User Stories

- **Mr Smith** is a junior government official whose job entails reviewing documents to identify sensitive information that should not be available to the public. He believes he would be able to perform his job more efficiently if he was provided with a *web-based tool* to speed up the reviewing process. He is not very tech-savvy so he would prefer if the tool *did not require excessive training* to understand and use. Mr Smith believes *having a login* to his own personal account - with *the ability to upload, delete and review documents* – would make his job far easier than the current paper based approach.
- **Mrs Johnson** is a senior government official in charge of coordinating a team of document sensitivity reviewers. She is under pressure from her senior supervisors to reduce staff budget spending on his department, however her team of staff is struggling to keep up with the current workload of document reviews. She recognises that current staff must complete document reviews quicker to keep up with the workload. She believes her staff would complete document reviews more efficiently if they had access to a tool that would *provide additional information accompanying each document* to assist the reviewing process. She feels that having *entities highlighted and tagged with their abstracts* would help provide much needed context to certain documents. Mrs Johnson believes that *additional plots breaking down frequency and probability of entity occurrences* would assist in judging sensitivity, as well as *knowledge graphs* linking information across the entire corpus.
- **Mr O'Shea** is a member of parliament who is currently under investigation for numerous disciplinary issues. He accepts the consequences of his actions however he is concerned that the government issued documents related to his offences will contain sensitive, personal information that is not currently available to the public. He is worried that the reviewers assigned documents relating to him may not correctly identify them as sensitive and feels the reviewers should be given a tool that *detect phone numbers and addresses*. He believes they should be assisted by an application that can *automatically recognise sensitive documents*, rather than having each document manually reviewed.

MoSCoW Requirements

Must Have

- Web-app based tool
- Upload documents
- View list of documents
- Delete documents
- Manually review sensitivity of document
- Tag named entities in each document
- Display abstracts relating to entities

Should Have

- Plots & graphs of analysis of documents
- Plots & graphs of analysis of corpus
- Login to account

Could Have

- Knowledge graphs from analysed documents
- Automatic suggestion of sensitivity (ML model)
- Recognise phone numbers & addresses

Won't Have

-

Non-Functional Requirements

- User-Friendly – The application should be easy to use for users without technical expertise or prior training
- Responsive – The application should provide regular updates throughout different steps of processing to keep the user informed
- Maintainable – The codebase should be extensively documented and engineered in a way such that future contributors will not need supervision to extend or maintain functionality
- Efficient – The application should perform at a speed such that is not hindering user experience
- Accurate – The application should identify entities, their attributes and the relevant sensitive information at a high accuracy

My view for the project is a web application designed to assist sensitivity reviewers in detecting and flagging sensitive documents. The homepage of the web app will feature background context for the motivation of the project and the desired use, as well as in depth instructions as to how to efficiently use the provided tools. Users will be able to upload a corpus of documents, with each document being individually analysed using named entity recognition tools. Each entity detected in a document will have its 'entity abstract' – a brief description of itself – scraped from a corresponding website and displayed available for the user to hover over and view. The process of entity recognition and web scraping may be computationally expensive so I will develop a database for the storage of entity abstracts, entity instances and document texts. This will allow the web app to perform a one off analysis of each document when it is uploaded and refer back to stored data for future references.

Once documents are uploaded and processed, users will be able to navigate to document or corpus analytic pages. The document analytics page will provide in depth analysis such as where entities occur in the document, what types of entities occur the most frequently and filtering by types of entity. The corpus analytics page will provide in depth analysis on the set of documents uploaded as a whole – Which entities occur most frequently in a document for a given entity and analysis of conditional probability of which entities are most likely to appear in a document, given it is deemed sensitive and vice versa.

https://www.figma.com/file/HhZTX2XQglS0nDlXdQi19G/L4-Project-Wireframes?node-id=0%3A1