Linked Papers

本次作业需要同学们使用前后端技术、数据库技术、大数据算法等搭建一个论文检索平台 Linked Papers，该系统需在 **隐私保护**、**访问控制**、**系统性能** 这三个方面有所体现。可参考的论文检索平台网站：Connected Papers、Semantic Scholar。

数据描述

本次作业提取了 169,343 篇 arXiv 上计算机科学相关的论文，这些论文涉及到 40 个子领域。每篇论文提供了标题与摘要，以及一个 128 维的特征向量表示，该向量通过对标题和摘要中所有单词的 skip-gram embedding 进行平均而获得。我们也提供了论文间的引用关系，论文发表的年份。数据集中的论文按年份分为训练集、验证集和测试集：2017年及其之前的论文作为**训练集**，2018年的论文作为**验证集**，2019年及其之后的论文作为**测试集**。我们提供了训练集和验证集论文的标签，但未提供**测试集**的标签（需要自己实现类似于作业二中的分类算法来预测）。

作业内容

基本功能 20'

- **论文检索**：当用户输入关键词进行搜索时，系统应能够展示出与关键词相关的候选论文列表。
- **论文展示**：用户选定某一论文时，系统应能够展示该论文的标题+摘要、引用论文列表、相似论文列表、同类论文列表（测试集的论文标签需要自己先预测）。

其它细节 60'

这里给了如何体现 **隐私保护**、**访问控制**、**系统性能** 的例子，同学们不必局限于此，**合理即可**。

1. **隐私保护** 20'
    - 该系统需要体现隐私保护功能。一个最简单的例子是 用户注册时至少需提供 username, email, password，此时系统应实现 email 和 password 的加密存储。
2. **访问控制** 20'
    - 利用鉴权技术根据用户的角色（如普通用户、VIP 用户）来分配权限/功能。
    - 权限功能可以自己定义，比如用户查询某一论文时，普通用户只展示论文的标题+摘要、引用论文列表；而对于 VIP 用户，还可以提供相似论文列表，额外展示最相似的几篇论文或者同类论文列表。
3. **系统性能** 20'
    - **准确性**：使用更准确的分类算法对测试集论文进行分类，以提高搜索结果的相关性。
    - **高效性**：可以利用向量数据库等技术加速向量查询，近似近邻搜索等算法加快相似性搜索，确保系统响应迅速，提升用户体验。

可以探索的扩展功能

这里是 bonus，不做没啥影响，做了可以额外加分，但上限 100

- **推荐系统**：基于用户的浏览和搜索历史，推荐相关的论文。
- **可视化工具**：提供图表和图形化工具，帮助用户更好地理解论文之间的关系和领域趋势。

本次作业的重点是如何体现 **隐私保护**、**访问控制**、**系统性能**，具体实现路线不限。

作业汇报提交 20'

- 作业组长一人交即可，上交时只需将源码（前后端、分类算法等）和 PPT（不要传 PDF） 打包上交，数据集不要再上传

关于本次作业有任何问题，可以发邮件给助教：刘云辉 yunhui.liu@smail.nju.edu.cn