

Studying the Temporal Dynamics of Word Co-Occurrences: An Application to Event Detection

Daniel Preoțiuc-Pietro
Srijith P.K., Mark Hepple, Trevor Cohn
LREC 2016

Positive Psychology Center
University of Pennsylvania

 **Penn** | World Well-Being Project

22 March 2016

Word Co-occurrence

Discover events based on temporal text variation

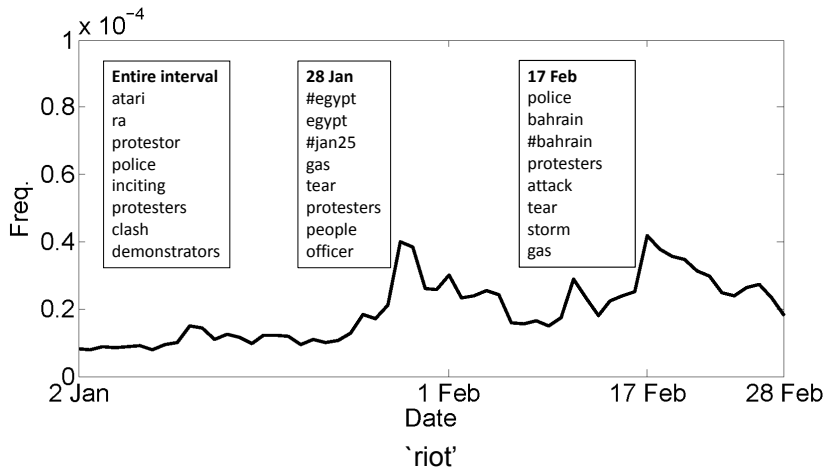
Word co-occurrence (e.g. NPMI) computed over large, static corpora:

- similar concepts or collocations

Computed over data from social media that reflects timely events (e.g. Twitter)

- current events & news

Example



Method

Cluster words (cf. messages) in a time interval

Fix time frame and compute word co-occurrence metrics

$$\text{NPMI}(X, Y) = -\log P(x, y) \cdot \log \frac{P(x, y)}{P(x) \cdot P(y)},$$

(NPMI) is bounded in the $[-1, 1]$ and has a easy interpretation:

- word pairs with a negative NPMI co-occur less often than expected under independence
- positive NPMI means more often than independence
- the maximum value $\text{NPMI}=1$ implies that both words appear exclusively together

Normalized PMI

Word1	Word2	NPMI	Type
arrests	yemen	0.699	news
publish	trailers	0.678	news
bestfriends	forming	0.678	news
g-slate	spotted	0.675	news
activist	arrests	0.674	news
china's	stealth	0.674	news
blake	griffin	0.672	proper name
magazines	merchandise	0.669	news
activist	yemen	0.667	news
actors	showcase	0.667	news
cameras	g-slate	0.664	news
angeles	los	0.662	proper name

Table: Top NPMI values for 23 Jan 2011, 9-10am. Word1, Word2 are in alphabetical order.

Clustering

Cluster words in a time frame to uncover terms that uniquely characterize an event

We use spectral clustering (*Shi & Malik, 2000*) using NPMI as similarity measure

Spectral clustering partitions a graph where edges are weighted using word co-occurrences.

We have to specify k – number of clusters

Sample Events



Query: Kubica crash

Label: Formula 1 driver Robert Kubica injured in rally crash

Coherence: 0.47, Magnitude: 140

Date: 06 Feb 2011, 12-1pm

Sample Events

BBC Sign in News Sport Weather Capital TV Radio Mo

SPORT **FORMULA 1**

Sport Homepage Page last updated at 00:00 GMT, Monday, 7 February 2011

Formula 1

Results

Standings

Race Calendar

Drivers & Teams

Gossip

BBC F1 team

Circuit Guide

A-Z of Sports ▾

Related BBC sites

News

Weather

Sport Relief

E-mail this to a friend

Printable version

Robert Kubica's F1 career at risk after rally crash



Kubica's car shows severity of crash

SEE ALL

- Can K 07 Fe
- Robe 19 Nc
- Kubic 03 Fe
- F1 te: 03 Fe
- Braw 03 Fe
- Thurs 03 Fe
- Alons 02 Fe
- Team 31 Ja
- Renai 14 Nc

Robert Kubica's Formula 1 career is under threat after his right hand was partially severed in a high-speed rally crash in Italy on Sunday.

NEWS

Home

Video

World

US & Canada

UK

Business

Tech

Science

Magazine

Ent

World

Africa

Asia

Australia

Europe

Latin America

Middle East

Moscow bombing: Carnage at Russia's Domodedovo airport

© 24 January 2011 | [Europe](#)



Sample Events



Query: Oprah Winfrey half-sister

Label: Oprah Winfrey has a half-sister

Coherence: 0.29, Magnitude: 43

Date: 24 Jan 2011, 9-10pm

Entertainment & Arts

Oprah Winfrey reveals half-sister on TV show

🕒 25 January 2011 | [Entertainment & Arts](#)

Oprah Winfrey has revealed she has a half-sister she never knew she had.

Winfrey introduced her sibling to viewers on her show, who was given up for adoption by Winfrey's mother nearly 50 years ago.

Patricia, 48, had been trying to trace her birth mother for years and eventually discovered her via a Wisconsin adoption



Winfrey's half-sister was only introduced as Patricia

Quantitative Analysis

First Story Detection Dataset:

2,228 annotated tweets – 27 events during June 2011 to September 2011 (*Petrovic et al. 2012*)

85,000 tweets from the same interval as background

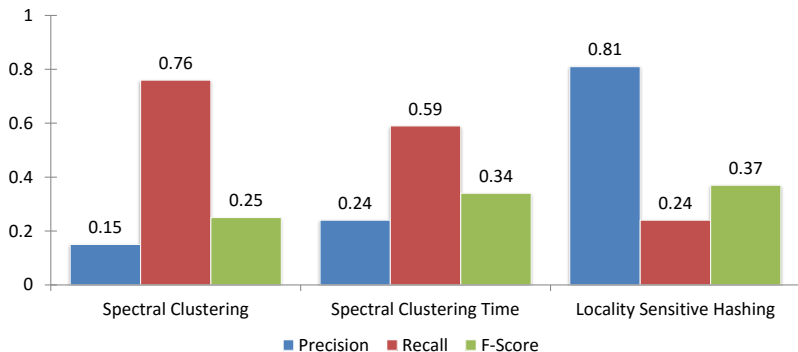
London Riots Dataset:

10,000 annotated tweets – 7 events during 6–16 August 2011

2,5M tweets as background

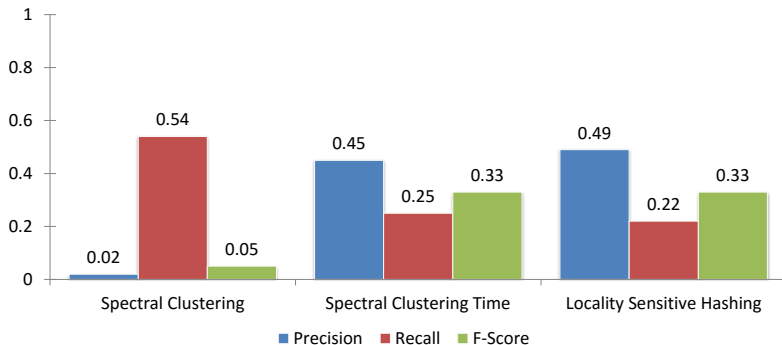
Events all about the London riots – more challenging as the events are related

Quantitative Analysis



First Story Detection Dataset

Quantitative Analysis



London Riots Dataset

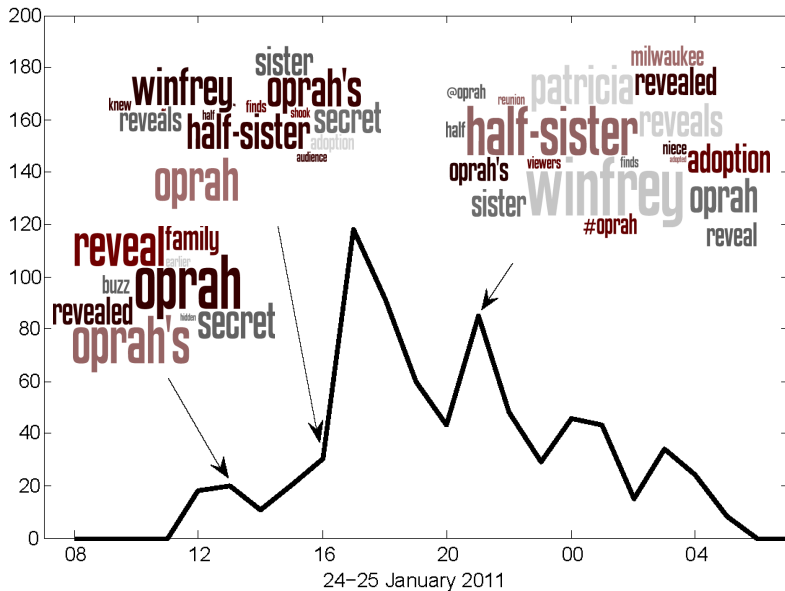
Longitudinal Analysis

evolutionary spectral clustering methods

create consistent clusters across consecutive time windows

discover event evolution and content drift over time

Longitudinal Analysis



Take Aways

Social Media data is highly time dependent text has different properties conditioned on time

By modelling time we gain a better understanding of real world effects

Social Media can be used to uncover real world events

Thank You!

Thank you!
Questions?