

**课 程 实 验 报 告**

**课程名称： 大数据分析**

**专业班级：**

**学 号：**

**姓 名：**

**指导教师：**

**报告日期：**

**计算机科学与技术学院**

**目录**

[实验二 PageRank算法及其实现 1](#_Toc57042479)

[**1.1实验目的** 1](#_Toc57042480)

[**1.2 实验内容** 1](#_Toc57042481)

[**1.3 实验过程** 1](#_Toc57042482)

[1.3.1 编程思路 1](#_Toc57042483)

[1.3.2 遇到的问题及解决方式 1](#_Toc57042484)

[1.3.3 实验测试与结果分析 1](#_Toc57042485)

[**1.4 实验总结** 2](#_Toc57042486)

# 实验二 PageRank算法及其实现

## **1.1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。

## **1.2 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姓名，但这些姓名包含许多别名，思考如何对邮件中人名进行统一并映射到唯一id？（提供预处理代码preprocess.py以供参考）。

完成这些后，即可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

实验进阶版考虑加入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。

输出人名id及其对应的pagerank值。

## **1.3 实验过程**

### 1.3.1 编程思路

### 1.3.2 遇到的问题及解决方式

### 1.3.3 实验测试与结果分析

## **1.4 实验总结**