Educated guesses and equality judgements?

PAN 12

Lee Gillam, Neil Newbold, Neil Cooke

with contributions from Peter Wrobel, Henry Cooke, Fahad Al-Obaidli

University of Surrey

www.surrey.ac.uk

Talk outline

UNIVERSITY OF SURREY

It was suggested that we spend time talking about the approaches taken to the tasks.

We had other ideas.

Our Challenge

UNIVERSITY OF SURREY

How do you do efficient plagiarism detection that can scale to the entire (deep) web AND be useful across (private) corporate resources and across (private) corporates?

We want a good answer quickly … at the speed of search?

The Corporate Security Problem

UNIVERSITY OF SURREY

- £9.2bn of IP theft per year? “this cyber criminal activity is greatly assisted by an ‘insider’”
  - X is a secure system, Y is not a secure system; wetware bridge works around a data transfer issue. Can’t build a bridge between, so need a proxy. If such a proxy can exist, we must be able to use it in the open.
    - How to find out if X data has been exposed without exposing data about X? [#superinjunction]
    - Or …. **How to search without revealing a query**, or using expensive techniques such as homomorphic encryption?

The Corporate Security Problem

UNIVERSITY OF SURREY

- Smells like plagiarism
  - but common plagiarism approaches can't get you there – have to expose the queries, or somehow "lock them up" (hash/encrypt).
  - very difficult to reverse engineer our patterns – highly lossy compression – yet still good match (vs e.g. most/least significant bit-drop type approaches).

Our method is…

- Covered by a kind of **superinjunction** for the time being.
- Licensed to a department of UK Government
- In commercialisation discussions under NDA with parties including a large automotive.

Common approaches

UNIVERSITY OF SURREY

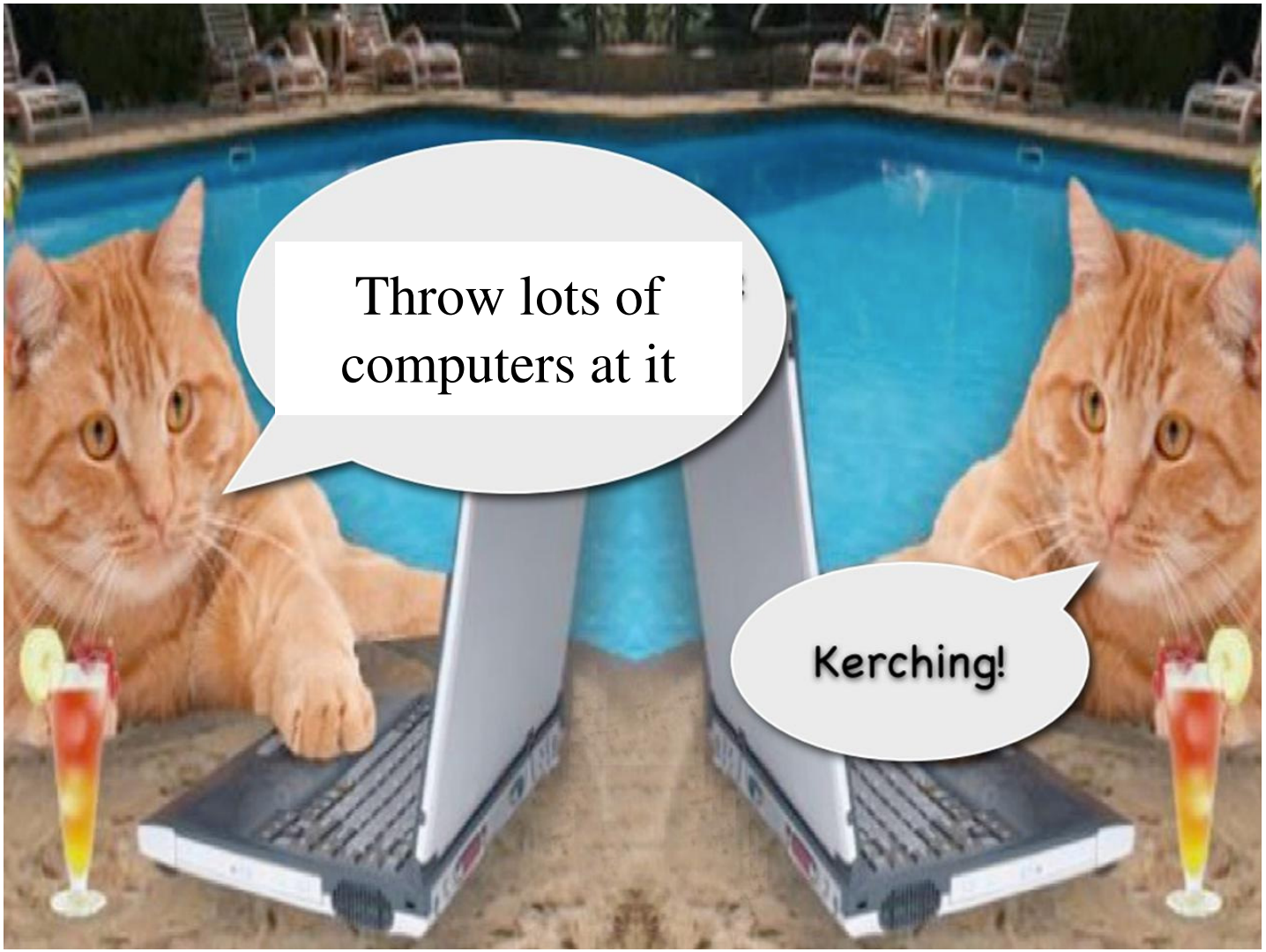
- Remove stopwords
- Use stemming
- Use POS tagging
- Bigrams, trigrams, …
- Use (uniquely) resolvable encodings
- ….

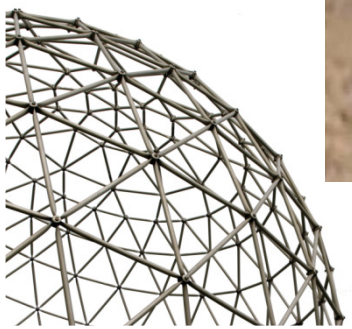
Common approaches

UNIVERSITY OF SURREY

- Remove stopwords – **loss of structure**
- Use stemming – **well, you can but what gain**
- Use POS tagging – **slows things down**
- Bigrams, trigrams, … **- straight to 50-grams?**
- Use (uniquely) resolvable encodings – **computational cost**
  - **also, brittle, susceptible to brute force and key proximity not necessarily indicative of data similarity…..**
- Scale?

Solving scale - fat cat consultants?



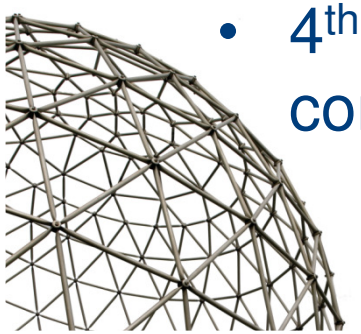


As Simon Wardley, Leading Edge Forum, might present it

At scale?

UNIVERSITY OF SURREY

- In 2011, we used **one** virtual core in a **single** High-Memory Quadruple Extra Large Instance (m2.4xlarge) instance.
  - Spec: *68.4 GB of memory; 26 EC2 Compute Units (8 virtual cores with 3.25 EC2 Compute Units each); 1690 GB of instance storage; 64-bit platform.*
  - $2 per hour; first run: $4.
  - ***We had to wait for result submission to open.***
- **If we had been given a 15 minute talk then, we could have demonstrated (the core of) our system live.**
  - **750,000 by 750,000 RCV1 documents took us 36 minutes, so we'd need a bit longer on that.**
- 4th in external, and between the 5th and 6th placed competitors from PAN 10.

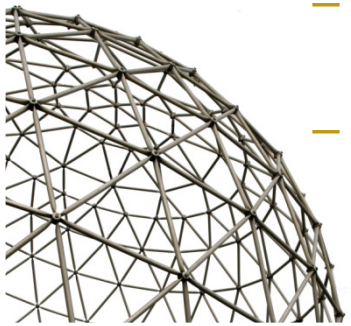


In 2012?

UNIVERSITY OF SURREY

- Competition changed completely
  - Use a search engine … have to **expose the queries** AND retrieve complete documents!
  - Pairwise match on results … computationally costly if you could get good matches at the right grain directly from the index.

- Not the direction we want to go in
  - Today (literally) we're building our approach using the ClueWeb09 dataset. **Really scale! (but still quite small?)**
  - Will take an estimated 2.5 weeks to create our first full index of the English portion.
  - Index estimated to be < 6GB. SATA III SSD speed 6GB/s; 6GB memory on a laptop?
  - Then, evaluate using PAN12 CR collection? (Where are the answers?)
  - Should easily be reportable next year.



For PAN 2012

UNIVERSITY OF SURREY

- Educated guesses? Candidate Retrieval in one quite simple (elegant?) equation and relatively few steps:

$$ew = \frac{N_{GL} f_{SL}^{\ 2}}{(1 + f_{GL}) N_{SL}^{\ 2}}$$

For each suspicious text, ***T***:

Split to sub-texts ***S*** by number of lines *l (=25)*.

For each sub-text in ***S***, generate queries ***Q*** by:

Rank by ***ew***.

Select the top 10 terms, and re-rank by frequency

top frequency-ranked word paired with the next ***m (=4)*** words

Retrieve texts for each query in ***Q***.

Pairwise match to find real results

- Equality judgements? Our approach remains under wraps for now.
  - Better speed definitely possible – double-processing.
  - Also, quite a simple (elegant?) approach

CR - Who won?

UNIVERSITY OF SURREY

- **Candidate Retrieval:**

Candidate Retrieval Task*												
	Total Workload		Time to 1st Result		No result	Reported Sources		Downloaded Sources		Retrieved Sources	
Team	Queries	Downloads	Queries	Downloads		Precision	Recall	Precision	Recall	Precision	Recall
Gillam et al. University of Surrey, UK	63.44	527.41	4.47	25.88	1	0.6266	0.2493	0.0182	**0.5567**	0.0182	**0.5567**
Jayapal University of Sheffield, UK	67.06	173.47	8.78	13.50	1	**0.6582**	**0.2775**	0.0709	0.4342	**0.0698**	0.4342
Kong Leilei Heilongjiang Institute of Technology, China	551.06	326.66	80.59	27.47	2	0.5720	0.2351	0.0178	0.3742	0.0141	0.3788
Palkovskii et al. Zhytomyr State University, Ukraine	63.13	1026.72	27.28	318.94	6	0.4349	0.1203	0.0025	0.2133	0.0024	0.2133
Suchomel et al. Masaryk University, Czech Republic	**12.56**	**95.41**	**1.53**	**6.28**	2	0.5177	0.2087	**0.0813**	0.3513	0.0094	0.4519
Gillam et al. University of Surrey, UK	63.44	527.41	52.38	445.25	22	0.0310	0.0414	0.0016	0.0526	0.0019	0.0526
Jayapal University of Sheffield, UK	67.06	173.47	39.00	115.13	16	0.0328	0.0394	0.0079	0.0994	0.0108	0.0994
Kong Leilei Heilongjiang Institute of Technology, China	551.06	326.66	440.59	274.06	21	0.0280	0.0458	0.0019	0.0391	0.0015	0.0435
Palkovskii et al. Zhytomyr State University, Ukraine	63.13	1026.72	54.88	881.34	25	0.0246	0.0286	0.0002	0.0286	0.0002	0.0364
Suchomel et al. Masaryk University, Czech Republic	12.56	95.41	11.16	93.72	30	0.0208	0.0124	0.0007	0.0124	0.0003	0.0208

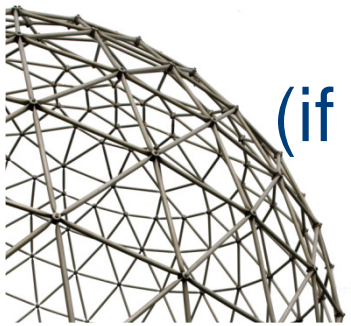
*) Values are averages over the 32 suspicious documents from the test corpus. The top half of the table shows performances when interpreting near-duplicates of the actual source documents as true positives; the bottom half of the table shows performances

Our Challenge

UNIVERSITY OF SURREY

How do you do efficient plagiarism detection that can scale to the entire (deep) web AND be useful across (private) corporate resources and across (private) corporates?

We want a good answer quickly … at the speed of search?



We might tell you how at PAN 13!

(if not too far away from our direction of travel)

UNIVERSITY OF SURREY

Keep It Stupid-Simple
(and don’t call people stupid)

Thank you.