Hackathon 主题四：简历智能解析与造假检测系统

核心需求

构建一个简历处理系统，能够自动解析简历文档，提取关键信息（如个人信息、教育背景、工作经历等），并通过多种方式检测简历中的潜在造假或不一致之处。系统需要能够处理各种格式的简历，识别常见的造假模式，并为HR提供风险提示。

使用场景示例

简历信息提取：

- 系统自动识别上传的简历（PDF、Word、图片等）。
- 自动提取个人信息（姓名、联系方式、邮箱等）。
- 自动提取教育背景（学校、专业、毕业时间等）。
- 自动提取工作经历（公司、职位、时间段、职责等）。
- 自动提取技能和证书信息。

造假检测：

- 检测时间线不一致（如工作经历时间重叠、毕业时间与工作时间矛盾等）。
- 检测学历造假风险（如检查学校名称是否真实、学位是否合理等）。
- 检测工作经历造假风险（如检查公司名称是否真实、职位描述是否合理等）。
- 检测内容夸大或虚假（如检查技能声称是否与经历相符、成就是否可验证等）。
- 检测文本异常（如检查是否存在明显的语法错误、格式不一致等）。

风险评分与建议：

- 系统为每份简历生成一个综合风险评分。
- 系统列出所有检测到的风险点，并提供具体的建议。
- 系统标记高风险简历，建议进行人工审核或背景调查。

功能需求

1. 系统能够接收各种格式的简历（PDF、Word、图片等）。
2. 系统能够准确提取简历中的关键信息。
3. 系统能够检测时间线不一致。
4. 系统能够识别常见的造假模式和风险信号。
5. 系统能够验证某些信息的真实性（如学校、公司等是否存在）。
6. 系统能够为每份简历生成风险评分和详细的风险报告。

7. 系统应该提供一个可视化的界面，展示提取的信息和风险点。
8. 系统应该能够学习和改进，根据用户反馈调整风险检测模型。

数据源

- **简历样本**：参赛者可以自己创建或收集示例简历（确保隐私和合法性）。
- **企业和学校数据库**：可以使用公开的企业名录和大学列表进行验证。
- **常见造假模式库**：参赛者可以基于公开的研究和案例构建。

评分标准

评分维度	权重	关键考察点
信息提取准确性	35%	- 关键信息提取的准确率如何? - 能否处理各种格式和质量的简历? - 提取的信息是否完整?
造假检测有效性	35%	- 能否准确检测常见的造假模式? - 误报率是否在可接受范围内? - 能否识别隐蔽的造假信号?
用户体验与呈现	20%	- 界面是否清晰易用? - 风险报告是否清楚? - 建议是否有帮助?
创新与效率	10%	- 是否有创新的检测算法或验证方式? - 处理效率如何?

提交要求

- 完整的源代码。
- 部署和运行的说明文档。
- 一份演示报告，展示系统在处理不同类型简历时的表现，包括信息提取和风险检测。

附录：技术参考

文档处理

- **PyPDF2 / pdfplumber**：用于PDF文件处理。
- **python-docx**：用于Word文档处理。
- **Pillow / OpenCV**：用于图像处理。
- **pytesseract / EasyOCR**：用于光学字符识别（OCR）。

信息提取与NLP

- **spaCy / NLTK**：用于自然语言处理和信息提取。
- **regex**：用于模式匹配和信息提取。
- **Pydantic**：用于数据验证和结构化。

数据验证

- **requests**：用于调用外部API验证企业、学校等信息。
- 可以使用公开的企业名录API或学校数据库进行验证。

可视化

- **Streamlit / Dash**：用于构建简洁的Web界面。
- **Plotly**：用于数据可视化。