Increasing the Accessibility of NLP Techniques for Defence and Security using a Web-Based Tool

Katie Paxton-Fear K.Paxton-Fear@cranfield.ac.uk

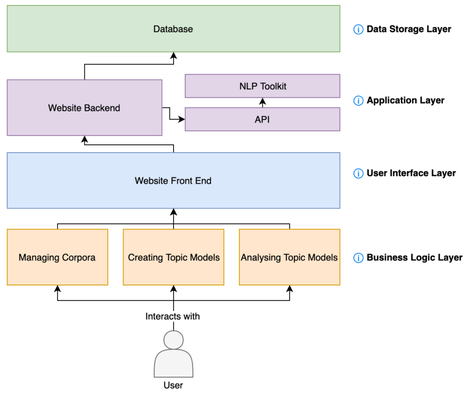
What is Natural Language Processing (NLP)?

NLP techniques are the mechanisms by which a machine can process and analyse text written by humans. These methods are used for a range of tasks including categorising documents, translation and summarising text. To accomplish this a large volume of text (corpus) is collected, processed and finally models can be trained on the corpus. With current methods, the ability to manage corpora is rarely considered, instead relying on researchers and practitioners to do this manually in their file system. To train models, researchers write code directly for one task or experiment, limiting their reusability and ability to generalise.

The Problem: Low Accessibility

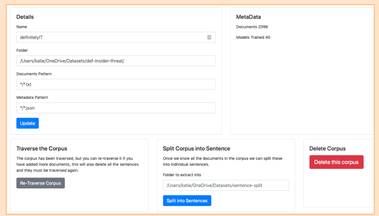
Although machine learning (ML) and natural language processing is becoming more common in Defence and Security, there are issues limiting its use. One such issue is the multidisciplinary nature of Defence and Security, with individuals from many backgrounds contributing to a single project. ML and NLP techniques require distinct specialisms to create and interpret a model. This is even more important when delivering research, where outputs may be operationalised and the accessibility can be a limiting factor in their deployment and use, limiting the impact of the work.

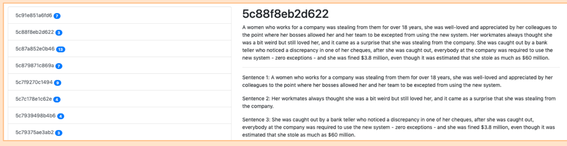
Solving the Problem



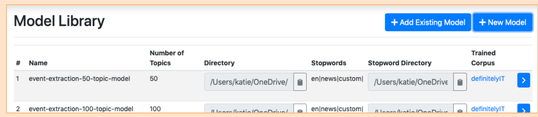
Managing Corpora

The corpora library allows a user to complete basic operations, including: creating a new corpus, changing the settings of an existing corpus and exploring the documents within a corpus. In addition it allows for complex topic modelling specific operations such as splitting a corpus into sentences and traversing a corpus.

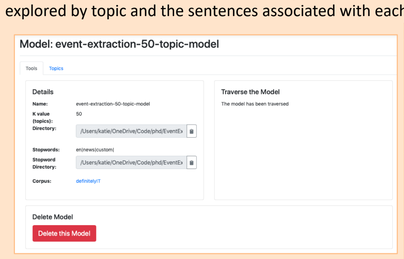


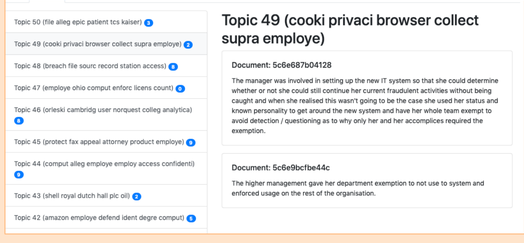


Creating Topic Models



Topic Models are models that attempt to split a piece of text into topics, this allows researchers to explore the content of text computationally. The topic model library allows for basic operations such as creating new topic models. However, users can also: import topic models that have already been created, traverse a topic model and delete a topic model. In addition models can be explored by topic and the sentences associated with each topic can be viewed.



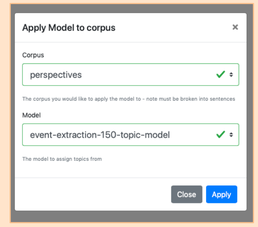


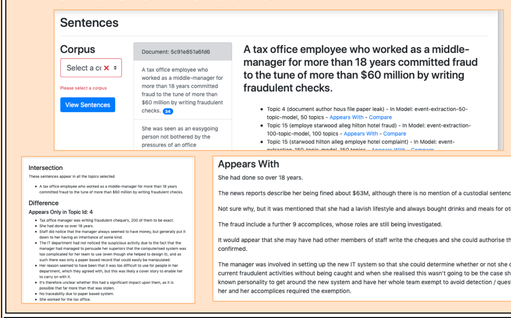
Analysing Topic Models

Finally topic models can be applied to a corpus and analysed. *The Application Tool* allows a user to apply a model and view the results either:

- At a topic level, viewing sentences assigned a certain topic from different documents
- At a document level viewing how different sentences have been assigned different topics.

The Analysis Tool compares topic models by evaluating the symmetrical difference and the intersection, visualising how different models evaluate the same data.





Jargon

- *Corpus/Corpora* – A body of documents used to create a model
- *API* – A piece of software that is designed to communicate with software rather than being used by a human
- *Training* – The process used to create machine learning models
- *Insider Threats* – Security threats that arise from an organisations own employees rather than externally
- *Topic Models* – An NLP model that attempts to automatically find topics in a piece of text using probabilities and key words

Summary

- Using natural language processing (NLP) tools can be difficult for non-specialists.
- Defence and Security involves many different people with different backgrounds
- Machine learning is becoming more common within the defence and security domain
- Therefore it is necessary to create a tool which offers the same functionality as ad-hoc code but is clear and approachable without requiring specialist skills
- **Solution: Web based tool with R application programming interface (API)**

Technology Stack

Front End – Bootstrap Framework
Back End – PHP7 and the Laravel Framework
NLP API – R using Plumber
Database - PostgreSQL

Insider Threat

This tool-support has been created as part of a project considering the use of NLP to better understand reports of insider threat attacks. These are security incidents where the attacker is a member of staff or another trusted individual. Insider threat attacks are particularly difficult to defend against due to the level of access these individuals gain during the regular course of their employment. The wider use of these techniques would generate greater impact both tactically in defending against these attacks and strategically in developing policy and procedures. There are tools available, however they are often complex and perform a single-task, limiting their use. To generate maximum impact from our research we have developed this web-based software to make the tools more accessible, especially to non-specialist researchers, customers and potential users.

Katie Paxton-Fear K.Paxton-Fear@cranfield.ac.uk †, Dr Duncan Hodges d.hodges@cranfield.ac.uk †,
Dr Oliver Buckley o.buckley@uea.ac.uk‡
† Cranfield University ‡ University of East Anglia
Centre for Electronic Warfare and Cyber, Cranfield Defence and Security, Defence Academy of the United Kingdom
www.cranfield.ac.uk