





# Identifying Changes in the Cybersecurity Threat Landscape using the LDA-Web Topic Modelling Data Search Engine

Thursday 13<sup>th</sup> July 2017 Multidisciplinary approaches to Cloud Crime HCII 2017, Vancouver Canada

Noura Al Moubayed, David Wall, and A. Stephen McGough

## Outline

## • The Problem

- Text Processing for Topic Modelling
- Using Topic Modelling for searching

## The Problem

- "90% of all the data in the world has been generated over the last two years"... IBM
- "85% of worldwide data is held in un-structured formats"…
  Berry and Kogan



- How can we understand it? ....or better still make use of it?
- How can we determine the most pertinent information? ...and then act on it?
- How can we find the needle if we are not sure what it looks like or what hay looks like?
- "Without labelling, you cannot train a machine with a new task"... IBM

## Outline

- The Problem
- Text Processing for Topic Modelling
- Using Topic Modelling for searching

# Topic Modelling: Latent Dirichlet Allocation (LDA)

## Topics

smart country attacks

military government mational hacker industry technology bonds

defendant company state with a sanctions companies security sillion defendant company state with a sanctions clinton quantum system information people information people state world security military state korea satellite customer states

malware clinton pakene intelligence monday accounts intelligence fancy accounts intelligence korea in clinton pakene intelligence korea intelligence kore



Select topics



monday accounts monday accounts fancy accounts fancy accounts trump soverment sove

# Select words from topics

US Govt Data Shows Russia Used Outdated Ukrainian PHP Malware

The United States government earlier this year officially accused Russia of interfering with the US elections. Earlier this year on October 7th, the Department of Homeland Security and the Office of the Director of National Intelligence



## Outline

- The Problem
- Text Processing for Topic Modelling
- Using Topic Modelling for searching

## Demo – Topic Modelling





Search Engine based on Topic Modelling

http:\\www.pravdareport.com\world\americas\31 -10-2016\136031-china_spies-0\	43%	•••	$\overline{}$
http:\\www.theregister.co.uk\2016\09\22\homela nd_security_launches_iot_campaign\	42%	•••	
<u>https:\\www.rt.com\news\354226-china-sea-</u> <u>navy-drills\</u>	37%	••	
<u>https:\\www.Hackread.com\anonymous-opicarus-</u> DDoS-bilderberg-website\	30%	••	
http:\\www.mondaq.com\unitedstates\x\540492\ Insurance\Insurance+Regulators+Fine+Tuning+Cy bersecurity+Guidance	29%	••	
http:\\www.homelandsecuritynewswire.com\dr201 60609-Hackers-could-easily-cause-drones-to- ignore-human-controllers-or-crash	16%	•••	

	Durham University Search Engine based on Topic Modelling	
-		
Þ	Upload file Choose File no file selected	
	User Query	
	No. Displayed Results	
	100 Search	
	Return to Home Page	



## Potential Applications:

- Security Applications:
  - Terrorist activity tracking
    - Acting out of character, predicting activity
  - Police crime database
    - Criminal profiling, acting out of character
  - Unwanted information release
    - Topic changes, specific damaging subjects

## • Other

- Student applications
  - Identifying bogus attempts for visa
- Social Media tracking
  - Social grooming, political persuasion, product complaints
- Fake News identification and tracking
- Sentiment tracking
  - .....Your use-case here

We Are recruiting:

- 1 PostDoc (Machine Learning / NLP)
- 1 PostDoc (Parallel Programming)
- Always looking for good PhD Candidates

<u>stephen.mcgough@newcastle.ac.uk</u> <u>noura.al-moubayed@dur.ac.uk</u> D.S.Wall@leeds.ac.uk