实验二 PageRank 算法及其实现

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β的作用；
3、将 PageRank 算法运用于实际，并对结果进行分析。

实验内容

提供 PageRank-data.zip，其中包括数据集邮件内容（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？（PageRank-data.zip 中还提供预处理代码 preprocess.py，以及处理后的 sent_receive.csv 数据以供参考）。

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，实现如下：

1) 不考虑 teleport，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。
2) 在功能 1）的基础上，加入 teleport β=0.8，用以对概率转移矩阵进行修正，解决 dead ends 和 spider trap 的问题，迭代直到误差小于 10^{-8}。