

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)?	Managing Corpora
NLP techniques are the mechanisms by which a machine can process analyse text written by humans. These methods are used for a range of ta including categorising documents, translation and summarising text. accomplish this a large volume of text (corpus) is collected, processed finally models can be trained on the corpus. With current methods, the ab to manage corpora is rarely considered, instead relying on researchers practitioners to do this manually in their file system. To train mode researchers write code directly for one task or experiment, limiting to reusability and ability to generalise.	asks To and idefinitelyIT '/'txt //Jsers/katie/OneDrive/Datasets/Perspinely/ and idefinitelyIT '/'txt '/'json //Jsers/katie/OneDrive/Datasets/Perspinely/ and ility idefinitelyIT '/'txt '/'json //Jsers/katie/OneDrive/Datasets/def-in and ility initian initian '/'txt '/'json //Jsers/katie/OneDrive/Datasets/def-in and inity inity '/'txt '/'json //Jsers/katie/OneDrive/Datasets/def-in and inity inity inity '/'txt '/'json //Jsers/katie/OneDrive/Datasets/def-in and inity inity inity inity inity inity and inity inity
The Problem: Low Accessibility	Details MetaData Name Documents 2096 definitely!T III
Although machine learning (ML) and natural language processing is become more common in Defence and Security, there are issues limiting its use. It such issue is the multidisciplinary nature of Defence and Security, we individuals from many backgrounds contributing to a single project. ML NLP techniques require distinct specialisms to create and interpret a mo This is even more important when delivering research, where outputs may operationalised and the accessibility can be a limiting factor in the deployment and use, limiting the impact of the work.	Documents vattern '/t.txt '/t.txt Metadata Pattern '/t.txt '/*json '/t.tx '/*json '/t.tx '/*json '/t.tx '/*json '/*json '/*json '/t.tx '/*json
Solving the Problem	5c91e851a6fd6 ? 5c88f8 5c88f8eb2d622 () A women who we the point where he sa a bit weird be teller who noticer new system - zero new system
Database (i) Data Storage Layer	5c879871c869a ? Sentence 1: A wo colleagues to the 5c7f9270c1494 ? 5c7f9270c1494 ? Sentence 2: Her the company. 5c7939498b4b6 ? Sentence 3: She everybody at the estimated that she estimated that she colleagues to the shimated that she estimated that she colleagues to the colleagues to the estimated that she colleagues to the estin the coleagues to the estimated that she colleagues t
Website Backend	 Jargon Corpus/Corpora – A body of documents used to a API – A piece of software that is designed to compare the software the software that is designed to compare the software the s
Website Front End	 human Training – The process used to create machine le Insider Threats – Security threats that arise from Topic Models – An NLP model that attempts to au probabilities and key words
Managing Corpora Creating Topic Models Analysing Topic Models Interacts with User User	 Using natural language pr Defence and Security invo Machine learning is becomin Therefore it is necessary to created but is clear and at the solution: Web based to be the solution of the

Katie Paxton-Fear K.Paxton-Fear@cranfield.ac.uk⁺, Dr Duncan Hodges d.hodges@cranfield.ac.uk⁺, Dr Oliver Buckley <u>o.buckley@uea.ac.uk</u>[‡]

[†] 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

	Creating Topic Models				Analysing Top	oic Models		
sentence Split Corpus rspectives/ //Jsers/katie/OneDrive/Datasets/perspectives-split/ Inspect 7 I-Insider-threat/ //Jsers/katie/OneDrive/Datasets/sentence-split/ Inspect 7 -articles/ Inspect 7 ro complete basic operations, including: ettings of an existing corpus and exploring n addition it allows for complex topic as splitting a corpus into sentences and for (Deers/katie/OneDrive/Datasets/all-articles/ Name	Model Library + Add Existing Model + New Model # Name Topics Directory Stopwords Stopword Directory Trained 1 event-extraction-50-topic-model 50 /Users/katie/OneDrive/ enlnews[custom] /Users/katie/OneDrive definitely/1 2 2 event-extraction-100-topic-model 100 R Leven/kestic/OneDrive/ enlnews[custom] R Leven/kestic/OneDrive/ definitely/1 2 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.		his pic els. ed, be	 At a documents At a document level viewing how different sentences have been assigned different topics. 		Apply Model to corpus Corpus perspectives Image: the corpus you would like to apply the model to - note must be broken into sentences Model event-extraction-150-topic-model Image: the model to assign topics from Close Apply		
here Delete Corpus Delete Corpus Delete this corpus */*.txt A pattern that describes where the metadata is stored, see readiext for more info Metadata Pattern */*.json A pattern that describes where the metadata is stored, see readiext for more info Close Add 88b2d622	Model: event-extraction-50-topic-model Tools Tools Tools Tools Tools Details Name: event-extraction-50-topic-model Kvalues Kvalues Topics Directory: /Users/katie/OneDrive/Code/phd/EventEx Topics: Corpus: definitelyT Delete Model Delete Model		Create a New Model × Name event-extraction-50-topic-model r The name of this model note: without the .rds extension c Corpus © © The comp of this model on, note this does not generate post posterior c Decomp to train this model on, note this does not generate post posterior c Decomp to train this model on, note this does not generate post posterior c Decomp to train this model on, note this does not generate post posterior c Decomp to train this model on, note this does not generate post posterior c Decomp to train this model on, note this does not generate post posterior c Decomp to train this model on, note this does not generate post posterior c Decomp to train this model on, note this does not generate post posterior c Number of Topics (K) c Hours/Default/Models/ c Full atta to this fielder c Stopword list directory c English News Specific Custom Copiii (Coupling) c c Definit News Specific Custom c		and the intersection, visualising how different models evaluate the same data. Sentences Corpus Select a ct X : Fues sterict a copy Vew Sentences New Sentences She was seen as an easypoing person not bothered by the pressures of an office New Sentences She was seen as an easypoing Person not bothered by the pressures of an office New Sentences She was seen as an easypoing Person not bothered by the pressures of an office New Sentences She was seen as an easypoing Person not bothered by the pressures of an office New Sentences She was seen as an easypoing Person not bothered by the pressures of an office New Sentence Status of Sentence Appears With - Compare She was seen as an easypoing Person not bothered by the pressures of an office New Sentence Status of Sentence Appears With - Compare Status of Sentence Appears With - Compare Status of Sentence Appears With - Compare			
Devoks for a company was stealing from them for over 18 years, she was well-loved and appreciated by her colleagues to re her bosses allowed her and her team to be excepted from using the new system. Her workmates always thought she d but still loved her, and it came as a surprise that she was stealing from the company. She was caught out, by a bank iced a discrepancy in one of her cheque, after she was caught out, everytody at the company was required to use the point where her bosses allowed her and her team to be excepted from using the new system. Works for a company was stealing from them for over 18 years, she was well-loved and appreciated by her the point where her bosses allowed her and her team to be excepted from using the new system. Works for a company was stealing from them for over 18 years, she was well-loved and appreciated by her the point where her bosses allowed her and her team to be excepted from using the new system. Her workmates always thought she was a bit weird but still loved her, and it came as a surprise that she was stealing from the company was required to use the new system - zero exceptions - and she was fined \$3.8 million, even though it was to a second of the company was required to use the new system - zero exceptions - and she was fined \$3.8 million, even though it was to as \$60 million. Or create a model municate with software rather than being used by a learning models n an organisations own employees rather than externally automatically find topics in a piece of text using	Delete this Model Topic 50 (file alleg epic patient tcs kaiser) Topic 49 (cooki privaci browser collect supra employe) Topic 49 (cooki privaci browser collect supra employe) Topic 48 (breach file sourc record station access) Topic 47 (employe ohio comput enforc licens count) Topic 46 (orleski cambridg user norquest colleg analytica) Topic 45 (protect fax appeal attorney product employe) Topic 44 (comput alleg employe employ access confident Topic 43 (shell royal dutch hall plc oil) Topic 42 (amazon employe defend ident degre comput)	Document: 5c6e687b04128 The manager was involved in setting up the new IT whether or not she could still continue her current caught and when she realised this wasn't going to known personality to get around the new system a avoid detection / questioning as to why only her ar exemption. Document: 5c6e9bcfbe44c The higher management gave her department exemption.	T system so that she could determine t fraudulent activities without being b be the case she used her status and and have her whole team exempt to and her accomplices required the		<section-header><section-header><section-header><text><text><section-header><section-header><section-header><list-item><list-item><section-header></section-header></list-item></list-item></section-header></section-header></section-header></text></text></section-header></section-header></section-header>	She had done so over 18 years. The news reports describe her being fined at Not sure why, but it was mentioned that she l The fraud include a further 9 accomplices, w It would appear that she may have had other confirmed. The manager was involved in setting up the r current fraudulent activities without being ca known personality to get around the new sys her and her accomplices required the exemp or created as part of reports of insider	members of staff write the cheques and she could authoris new IT system so that she could determine whether or not sl uight and when she realised this wasn't going to be the case tem and have her whole team exempt to avoid detection / qu tion.	r others. e them however this is not he could still continue her a she used her status and uestioning as to why only g the use of are security
Summary		Technology Stack Front End – Bootstrap Frame	work		Insider threat attacks are of access these individual	s gain during the re	•	nployment.

rocessing (NLP) tools can be difficult for non-specialists. olves many different people with different backgrounds ing more common within the defence and security domain eate a tool which offers the same functionality as ad-hoc code approachable without requiring specialist skills tool with R application programming interface (API)

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



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.

