Pseudalis

使用和分享文档，而不暴露其中的个人数据

Pseudalis 检测您文档中的个人数据，并把每个值替换为一个以加密形式包含该值的标签。只有您电脑上的密钥才能还原文档，或任何由它派生的文本。

版本 1.0.11　　平台 Windows 11, x64，无需驱动

许可证 Boost Software License 1.0

示例　　带标签

主题：合同续签

[[PERSON:Rnifypn Eouiitlrwky#EKR]] 您好：

继 [[DATE:28/08/2018#VLC]] 通话之后，现随函附上 [[ORG:Xkekwmufwmcj Eouiitlrwky et Wldwdst SARL#T4N]]（[[LOCATION:Saint-Ntaahgp#QWB]]）的续签文件。
退款将汇入账户 [[IBAN:GB45 NWBK 6152 7186 9288 16#K9P]]。

您可以通过 [[PHONE:06 07 09 00 95#M2X]] 或 [[EMAIL:rnifypn.eouiitlrwky@pspbshklkspj.com#7HD]] 与我联系。

0.946
检测器在五种语言测试集上的 F1（1.0.11 版）

41,000
两个核心上每秒读取的字符数

40
个国家的身份类号码可校验

17 MB
自带一切、无需驱动的安装程序

概览

标签自带其值
每个值都变成一个标签，其中保存着用基于 AES-256 的 FF1 加密后的值，并带有校验。不保留任何对照表。

带标签或不带标签
也可以只写入替换值，让文件读起来像一份普通文档，适合用于提示、测试数据或对外发送的文件。

强制词条与列
管理员最多可以指定 100 个词条和 100 个 CSV 列，在读取其他内容之前先行替换，从而保证导出表格之间的关联仍然成立。

由 Albireo 描述图片
如果同一台电脑上装有 Albireo，每张图片都会被替换为一段不指名任何人的描述。

文档保持版式
Word、Excel、PowerPoint、PDF、文本、CSV、Markdown、HTML、XML、JSON 和 YAML 文件保持原有版式，PDF 保持原有页面。

名称与属性
文件名、压缩包条目和文档属性同样被替换，并在还原时恢复。

原件加密保存
原件进入一个在时间和空间上都有上限的加密保险库，并可凭匿名化文件的任意副本按需取回。

一致的替换值
相同的值在各处得到相同的替换值，受保护的 IBAN 仍能通过校验，日期则按一个保密的偏移量整体平移。

融入您日常使用的工具

一个虚拟盘
复制到 Pseudalis 虚拟盘的文件或文件夹，一旦写入就会被匿名化，连名称在内。无需安装任何驱动。

资源管理器命令
对文件、文件夹以及最多 100 项的选择执行“匿名化”和“还原”，可带标签或不带标签，可原地处理或另存副本。

一个网页应用
拖入文件、压缩包和文件夹，或粘贴一段文本，即可取回匿名化或还原后的结果。

一个本地接口
位于回环地址上、带文档的 HTTPS 接口，供其他程序使用：文件、文本、名称、记录和设置。

四种检测机制

规则识别有结构的值；两个经过训练的小模型则根据上下文识别规则遗漏的值。

带校验的扫描器

电子邮件和 IP 地址、电话号码、IBAN、银行卡，以及约四十个国家的身份类号码，校验位验证通过才予以接受。

结构化 SVM 标注器

一个基于词的序列标注器，根据上下文识别姓名、地点、地址和组织。

词表

五种语言的 44 张词表：称谓、字段标签、法律形式、名字、公共机构以及约两千个组织。

字节级 transformer

一个 116,791 个参数的片段模型，能读取任意拼写，并报告一个值的起止位置。

检测器的实测数据

完整检测器，精确片段，所有类别（Pseudalis 1.0.5）

测试集	精确率	召回率	F1
测试集，五种语言	0.955	0.935	**0.945**
训练中从未见过的值	0.951	0.935	**0.943**
六倍噪声	0.941	0.897	**0.918**
完整文档	0.961	0.943	**0.952**
仅国外地点	0.960	0.939	**0.949**

个人数据类别

类别	状态
姓名、昵称、登录名	● 已覆盖
地址、城市、带城市的邮政编码	● 已覆盖
电话、电子邮件、日期	● 已覆盖
称谓、性别、国籍	● 已覆盖
带校验的标识符：社会保障号、IBAN、银行卡，约 40 个国家	● 已覆盖
IP、MAC、UUID、IMEI、车牌、VIN、GPS	● 已覆盖
组织和具名公共机构	● 已覆盖
身份证、驾驶证、居留许可	● 部分覆盖
客户号、案件号、合同号和员工号	● 部分覆盖
以完整句子写出的健康、信仰或犯罪事实	○ 未覆盖
图片中的文字	○ 未覆盖

Pseudalis 的局限

- 在学习型类别上，检测是概率性的：受保护的文档在分享前应当再读一遍。
- 信任模型以电脑为单位：这台电脑的所有用户共享它的密钥、原件和管理权限。
- 没有文本层的扫描页会被保留为灰色方块：不执行 OCR。
- 如果文档的语言不在识别器支持的五种语言之内，会有更多的值保持可读。

规格

文档	Word、Excel 和 PowerPoint（.docx、.xlsx、.pptx 及其带宏和模板的变体）、PDF
文本	三十多种扩展名，包括 .txt、.csv、.md、.json、.xml、.html、.yaml 和 .sql，以及任何内容为文本的文件
可识别的语言	法语、英语、意大利语、西班牙语和德语
值的加密	基于 AES-256 的 FF1，每个标签以 HMAC-SHA-256 认证
原件保险库	AES-256-CTR 配合 HMAC-SHA-256；主密钥由 DPAPI 保护；密钥以 HKDF-SHA-256 派生
本地接口	https://127.0.0.1:8444，以 Swagger 提供文档
依赖	无：安装程序自带一切，程序只导入 Windows 自身的库

标准与参考

NIST SP 800-38G	值的 FF1 保留格式加密
AES-256, HMAC-SHA-256	标签和保险库的加密与认证
HKDF-SHA-256, PBKDF2-SHA-256	密钥派生以及导出密钥的保护
ECDSA P-256	更新源的签名
GDPR 第 25 条和第 32 条	假名化与加密，作为设计阶段的数据保护措施和安全措施
EDPB 指南 01/2025	假名化，依据 2026 年 11 月时的文本

免费且开源，采用 Boost Software License 1.0。© 2026 Cydral Technology，法国。