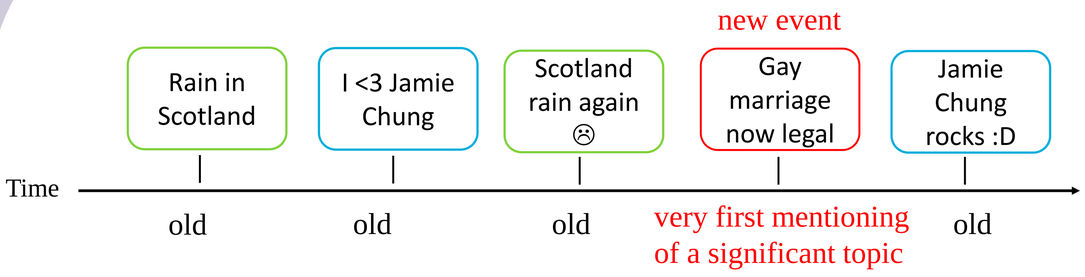
Twitter-scale New Event Detection via k-term hashing

Dominik Wurzer
d.s.wurzer@sms.ed.ac.uk
University of Edinburgh

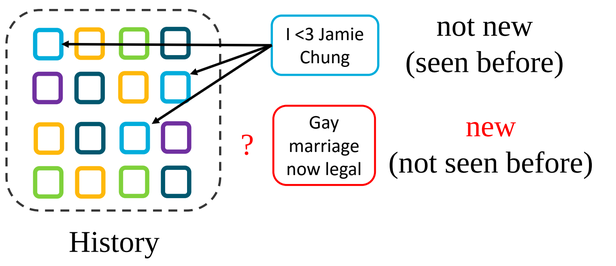
Victor Lavrenko
vlavrenk@inf.ed.ac.uk
University of Edinburgh

Miles Osbrone
mosborne29@bloomberg.net
Bloomberg London

Task: New Event or First Story Detection (FSD)



How people do it



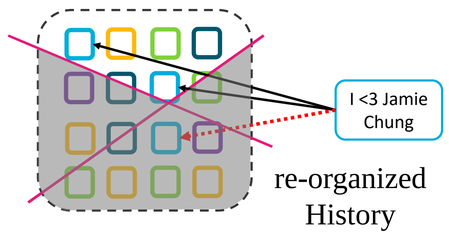
not new (seen before)

new (not seen before)

Exhaustive (UMass): (Allan et. al, 2000)

+ state-of-the-art accuracy

- $O(n^2)$ runtime



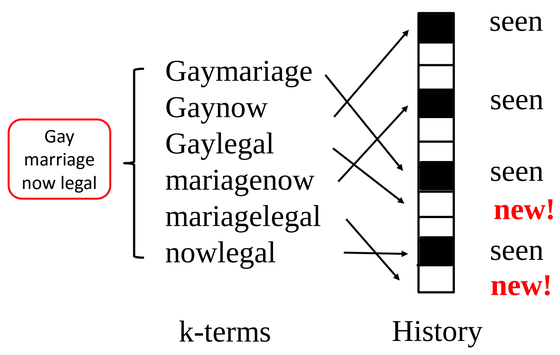
LSH: (Petrovic et. al, 2010)

+ gain in efficiency

- drop in effectiveness

How we do it

In contrast to previous approaches to FSD, we do not compare each new document to those previously seen. Instead, we maintain a single look-up structure that captures the history. Novelty is estimated based on counting what fraction of a document's content is new with respect to the history.



$$Novelty(d_n) = \sum_{t \in c_n} \alpha_{|t|} \binom{|d_n|}{|t|}^{-1} \begin{Bmatrix} 1 : t \in H_{\{n-1\}} \\ 0 : t \in H_{\{n-1\}} \end{Bmatrix}$$

d_n	= n[th] document in the stream
t	= non-empty set of up to k distinct terms
c_n	= set of all k-terms for document d_n
$\alpha_{\|t\|}$	= weight associated with k-terms of size t
$H_{\{n-1\}}$	= history H capturing information of documents before d_n

For each new document d_n , we form all k-terms - compounded sets of length k based on terms in d_n - and check their membership in history $H_{\{n-1\}}$.

To ensure constant look-up and update time, we maintain the history H as a Bloom Filter (Bloom, 1970) and hash k-terms onto a fixed length array. Constant error rate is guaranteed by placing an upper limit on the load-factor.

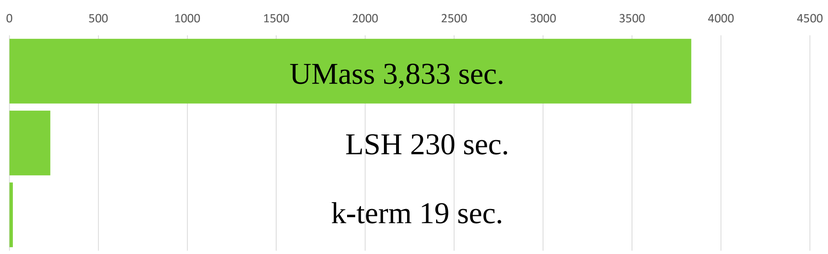
Experiments

We compare our approach using the official TDT evaluation procedure (Allan, 2002) with two **baselines** based on the **data set** by Petrovic (2013), which consists of 116,000 tweets and 27 topics :

- **UMass** – state-of-the-art high accuracy FSD system (Allan et. al, 2000)
- **LSH** – state-of-the-art high efficiency FSD system (Petrovic et. al, 2010)
- **k-term** – our approach based on k-term hashing

Results: Runtime

When processing 116,000 tweets k-term hashing performs 197 times faster than UMass and 12 times faster than LSH based hashing.

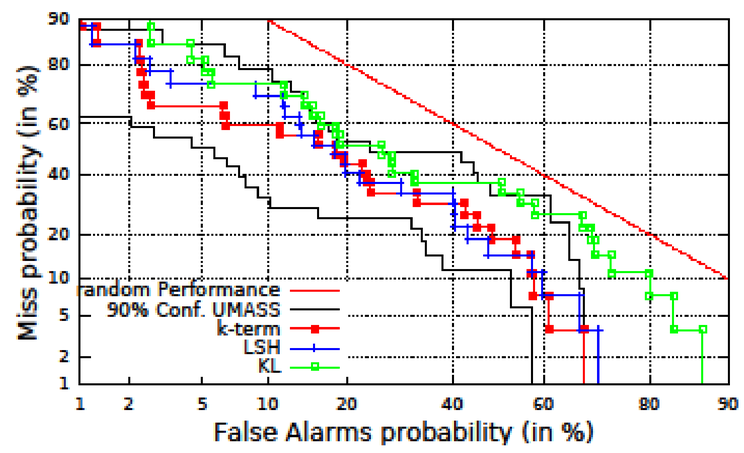


Result: Accuracy

We evaluated accuracy by the normalized Minimum Detection Cost (Cmin) , the official TDT (NIST, 1998-2004) accuracy metric (lower = better);

Algorithm	C_{min}	% - diff
UMass	0.7981	-
LSH	0.9061	- 13%
k-term	**0.7966**	+0.2%

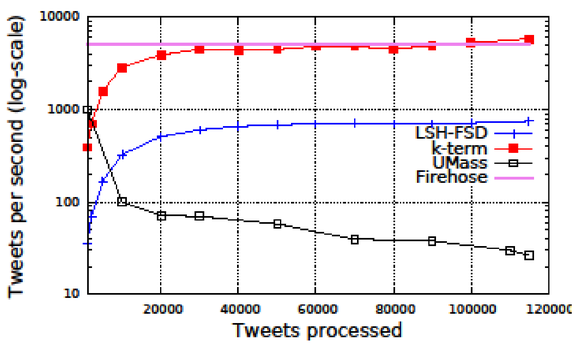
UMass shows state-of-the-art performance (C_{min} = 0.79) and exceeds LSH (C_{min} = 0.90), which trades accuracy against gains in efficiency by 13%. In contrast, k-term (C_{min} = 0.79) operates on par with UMass, while being 197 times faster.

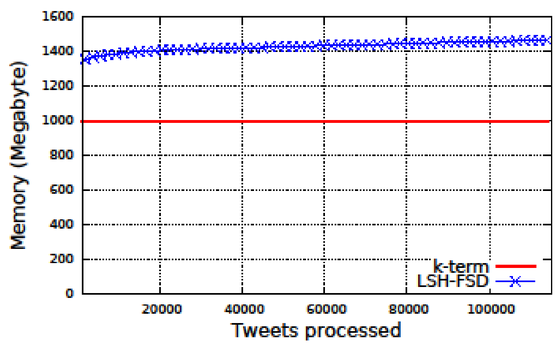


The Detection Error Trade-off plot (Allan, 2002) reveals that the observed accuracy of k-term is statistically indistinguishable from UMass for any Miss/False-Alarm trade-off point.

Result: Constant Time and Space

New Event Detection on streams requires operation in constant time and space to keep computation feasible.





While both, k-term and LSH operate in constant time, the throughput of k-term exceeds LSH by more than an order of magnitude. UMass becomes progressively slower over time

k-term remains in strictly constant space, while the memory footprint of LSH-FSD gradually increases over time.

Conclusion

We presented a novel approach to FSD based on k-term hashing, which allows scaling to high volume streams.

Computing novelty based on the fraction of a document's new k-terms results in accuracy which is statistically indistinguishable from UMass while performing 197 times faster.

Although running on a single core of modest hardware, k-term hashing is capable of processing the entire Twitter stream with 5,787 tweets/sec without sacrificing accuracy.

References

- James Allan, Ron Papka, Victor Lavrenko. 1998. Online new event detection and tracking; ACM
- James Allan, Victor Lavrenko, Hubert Jin. 2000. First story detection is hard. ACM
- James Allan. 2002. Topic Detection and Tracking: Event based Information Organization; Kluwer
- Burton H. Boom. 1970. Space/time trade-offs in hash coding with allowable errors; ACM
- Sasa Petrovic, Miles Osborne, Victor Lavrenko. 2010. Streaming first story detection; HLT
- Sasa Petrovic. 2013 Real-time event detection in massive streams. PhD thesis; University of Edinburgh
- TDT by NIST 1998 -2004 www.itl.nist.gov/iad/mig/tests/tdt (Last Update: 2008)