# ANU Computing Internship

Statement of Work

Parsing custom templates for fake text generation.

Thesis….

Student project COMP3820 / COMP4820 / COMP8830

|  |  |  |  |
| --- | --- | --- | --- |
| Student Name: | Luke Wanless | Student ID: | U7120506 |
| Degree Program: | Bachelor of Advanced Computing (Research and Development) (Honours) | Course Code: | COMP3820 |
| Host Organisation: | Penten | Host Supervisor: | Isaac Jackson |
| Project Title: | Parsing custom templates for fake text generation | | |

# Project Vision and Objectives

*Penten are a cyber security company that provide security solutions to private sector and government organisations. Specifically, the Applied AI department utilise modern AI tools to create Honey files for cyber deception. These honey files are designed to alert system administrators to the presence of hackers within their filesystems. As a part of this customers need to be able to generate honeyfile documents that appear realistic. This has motivated the addition of a feature which allows clients to upload internal organisation templates which can then be populated with tracers and fake generated content.*

*My project concerns parsing these templates in a robust manner to allow for automated generation with minimal supervision and avoid unwanted side effects. Specifically, templates may be populated with stub text or unpopulated. Text will appear in headings, paragraphs, emails and in other locations. Similarly, these templates may be entirely empty. In order to generate text in an automated and robust manner it is necessary to first classify these document elements such that the appropriate replacement text can be generated. My project aims to use large language model API to parse the XML of .docx files in an unsupervised manner and return a complete honeyfile.*

*This is an important extension to the hosts current functionality as it improves the realism of generated honeyfiles and removes the requirement of manually uploading and onboarding the template specifications of each new client; a process that is not scalable or compatible with widely distributed software.*

*The end goal for the project will be uploading an array of custom templates with random fields, parsing this template accurately and returning a contextually consistent honeyfile using large language models.*

# Project Stakeholders and Benefit

*My project stakeholders, their respective requirements and the benefits they derive from my project are as follows:*

1. *Penten management – Senior product managers will provide me with a high-level description how my project should operate from a user perspective. These requirements are not aimed at an implementational level and specify how a user will interact with the custom templates and what functionality should be supported at a minimum. The requirements I have received from them are to support the detection of the most common elements found in a templated document. These are as follows:*

* *Titles/headings*
* *Email/phone numbers*
* *Tables*
* *Images (non-strict requirement)*
* *Text paragraphs*

*Following on from the detection of the above document elements, my project must then support realistic content replacement using OpenAI’s large language models.*

1. *Penten developers – The developers at Penten will be responsible for integrating the work I complete into their software product to deliver to clients. As a stakeholder they require that the work I generate supports packages that may be integrated easily into the existing codebase. This not only reduced the overhead associated with including my work however affords them a greater capacity in assisting me troubleshoot potential problems with frameworks they are familiar with. The steps I have taken to ensure compatibility are as follows:*

* *Running the same python version*
* *Using OpenAI API for text generation*
* *Using lxml python standard library for XML parsing*

*The benefit they will derive from this project is the functionality to support custom template parsing and in a manner that is compatible with the existing codebase.*

1. *Clients – Finally the client will be the end user of the custom template functionality. Clients are large organisations with a need for information security. As previously mentioned, a custom template parser will allow users of the product to upload their organisations templates with minimal supervision and human input. This benefits them as a stakeholder by facilitating the generation of more targeted and thus realistic honeyfiles. Furthermore, the intelligent parsing of documents reduces the overhead of bulk template enrolment. The custom template feature will increase the specificity of the content that can be generated as it can be conditioned on domain specific metadata and is generated around company specific filetypes.*

# Stakeholder interaction

*Due to the nature of Penten as a high security company, stakeholder interaction is interfaced through management. As such my only interactions are with internal Penten staff. This reduces my control over how well I can elicit requirements from the user. However, this workflow yields the added benefit of allowing me more frequent opportunities to clarify my project requirements as I will be working with management for the duration of the project.*

# Project Plan

|  |  |  |
| --- | --- | --- |
| ***Milestone*** | ***Date / Week*** | ***Requirements*** |
| *Begin internship* | *Monday 20 February 2022 / Week 1* | *Start working at placement organisation.* |
| *Create environment and tooling* | *Tuesday 28 March / Week 2* | *Create a Github repo.*  *Configure development environment on personal laptop.* |
| *Elicit project requirements* | *Tuesday 7 March / Week 3* | *Complete list of requirements from product managers and developers.* |
| *SoW* | *Friday 17 March / Week 4* | *As defined in course outline.* |
| *Text detection and classification* | *Tuesday 21 March / Week 5* | *Complete functionality for collecting all document text from a custom template. Then using text detection, classify each text field in XML as a document element (title, name, email, table, etc…).* |
| *Full text replacement* | *Tuesday 28 March / Week 6* | *Allow classified text to be replaced with generated text that is suitable to the section of the document.* |
| *Image generation* | *Tuesday 18 April / Week 7* | *Allow images to be generated and inserted into the document that are conditioned on the newly generated text content.* |
| *Front end development* | *Tuesday 2 May / Week 9* | *Allow users to interact with the custom template parser through a front-end web app that supports custom template uploading, parsing and replacement.* |
| *Poster and Pitch* | *Friday 12 May / Week 10* | *As described in course outline.* |
| *Portfolio* | *Friday 2 June / Exams* | *As described in course outline.* |
| *Final reflection* | *Friday 9 June / Exams* | *As described in course outline.* |

# Project Resources and Tooling

*Development tools:*

* *Neovim: Text editor and IDE*
* *Lazygit: Project version control interface*
* *Pre-commit: Linting, automated testing and code quality management*
* *Github: Version control tool*
* *Python: Backend language*
* *Django: API framework to connect back and front-end*
* *HTML: Front end markup language and GUI*
* *Jupyter notebooks: For testing proof of concept code*
* *OpenAI: Text generation API*
* *Tmux: A terminal multiplexer that allows for more efficient terminal development and session sharing with developers*

*Project management:*

* *Microsoft teams: Communicating with various stakeholders*
* *Jira: Agile Kanban board for task management and scheduling work into manageable tickets*

*Resources and tools:*

* *OpenAI API key: Fund API text generation (Penten provisioned)*
* *Work Laptop: For development (Penten provisioned)*
* *Personal laptop: To allow working from home*
* *Development servers: SSH access to remote dev servers (Penten provisioned)*

# Project Constraints and Risks

*Project constraints:*

* *Access to remote server localhosts: Due to security reasons access to localhost IP’s on remote servers is restricted for personal projects. Front end development testing will be done on my personal machine.*
* *No Unix shell sudo access on Penten provisioned machine: Due to security risks my Penten laptop has no sudo privileges. This may slow down development and packages that need to be installed. To mitigate this, I have tried to limit the tools I require to those that don’t require sudo access.*

*Project risks:*

* *Work balance: Due to also being a developer at Penten in addition to the internship project, managing time effectively and not neglecting the project will be important to ensure its completion.*
* *XML non-determinacy: As XML is largely non-deterministic (Sentences can be randomly split amongst multiple XML nodes) finding a robust solution may prove difficult and stall progress. To mitigate this, multiple avenues of text classification will be explored in appropriate breadth to increase the likelihood of finding a workable solution.*
* *Front-end experience: Having no prior front-end experience may lead to slower than usual development. This must be managed appropriately and not left until the last minute due to scope creep in back-end functionality. To mitigate front end development will be done in parallel to the backend to allow sufficient time to learn Django and HTML.*

Supervisor Approval

*I [Supervisor name] confirm that I have read and agreed to the scope of works set out in this Statement of Work document.*

|  |
| --- |
| *Supervisor signature Date* |