PROJECT REPORT

Project:RE-DACT(SIH1678)

Team Name: Team Rookie

Team Members:

1) Kushal Parihar(Team Leader)
2) P. Varshith(Front End Dev+Documentation)
3) L. Ronith Dhan raj(Back End Dev)
4) Sree Vardhan(Back End Dev)
5) M. Chandana(Front End Dev+Documentation)
6) N. Sriya(Back End Dev)

Problem Statement:

Easy to use and secure redaction tool
“RE-DACT” which allows redaction/masking/anonymization on various input formats based on a gradational scale defined by the user and providing customized output. Over a time, model will learn and have the ability to generate realistic synthetic data in any sought format.

Proposed Solution:

The proposed solution is a natural language processing (machine learning) based redaction tool. The tool will redact or obfuscate from original data leaving the output structurally/logically the same but stripped of key identifiers and other content which may in any way allow the identity, actual data, markers or issues in the input content to be revealed. The correlational logic may be appropriately obfuscated based on the degree of redaction. This will have an easy to use GUI and will be available for use on online and offline systems. The degree of the redaction will be up to the user- the higher the degree set by the user, the more the degree of redaction. This will work with all different commonly used formats for text and data sets. Security of data will be assured by ensuring that the input data is not stored or retrievable in any fashion by third party entities. User will have complete control over the input data. It is also an important aspect that sometimes data may be required to be stored or submitted, however specific sensitive details may not necessarily be required. In such a situation- anonymized data authenticated as having being redacted from original would suffice.

Objectives:

To make a redaction tool ‘RE-DACT’ which allows redaction/masking/anonymization on various input formats based on a gradational scale defined by the user and providing customized output.
To build a easy to use/ User friendly Interface for the RE-DACT tool, which can be in both online and offline mode, with minimal API dependency.

Data Collection and Preprocessing:

Data Acquisition:-Gathering a wide range of publicly available datasets, including text files, images, and PDFs, containing identifiable personal details such as names, email addresses, phone numbers, and geographic data.
To Include images with diverse text fonts, sizes, and languages to enhance the accuracy of Optical Character Recognition (OCR).

Data Preprocessing:-
1) Text Files: Standardize and format the text data to ensure consistency. Eliminate unnecessary characters and ensure the text structure is uniform.

2) Images: Preprocess images by converting them to grayscale or binary formats to improve OCR efficiency. Resize and normalize images to maintain consistency.

3)PDFs: Extract text from PDFs, ensuring proper handling of multi-page documents. Clean the extracted text in a manner similar to standard text files.

Model Development:

Named Entity Recognition (NER):-
Utilize the spaCy library to identify and categorize entities such as names, organizations, and locations within text documents.
Fine-tune the NER model on a custom dataset to improve detection accuracy for specific redaction needs.

Optical Character Recognition (OCR):-
Implement pytesseract to extract text from images. Enhance OCR accuracy by training on customized datasets featuring various fonts and text layouts.

Pattern Matching:-
Employ regular expressions (re library) to identify and redact patterns like email addresses and phone numbers that may not be detected by the NER model.

Redaction Strategy:-
Develop a flexible redaction strategy that adapts based on user preferences, ensuring effective obfuscation of sensitive information while maintaining the necessary data structure.

Model Evaluation and Testing:

Evaluation Metrics:-

1) Text Redaction:- Evaluate the NER model's performance using metrics like Precision, Recall, and F1 Score on publicly available testing datasets.

2) Image Redaction:- Assess OCR accuracy and redaction effectiveness through both visual inspection and automated verification.

3) PDF Redaction:-Measure the tool's ability to accurately identify and redact sensitive information in complex multi-page PDFs.

Testing Methodology:-

Conduct thorough testing across diverse datasets to ensure robust performance with different languages, fonts, image resolutions, and document formats.
To Validate the tool's efficiency at different redaction levels, from basic to advanced level.

Application Development:

User Interface:-Design an intuitive graphical interface using tkinter for the desktop version, allowing users to easily select files, set redaction preferences, and save the output.
Provide options for selecting different file types (text, images, PDFs) and adjusting redaction levels.

Backend Logic:- Integrate the NER, OCR, and pattern-matching components into the backend, ensuring seamless redaction across various file formats.

Web Application:- Develop a web-based version using HTML, CSS, and JavaScript for the front-end, coupled with Flask or Django for the back-end, enabling users to access the tool remotely.

Deployment and Scalability:

Deployment Strategy:- Package the desktop application using PyInstaller or similar tools to create standalone executables for different operating systems (Windows, macOS, Linux).
To Deploy the web application on cloud platforms like AWS or Azure, ensuring it can manage a large user base.

Scalability:-
Optimize the redaction process for large datasets to ensure the tool remains efficient and accurate as data volume increases.
Implement load balancing and caching strategies for the web application to handle high traffic and multiple redaction requests simultaneously.

Security Considerations:-
Ensure the application does not retain any sensitive data after redaction and complies with relevant data privacy regulations.
Incorporate secure coding practices and employ encryption for any temporary file storage during the redaction process.

Challenges:

1)Data Privacy and Security: Ensuring that the redaction process effectively protects sensitive information while still maintaining data utility is a significant challenge. There is always a risk of incomplete redaction or accidental exposure of confidential data, especially when dealing with complex formats or unstructured data.

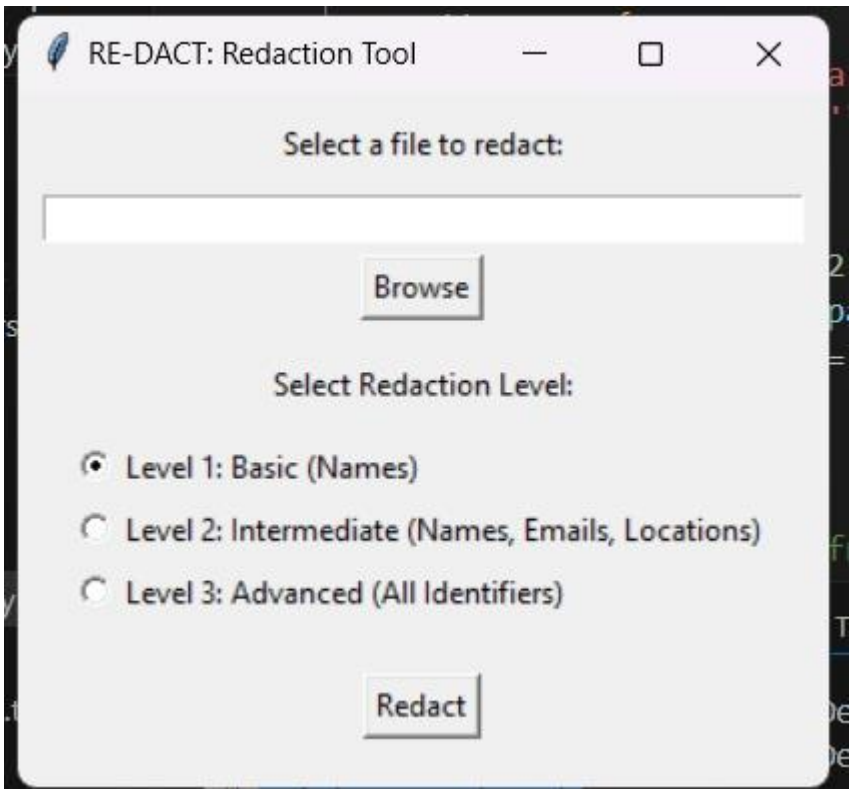
2)Accuracy and Consistency: Maintaining the accuracy of redactions across different document types and formats can be difficult. Inconsistent redaction can lead to information leakage or incorrect data masking, reducing the effectiveness of the tool.

3)Scalability: As the tool begins to handle larger volumes of data, ensuring that it can scale effectively without compromising performance or accuracy is a major technical challenge.

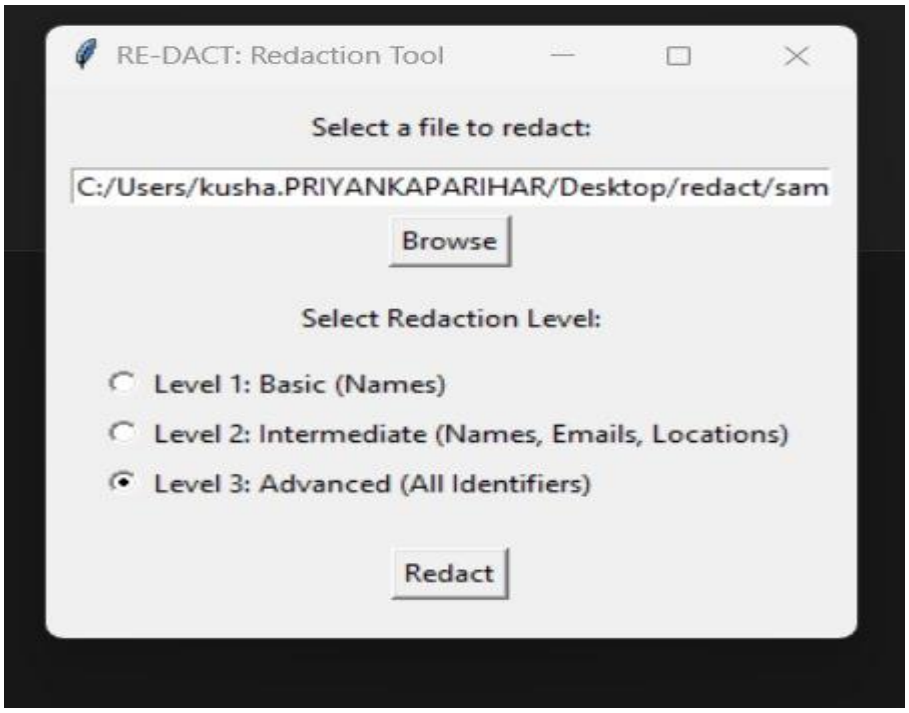
Expected Outcome:

1) Creating a sample text using a random name, with random details such as email, phone number etc.The text file also contains some details about the person.

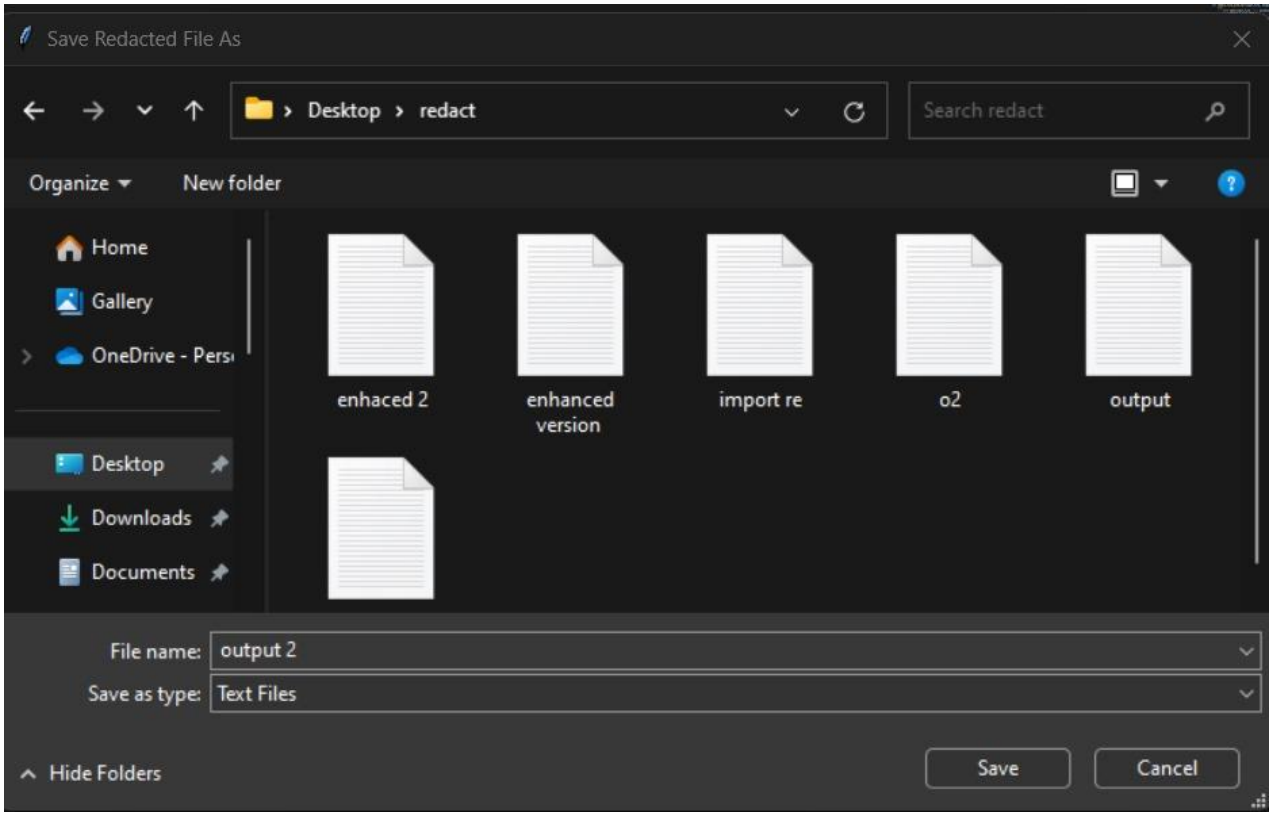
2) Using the redaction tool



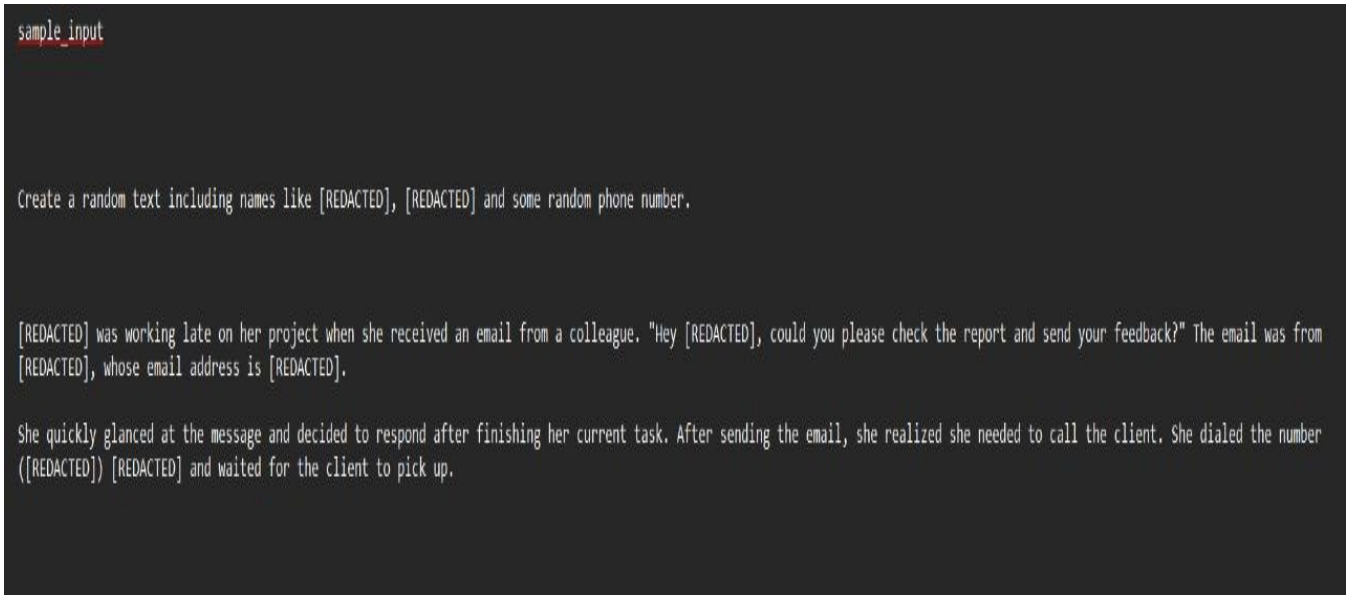
3) Selecting the text file and using the tool to redact the data, the output is saved as output2.txt



4)Selecting the file output2.txt



4) The required output:



Future Scope:

1) **Integration with Enterprise Systems:**The tool could be integrated with existing enterprise content management systems, allowing for automated redaction processes within larger workflows. This would enhance efficiency and reduce manual intervention.

2) **Real Time Redactions:**The development of real-time redaction capabilities could enable the tool to redact live data streams, such as video or audio, expanding its use cases to areas like live broadcasting or real-time communication platforms.

3)**Multi-language support:**Expanding the tool's capabilities to support multiple languages and regions would make it more versatile and applicable in a global context, meeting the needs of international organizations.

Conclusion:

The RE-DACT application/tool represents a significant advancement in the field of data redaction and anonymization, offering users a flexible and secure way to handle sensitive information across various formats.RE-DACT can become an indispensable tool for organizations aiming to protect privacy and comply with data protection laws. As data security becomes increasingly critical, tools like RE-DACT will play a crucial role in safeguarding information while enabling the safe sharing of data.