Many government documents contain sensitive information that must be identified and protected before the documents can be released to the public. While manually reviewing such documents for sensitive information it can be important to determine contextual information about specific entities that are mentioned in the documents and whether the information that is discussed about these entities is already in the public domain. In this project, you will develop a system that can automatically identify external information about specific entities from publicly available knowledge graphs (e.g. Wikidata or DBpedia). The system should be able to assist human sensitivity reviewers by identifying entities that are referenced by different names in the collection (based on the entity's attributes) and whether personal information about named entities is in the public domain.

You will work with named entity recognition tools (e.g. spacy https://spacy.io/) along with entity linking tool such as ReFinED (https://github.com/amazon-research/ReFinED) or DBpedia Spotlight (https://www.dbpedia.org/resources/spotlight/). A graph databases such as Neo4j (https://neo4j.com/) will likely also be used to dynamically build a definitive view of the entities within the document collection.

- **Summary of what was agreed last week**
  - Topic modelling
  - Improve web app aesthetic

- **Progress made in the past week**
  - Added cached tokenised & pre-processed versions of each existing document into DB
  - Documents are tokenised and pre-processed as they are uploaded to reduce repeat processing
  - Created topic model example on homepage
  - Used Python word cloud module to display topics

- **Main questions for discussion**
  - Unsure as to whether I can link sensitivity values into topic modelling word cloud
    - Word cloud module does not support editing individual word colours
  - Change topic modelling tokens to just entities? Could improve the relevancy of word clouds if it just features entities

- **Feedback from meeting**
  -