EuroVis 2017 Machine Learning Methods in Visualization for Big Data 2017

12 June 2017, Barcelona, Spain

# Visualisation and Mining of Text Data

#### Daniel Archambault<sup>1</sup>

1 Swansea University

## **My Experience**

- Post-doctoral postion at Clique UCD
- One of two information visualisation researchers in a lab of machine learning researchers
- Worked on social media analytics and in particular text analysis

## All ICWSM and SocMedVis

- SocMedVis was run at AAAI International Conference on Weblogs and Social Media in 2012 and 2013
- Top conference in data mining community for social media analytics
- Explored good ways for data mining and information visualisation to work together in analysing social media

- Early work to visualise hierarchically clustered Twitter data
- Information is a text document per user
- Represented as a high dimensional vector (bag of words) and clustered based on similarity

Daniel Archambault, Derek Greene, Pádraig Cunningham, Neil Hurley. ThemeCrowds: Multiresolution Summaries of Twitter Usage. CIKM 3rd International Workshop on Search and Mining User-generated Contents (SMUC 2011), 77-84, 2011.





В		D	
С	Α	E	
		F	

- Topics Clustered Hierarchically
- Antichains used to control visualisation

- What about sentiment?
- The same idea as ThemeCrowds but applied to sentiment bearing words to find clusters of strong sentiment

Anthony Brew, Derek Greene, Daniel Archambault, and Pádraig Cunningham. Deriving Insights from National Happiness Indices. ICDM SENTIRE Sentiment Elicitation from Natural Text for Information Retrieval and Extraction, 53-60, 2011.





- After filtering out noise clusters events begin to emerge
- Events clarified through visualisation of text information at macroscopic scale

## **Strong Topics or Sentiment**



 If hierarchy does not matter, the topics can be represented using lists

		SentireCro	owds		
Search: osama		+	- +/-	100%	
U: 24539	U: 674	U: 948	U: 685	U: 583	
ave	laden	laden	laden	laden	obama
good	osama	osama	osama	osama	laden
time	dead	dead =	obama	obama	president
park	obama	obama	death	photos	ground ≡
york	killed	#obl	dead	release	great
love	news	news	ladens	dead	osama
one	president	death	killed	president	back
great	finally	killed	one	ladens	show
night	#osama	world	pakistan	death	check
street	death	president	#obl	obl	#fb
≡	U: 557	U: 897	U: 24186	U: 25058	U: 25732
	obama	laden	ave	ave	mayo
	laden	osama	good	good	ave
	osama	dead	love	love	good
	mission	sea	time	time	happy
	dead	buried	york	one	love
	accomplished	#osama	one	york	time
	news	death	work	back	york
	killed	world	back	work	one
	president	body	great	great	back
	death	ladens	home	home	work
2011-04-30 20	U: 545	U: 600	U: 10685	U: 7275	U: 7551
	trump	death	good	ave	ave
	death	laden	love	york	york
	certificate	osama	time	blvd	blvd
	donald	ladens	one	chicago	san
	obama	obama	work	street	street
	ladens	dead	great	san	chicago

\_\_\_\_\_\_ Showing 61 to 66 of 82



— Showing 32 to 37 of 90

## Irish Blog Study

- PhD student Interested in English usage in Ireland
- Identify representative blogs in the Irish Blogosphere
  - extract blogroll links and historical posts
- Is there a way to dentify representative blogs through data mining and visualisation?

Karen Wade, Derek Greene, Conrad Lee, Daniel Archambault, and Pádraig Cunningham. Identifying Representative Textual Sources in Blog Networks. International AAAI Conference on Weblogs and Social Media (ICWSM 2011), 393-400.

## **Text Similarity Euler Diagram**



Discussion blogs need further decomposition

## **Blogroll Link Communities**



#### • Further topics based on community detection

## **Recommended Blogs**

Theme	Representative Blog
Beauty	** beaut.ie
Education/Law	** cearta.ie
Fashion	blanaid.com
Food	** icanhascook.wordpress.com
Gaelic	miseaine.blogspot.com
Humor	counago-and-spaves.blogspot.com
Movies	scannain.com
Music	** irishtimes.com/blogs/ontherecord
Personal	anonomousangel.wordpress.com
Photos	slkav.com
Politics	splinteredsunrise.wordpress.com
Sport	dangerhere.com
Technology	** mulley.net
Wine	firstpress.blogspot.com

- Fourteen Blog Recommendations
- Blogs would be missed if only degree ranking of blogs used

## **STAR Text Analytics**

F. Wanner, A. Stoffel, D. Jäckle, B. C. Kwon, A. Weiler and D. A. Keim. State-of-the-Art Report of Visual Analysis for Event Detection in Text Data Streams. EuroVis STAR Reports, 2014.

- Looks at visualisation and mining of text from a variety of perspectives
  - data sets analysed
  - data mining and analytics techniques used
  - visualisation approaches used
- Presentation of work is comprehensive and provides a very good overview