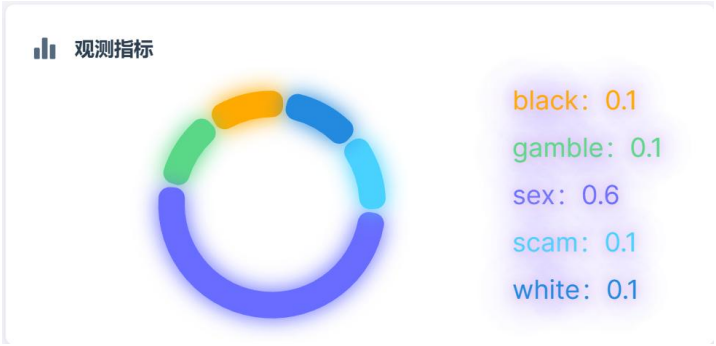
重庆大学大数据与软件学院

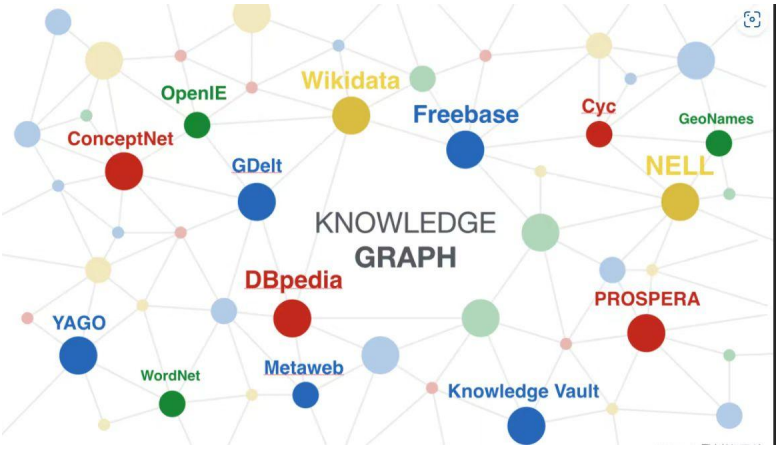
软件综合实践第三周进度及总结报告

<table>
<tr><td>项目名称</td><td colspan="4">基于深度学习和大模型的 android 恶意软件检测系统</td></tr>
<tr><td>项目成员</td><td>学号</td><td>20221982</td><td>姓名</td><td>潘铷葳</td></tr>
<tr><td rowspan="3">组内其他成员</td><td>学号</td><td>姓名</td><td>学号</td><td>姓名</td></tr>
<tr><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td></tr>
<tr><td>项目进展情况
(含验收情况，100 字以上)</td><td colspan="4">1.优化前端界面：
（1）首页地图美化：对首页的地图进行了美化，使其视觉效果更加友好，提升了用户体验。具体美化内容包括调整地图的配色方案，使高风险区域更加明显，并优化了地图加载和缩放的效果。
 

（2）检测报告可视化：在检测报告中添加了观测指标的可视化功能。具体地，我们对 APK 进行多分类检测后，得到的概率用于表示 APK 属于某类别的可能性。概率越大，APK 越有可能是该类别。我们提供了以下几种类别的可视化：black、white、sex、scam、gamble。每种类别的概率使用不同颜色的环形图表示，使用户能够直观地了解 APK 的风险类别和概率分布。
 
</td></tr>
</table>

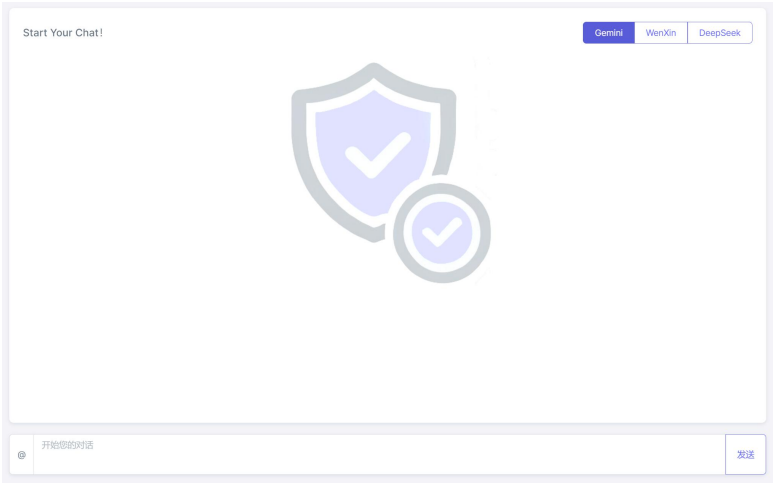
2.知识图谱可视化:

使用知识图谱对黑灰产 APK 之间的关系进行了可视化。这一功能通过展示 APK 之间的关联和网络结构，帮助用户更直观地理解和分析恶意软件之间的联系，提升了系统的分析能力。通过节点和边的形式展示 APK 之间的关系，并支持点击节点查看详细信息。



3. 智能问答功能:

实现了智能问答功能，用户可以通过提供特征信息来对 APK 进行分析，并得到详细的风险评估报告。



总结（总结项目得失，300 字以上）	本次项目**旨在基于深度学习和大模型构建 Android 恶意软件检测系统**。历时三周，我在数据处理、模型训练、前端优化等方面取得了显著进展，同时也遇到了诸多挑战。我们不仅提高了系统的准确性和性能，还增强了用户体验。 **1.项目进展** 第一周主要集中在数据的收集和初步处理。我收集了近六百个 APK 文件，并对其进行分类标签标注，涵盖了 gamble、scam、sex、white、black 五个类别。利用 Androguard 对 APK 文件进行解析，提取了包括应用名称、包名、主要活动、权限、证书信息等多种特征，并将这些特征存储到 CSV 文件中。接着，构建了一个包含 519 条数据的数据集，并设计了调用大模型进行 APK 特征分析的 Prompt 模板。 第二周我着重优化和扩展了数据特征。新增了应用的图标、域名、IP 地址、城市、国家及地理坐标等关键特征，对现有特征进行了清洗和优化。通过训练两个深度学习分类模型，一个针对文本特征，使用 BERT、LSTM 和 Transformer 进行对比，保留了准确率最高的 Transformer 模型；另一个针对图像特征，利用 CLIP 模型进行 Logo 图片的多分类。我还构建了多模态模型，能够同时处理文本和图像特征，实现更全面和精确的分类。 第三周我进行了前端界面的优化和知识图谱的可视化。对首页地图进行了美化，使其视觉效果更友好，并添加了高危地区 IP 的地图可视化功能。在检测报告中添加了观测指标的可视化功能，使用户能够直观了解 APK 的风险类别和概率分布。通过知识图谱，我们展示了黑灰产 APK 之间的关联和网络结构，帮助用户更直观地理解和分析恶意软件之间的联系。 **2.项目得失** **（一）成功之处:** 数据处理和特征提取: 成功收集和处理了大量 APK 数据，提取了多种关键特征，为后续的模型训练打下了坚实基础。 模型训练和优化: 利用先进的深度学习算法训练分类模型，并通过对比选择了最佳模型，显著提高了系统的分类准确性和效率。 前端优化和用户体验: 前端界面的美化和功能的增加，大大提升了用户体验，使系统的使用更加直观和高效。 知识图谱的应用: 通过知识图谱可视化恶意软件之间的关系，增强了系统的分析能力，帮助用户更好地理解和应对 APK 相关的安全风险。 **（二）面临的挑战:** 数据解析问题: 部分 APK 文件由于格式或内容问题导致解析失败，需要手动处理这些异常情况。收集的数据中有一些标签不完整或不准确，需要进一步清洗和校正。 大模型调用问题: 设计的大模型 Prompt 模板在调用时生成的结果有时不够准确或详尽，需要进一步优化以提高响应质

<table>
<tr><td></td><td>量。大模型的计算资源需求较高，在多次调用时容易造成系统负载，需要优化调用策略或增加计算资源 。
前端性能问题: 在加载大量数据时，前端页面的响应速度较慢，特别是在展示高危情况跟踪和风险 APP 总览时，需优化前端性能。
3.未来计划
（一）增强数据采集能力：
扩大数据源：目前的数据主要来源于已有的 APK 文件，未来我们计划扩大数据采集的范围，涵盖更多种类和来源的 APK 文件，以提高数据的多样性和全面性。
自动化数据采集工具：开发和完善自动化数据采集工具，实现对互联网中公开 APK 文件的自动抓取和分类，确保数据的实时性和更新频率。
（二）提高模型的鲁棒性和适应性：
异构数据融合：除了现有的文本和图像特征外，进一步引入更多异构数据，如网络流量数据、用户行为日志等，通过多模态数据融合，提升模型的全面性和适应性。
模型集成与优化：结合多个深度学习模型的优势，进行模型集成（如 Stacking、Blending），提升整体系统的鲁棒性和泛化能力。通过超参数优化和模型调优，进一步提高模型的性能。
（三）前端功能扩展与优化：
实时监控和报警系统：在前端页面中集成实时监控和报警功能，对高风险 APK 进行实时监控，并在发现异常时及时报警，帮助用户快速应对安全威胁。
交互式数据可视化：增强前端的交互式数据可视化功能，使用户可以通过拖拽、点击等方式动态查看和分析数据，提升用户的分析体验和效率。
（四）提升系统性能和扩展性：
分布式计算和云服务：采用分布式计算架构和云服务，提升系统的计算能力和扩展性，确保在高负载情况下系统仍能保持高效运行。
缓存机制和数据压缩：引入缓存机制和数据压缩技术，优化数据传输和存储效率，减少系统的响应时间和存储开销。</td></tr>
<tr><td rowspan="2">项目总体自评
（在相应栏目内打“√”）</td><td>优</td><td>良</td><td>中</td><td>基本合格</td></tr>
<tr><td>√</td><td></td><td></td><td></td></tr>
<tr><td>教师意见</td><td colspan="4">教师签字：
年 月 日</td></tr>
</table>

大数据与软件学院制表