THE CLEVER CODERS

LOGAN LIVIGNI
ROBERT WELLS
SANG NGUYEN
ROCCO SWANEY
GAGE FLORES

Stage 1 Presentation

Group Website

WHAT IS PII?

- “Personally identifiable information (PII) is any information connected to a specific individual that can be used to uncover or steal that individual’s identity.” – IBM
- Examples of PII are:
  - Social security numbers
  - Full name
  - Email addresses
  - Phone number
  - Medical Information

BENEFITS OF REDACTING PII

- There are many benefits to redacting PII, they include:
  - Preventing identity theft and fraud
  - Ensuring regulatory compliance
  - Mitigates insider threats
  - Protects against data leaks
  - Enhances document security in cloud storage
  - Boosts consumer and client trust

PII REDACTOR AND ITS PURPOSE

- PII Redactor – our program to automatically identify PII and redact it with a simple user interface for uploading input files and downloading output files.
- Makes PII redaction more efficient than manual redaction.
- Removes potential human error during manual redaction.
- Quickly redacts sensitive information such as names, emails, phone numbers, SSN, etc.
- Provide a simple, accessible interface for non-technical users.

PII REDACTOR LAYOUT & REQUIREMENTS

- The program for the PII redactor will be made using Python.
- The program should be able to read text files.
- The program should also be able to sort files into chosen categories.
- To identify and redact PII, the program will use pattern matching and natural language processing (NLP).
- Masking will be utilized for PII redaction.

INTERFACES

- Operating System required: Windows, macOS, or Linux.
- A regular computer (laptop or desktop) is required.
  - At minimum, computer with basic components.
  - No additional hardware like a graphics card.
- Program and user interaction is done through the console or terminal locally on the user's computer.
  - Running the program will open a console for users to interact with.
- An executable file will be provided for each of the 3 operating systems.
  - Python is not required to be installed.

CONSTRAINTS

- The program must be able to complete the redaction and output of the censored text document in under 1 minute.
- Memory is constrained by the users' memory size and the size of the text document that will input into the program.
- PII redactions must comply with Protected Health Information guidelines set by HIPAA and US Privacy Act of 1974.
- Input files must adhere to their respective file standards.
  - Examples: Tax records must adhere to the standard set by the Internal Revenue Service
  - Medical records must adhere to the legal medical record standards.

PLAN/TASKS

Setup:

- Make config file, basic python scripts, and choose GUI toolkit (if NOT console-run).

Detect and Redact:

- Write a detection function to detect PII using regexes
- Replace PII with tokens, then mask with asterisks.

GUI (if NOT console-run):

- Easy-access options to import files, show preview, batch process, open settings panel.

Command Line Interface:

- Commands to detect, redact, sort, or run full pipeline

Sorting:

- Auto-sort outputs by PII type.

Testing:

- Small test set, check accuracy, fix false detections, error handling.

INTERVIEW LOG SUMMARY

- Person 1
  - Doctor of Occupational Therapy student
  - The interviewee has moderate computer experience (4/10) and expects to redact PII in their future job. They trust automation (8/10 overall, 7/10 for redaction) and would apply it to various records, mostly in PDF form
- Person 2
  - Graduate student
  - The interviewee has moderate computer experience (6/10) and does not expect to redact PII in their future job. They have a moderate trust level in automation (4/10 overall, 4/10 for redaction) but currently do not currently redact PII
- Person 3
  - Nursing student
  - The interviewee has a high level of computer experience (8/10) and expects to redact PII in their future job. They trust automation (7/10 overall, 8/10 for redaction) and would apply it to medical records in PDF form
- Person 4
  - Attorney
  - The interviewee has a high level of computer experience (8/10) and redacts PII in their job. They have a moderate trust level in automation (5/10) but a low trust for redaction (3/10) and they apply it to various records, mostly in PDF form
- Person 5
  - Deputy CISO
  - The interviewee has a high level of computer experience (10/10) and redacts PII in their job. They have a high level of trust in automation (9/10 overall, 9/10 for redaction) and they apply it to various records in word document, PDF file, and image forms