阵

4. 初始化M矩阵：根据每个节点的出度，初始化EM矩阵数值

5. 初始化PageRank矩阵

6. 开始PageRank矩阵迭代：为了解决spider trap和dead end引入teleport

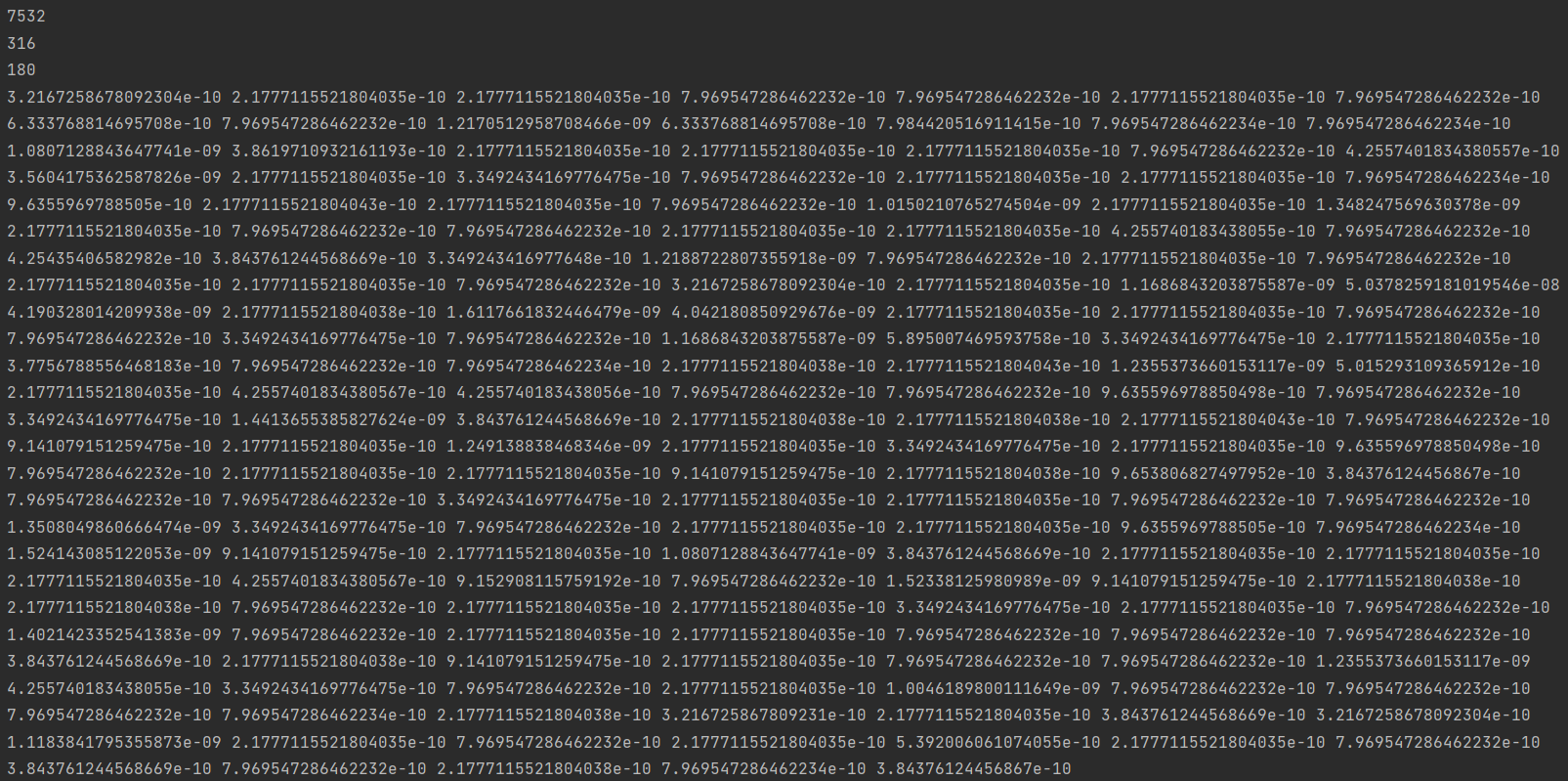
### 2.3.2 遇到的问题及解决方式

1. 使用numpy的array时，无法通过(==0)直接判断一个元素是否等于0

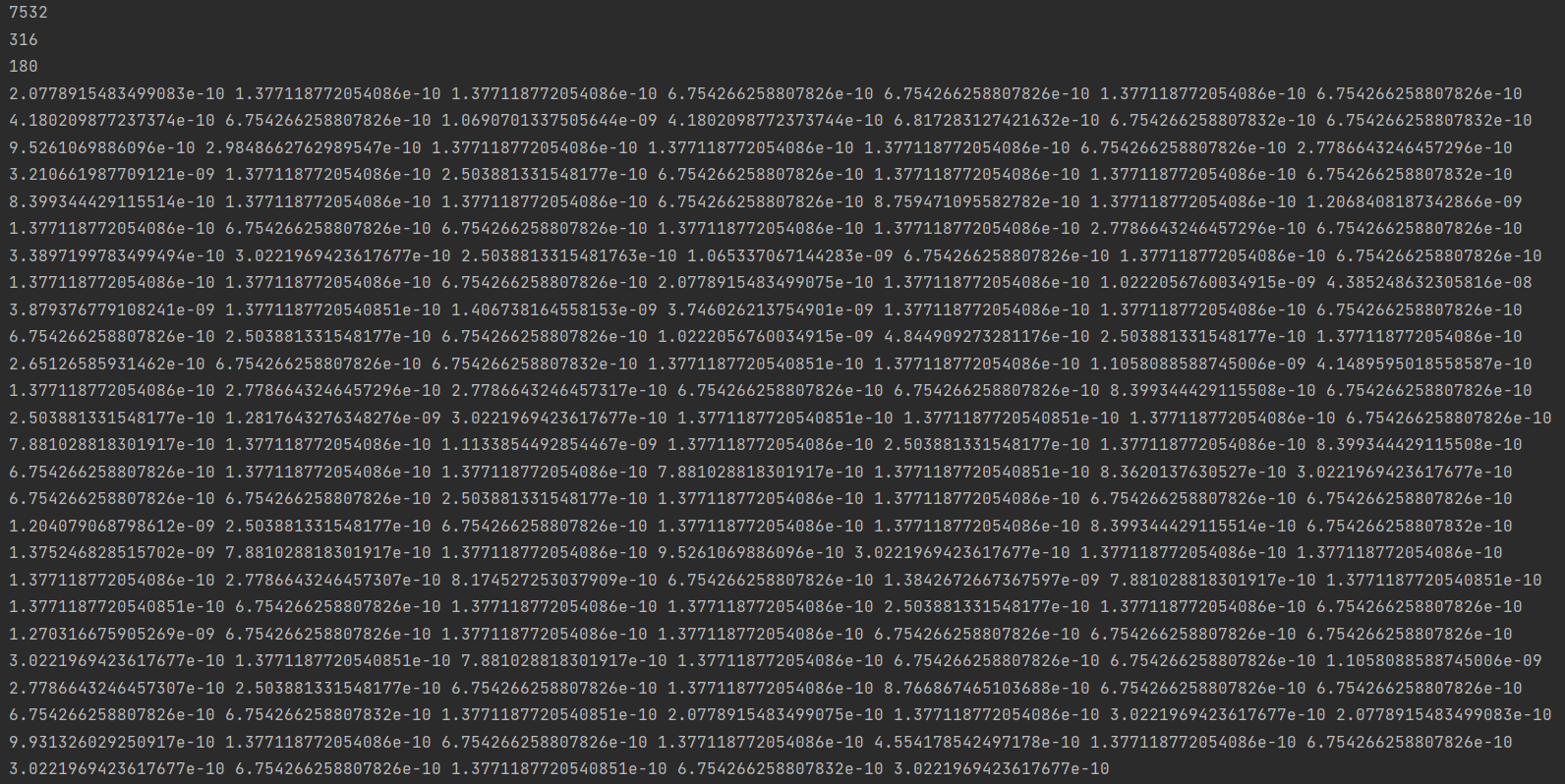
解决方法：将是否等于0的判断条件换位是否大于0

### 2.3.3 实验测试与结果分析

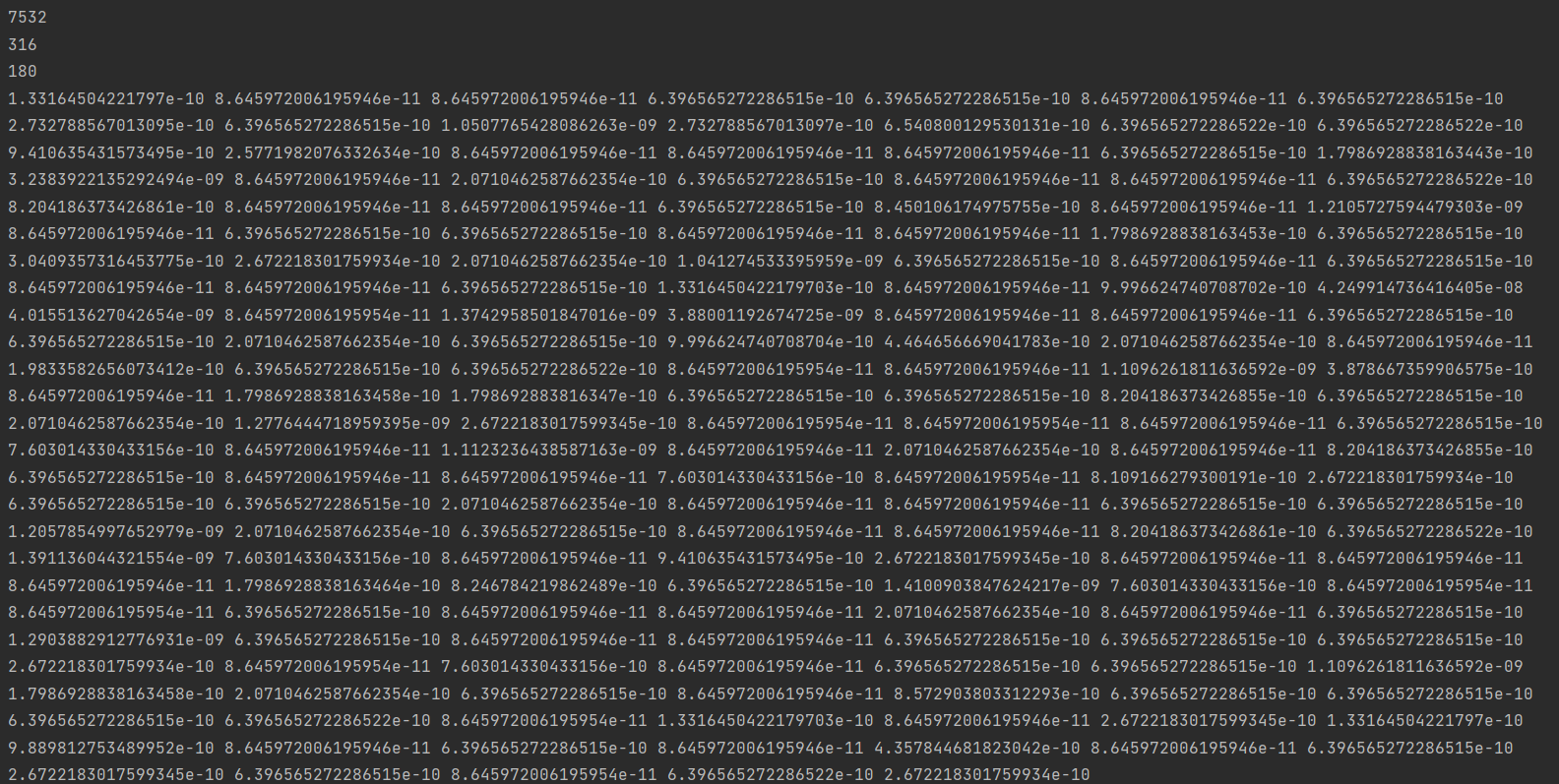
1. 取阿尔法等于0.8，生成结果如下图



2. 取阿尔法等于0.85，生成结果如下图



3. 取阿尔法等于0.9，生成结果如下图



## **2.4 实验总结**

1. 了解了pagerank算法及其推导方式

2. 学习了阻尼系数的作用

3. 亲自实现了pagerank算法，了解了实现的细节

4. 学习了ragerank算法存在的问题及其解决方法

5. 实现了升级后的pagerank算法