Compute the PageRank scores on the given dataset

- Dataset: Data.txt
- The format of the lines in the file is as follow:

  FromNodeID ToNodeID

- In this project, you need to report the **Top 100 NodeID** with their PageRank scores. You can choose different parameters, such as the teleport parameter, to compare different results.
- One result you must report is that when setting the teleport parameter to 0.85.
- In addition to the basic PageRank algorithm, you need **to optimize your memory use** via Block Matrix, Sparse Matrix and some other approches.
- Deadline: 2025.4.30

计算给定数据集的 PageRank 分数

- 数据集：Data.txt
- 文件中各行的格式如下：

  从节点 ID 到节点 ID
- 在这个项目中，您需要报告前 100 个 NodeID 及其 PageRank 得分。您可以选择不同的参数（例如 teleport 参数）来比较不同的结果。

- 您必须报告的一个结果是，将传送参数设置为 0.85 时。

- 除了基本的 PageRank 算法之外，您还需要通过块矩阵、稀疏矩阵和其他一些方法来优化内存使用。
- 截止日期：2025.4.30

Compute the PageRank scores on the given dataset

Specific Requirement:

- Language: You can use C/C++/Python.
- For C/C++, It's recommended to code in a single file and minimize the use of third-party libraries. You'd better to use gcc or g++ for compiling your code and give the compilation parameters in *compile-parameter.txt* with the following format.

  g++ -fopenmp PageRank.cpp -o PageRank.exe

- For Python, It's recommended to use commonly-used libraries or package (e.g. numpy and scipy is accepted). To confirm the availability of a package, you can email *bigdatacomputing@163.com*. It's recommended to code in a single file. Ensure the entry of your code should be named as 'main.py'.

计算给定数据集的 PageRank 分数

具体要求：

- 语言：您可以使用 C/C++/Python。
- 对于 C/C++，建议在单个文件中编写代码并尽量减少使用第三方库。最好使用 gcc 或 g++ 来编译代码，并在 compile-parameter.txt 中给出编译参数，格式如下。

```
g++ -fopenmp PageRank.cpp -o PageRank.exe
```

- 对于 Python，建议使用常用的库或包（例如 numpy 和 scipy 都可以）。如需确认包的可用性，您可以发送电子邮件至 bigdatacomputing@163.com。建议在单个文件中编写代码。请确保代码的入口命名为 'main.py'。

Compute the PageRank scores on the given dataset

Specific Requirement:

- Consider of dead-ends and spider-traps
- Optimize your memory use **as much as possible.**
- You can optimize the memory use in other ways**, but Block Matrix and Sparse Matrix(**https://zhuanlan.zhihu.com/p/557231877 **) optimization is compulsorily required**
- Your program is required to iterate until reaching convergence
- Using existing PageRank API is prohibited. (e.g. *networkx.pagerank* in Python)
- Your program need to read *Data.txt*, conduct PageRank and Return the Top-100 nodes in *Res.txt* with its score as following format:

  NodeID Score

计算给定数据集的 PageRank 分数

具体要求：

- 考虑死胡同和蜘蛛陷阱
- 尽可能优化你的内存使用。
- 您可以通过其他方式优化内存使用，但块矩阵和稀疏矩阵(https://zhuanlan.zhihu.com/p/557231877)优化是必需的

- 你的程序需要迭代直到达到收敛
- 禁止使用现有的 PageRank API。（例如 Python 中的 networkx.pagerank）

- 你的程序需要读取 Data.txt，进行 PageRank 计算，并返回 Res.txt 中的 Top-100 个节点及其分数，格式如下：

  节点 ID 分数

Compute the PageRank scores on the given dataset

Specific Requirement:

- You'd better make sure the maximum memory use during the whole life of your program should be lower than **80MB.**
- You can't to sacrifice too much time performance to minimize your memory usage. The program need to complete its runtime under **60s**.
- Maybe you can use some Parallel or/and Distributed Computing techniques. **But it's not the main content of this assignment**.

计算给定数据集的 PageRank 分数

具体要求：

- 您最好确保程序整个生命周期内的最大内存使用量低于 80MB。

- 你不能牺牲太多的时间性能来最小化你的内存使用量。程序需要在 60 秒内完成运行。
- 也许你可以使用一些并行或分布式计算技术。但这不是本次作业的主要内容。

Compute the PageRank scores on the given dataset

Submitting:

- *bigdatacomputing@163.com*
- *Report:* Include but not limited: description of dataset, key code details, **how to optimize the memory usage** and result with analysis
- *Code:* Meet the above requirements, Input Dataset is not required
- *Result:* Named as *Res.txt* and Meet the above requirements
- *Executable File:* Compile with all Static link library.
  - For C/C++, maybe you can use [-*static*] setting in gcc/g++
  - For Python, using third-party packages to generate and integrate the package used in your code.
  - **Ensure your code can run in other computers (and maybe other OS)**. (You can ask LLMs or TAs for more).

计算给定数据集的 PageRank 分数

提交中：

- bigdatacomputing@163.com
- 报告：包括但不限于：数据集描述、关键代码详细信息、如何优化内存使用情况及结果分析
- 代码：满足以上要求，输入数据集不作要求
- 结果：命名为 Res.txt 并满足上述要求
- 可执行文件：使用所有静态链接库进行编译。
  - 对于 C/C++，也许您可以在 gcc/g++ 中使用 [-static] 设置
  - 对于 Python，使用第三方包来生成并集成代码中使用的包。

  - 确保您的代码可以在其他计算机（也可能是其他操作系统）上运行。（您可以向 LLMs 或 TA 询问更多信息）。

Compute the PageRank scores on the given dataset

Submitting:

- Indicate team division of work and contributions.
- If no team division indicated, whole group will be given the same score.
- Compress your submission (zip format), with following naming:

  2000000_小张_2111111_小王_2222222_小林_第一次作业.zip

- Deadline: 2025.4.30
- **Failure to comply with the format will result in failure to grade.**
- **Code plagiarism and academic fraud are absolutely intolerable.**

计算给定数据集的 PageRank 分数

提交中：

- 表明团队分工和贡献。
- 若未注明团队分组，则整个小组将获得相同的分数。
- 压缩您的提交（zip 格式），并采用以下命名：

  2000000_小张_2111111_小王_2222222_小林_第一次作业.zip
- 截止日期：2025.4.30
- 不遵守格式将导致评分失败。
- 代码抄袭和学术造假是绝对不能容忍的。