实验一

使用局部敏感哈希进行相似性搜索

实验描述：

数据集是表示研究人员之间共同作者关系的无向图。

需要构建一个LSH方案，以进行相似性搜索：对于任何查询节点，应该找到与查询节点邻居集合的Jaccard相似度得分最高的前10个节点（不包括查询节点本身）。

在最终报告中，请说明相似性搜索的性能，如准确性、索引时间、查询时间和空间使用率等。

实验步骤

- 数据集预处理，数据集是一个无向图，每一行表示两个节点之间存在边。
- 定义MinHash哈希函数族，它包括多个不同的哈希函数。每个哈希函数都是一个随机置换。
- 对于每个节点，我们使用MinHash哈希函数族计算其邻居集合的签名向量。
- 将节点的签名向量分成b个分组，每个分组包含r行。将相同分组的节点映射到同一个桶中。
- 将查询节点对应桶中的节点添加到一个候选集中，并返回候选集中相似度最高的前k个节点。
- 评估LSH算法的性能，我们可以记录算法的索引时间、查询时间、空间使用和准确度等指标，并根据不同的参数设置进行比较和分析。

补充材料

- A Survey on Locality Sensitive Hashing Algorithms and their Applications
- Multi-Probe LSH: Efficient Indexing for High-Dimensional Similarity Search （VLDB-07）
- Modeling LSH for Performance Tuning （CIKM-08）
- Quality and Efficiency in High Dimensional Nearest Neighbor Search （SIGMOD-09）
- Bi-level Locality Sensitive Hashing for k-NearestNeighborComputation （ICDE-12）