EuroVis 2017
Machine Learning Methods in Visualization for Big Data 2017
12 June 2017, Barcelona, Spain

Visualisation and Mining of Text Data

Daniel Archambault[1]

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

AIII ICWSM and SocMedVis

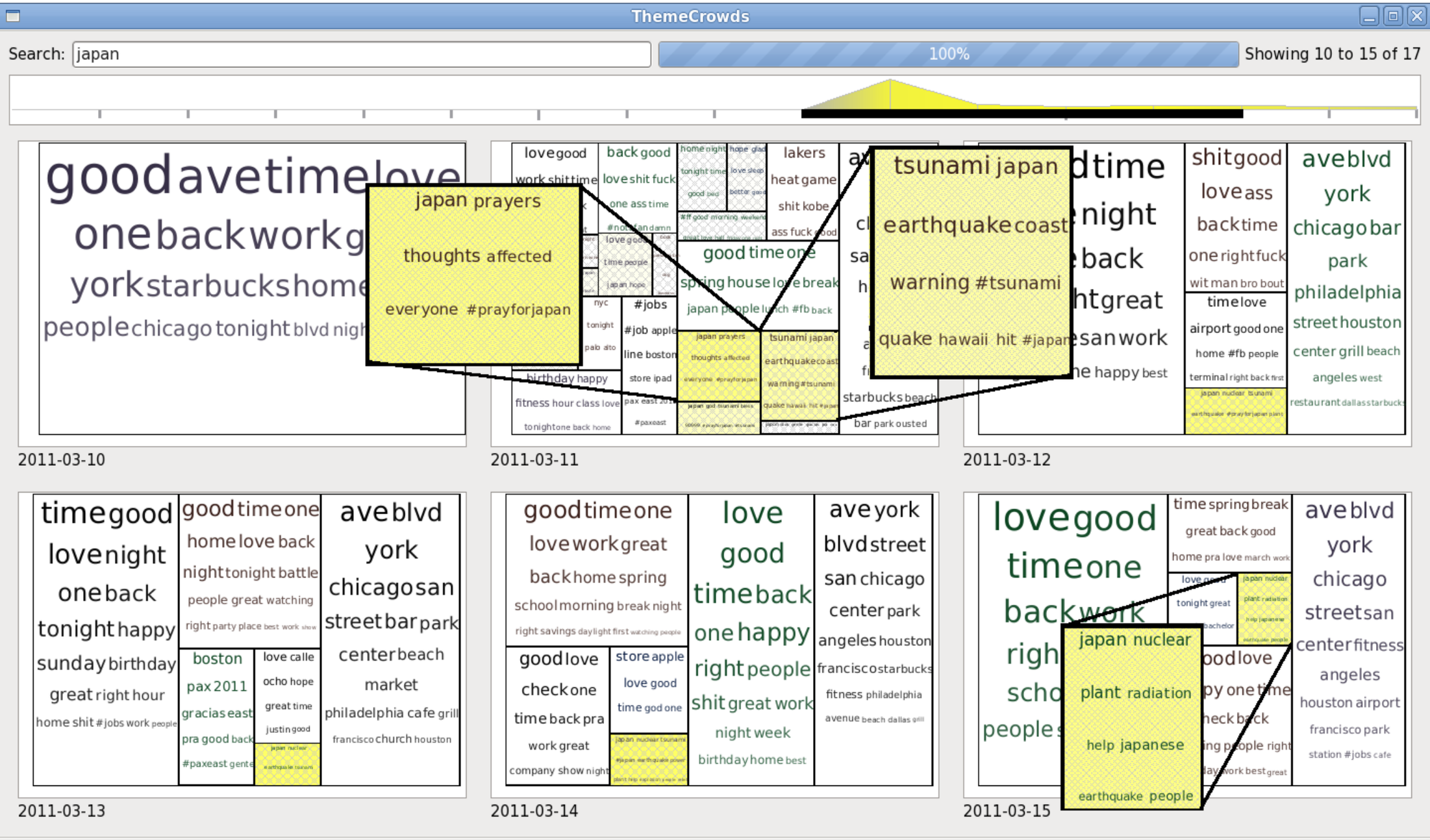
- SocMedVis was run at AAAI International Conference on Weblogs and Social Media in 2012 and 2013
- Top conference in data mining community for social media analytics
- Explored good ways for data mining and information visualisation to work together in analysing social media

ThemeCrowds

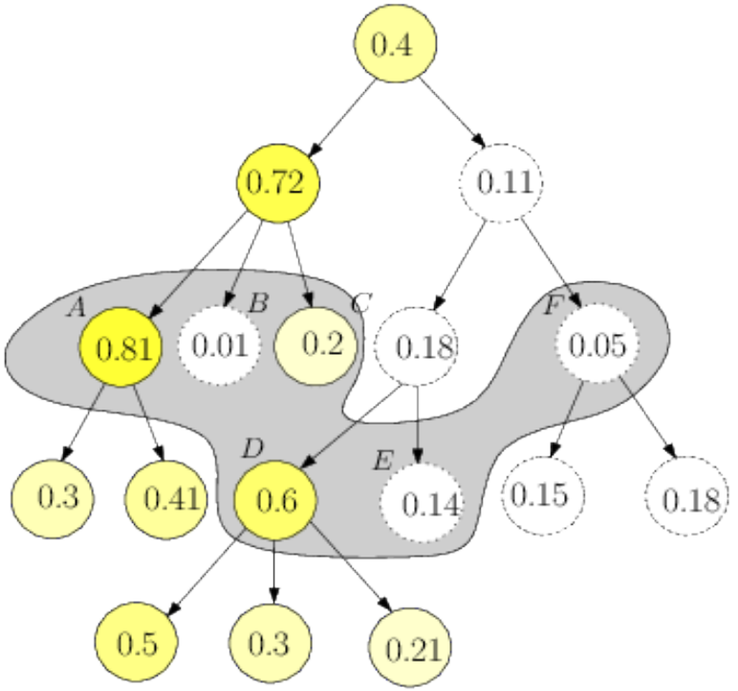
- Early work to visualise hierarchically clustered Twitter data
- Information is a text document per user
- Represented as a high dimensional vector (bag of words) and clustered based on similarity

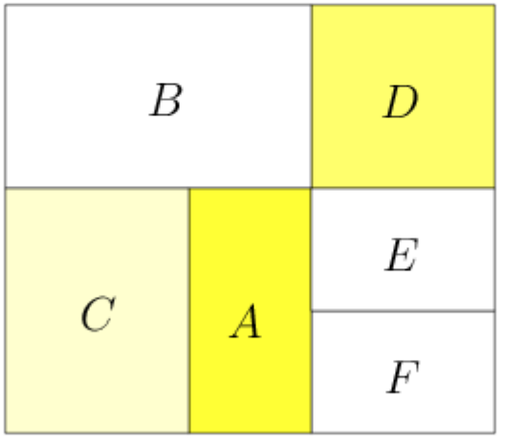
Daniel Archambault, Derek Greene, Pádraig Cunningham, Neil Hurley. ThemeCrowds: Multiresolution Summaries of Twitter Usage. CIKM 3rd International Workshop on Search and Mining User-generated Contents (SMUC 2011), 77-84, 2011.

ThemeCrowds



ThemeCrowds





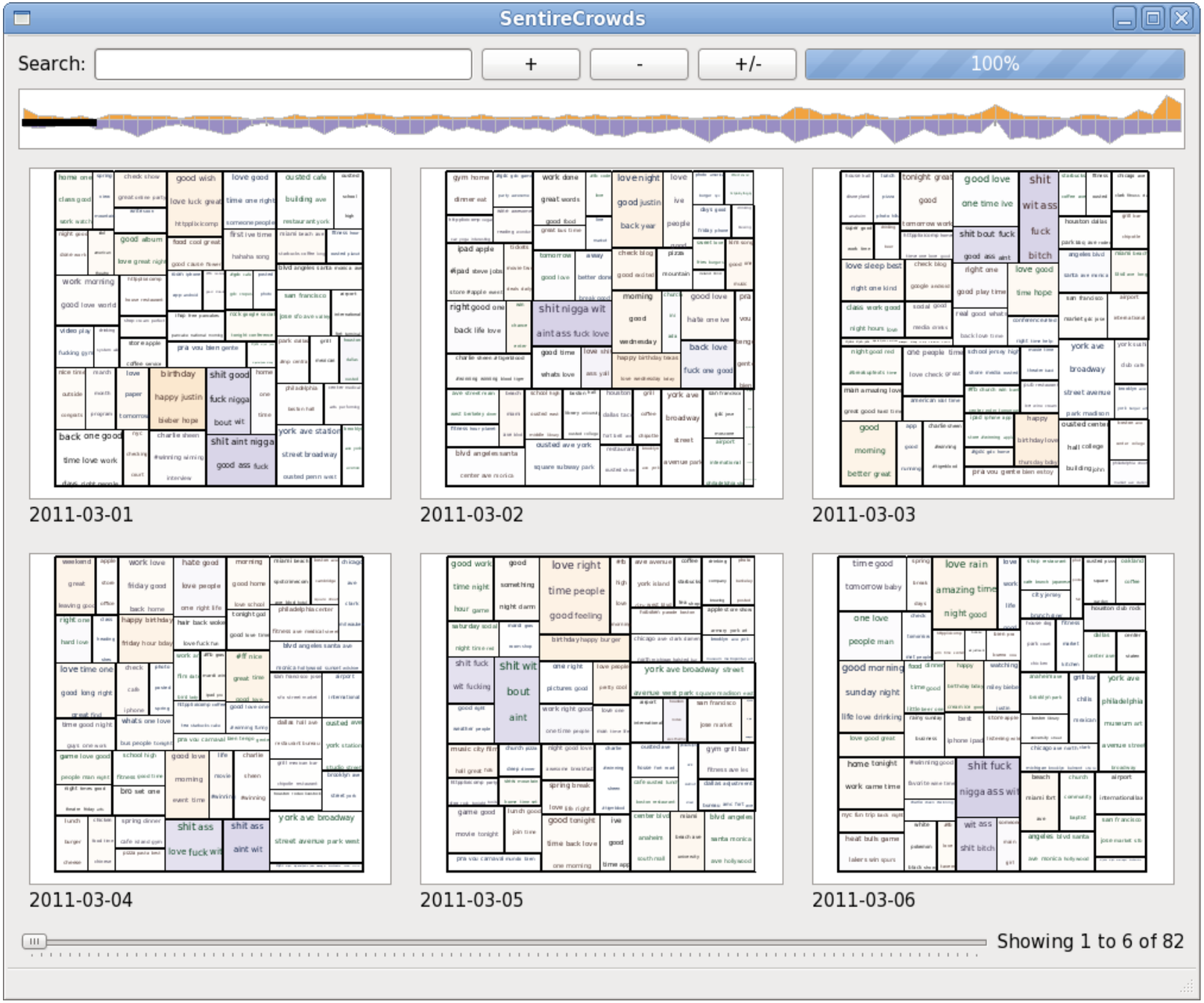
- Topics Clustered Hierarchically
- Antichains used to control visualisation

SentireCrowds

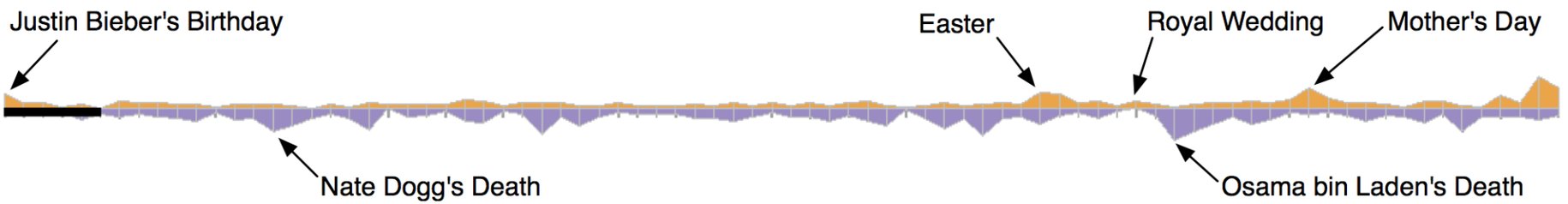
- What about sentiment?
- The same idea as ThemeCrowds but applied to sentiment bearing words to find clusters of strong sentiment

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.

SentireCrowds

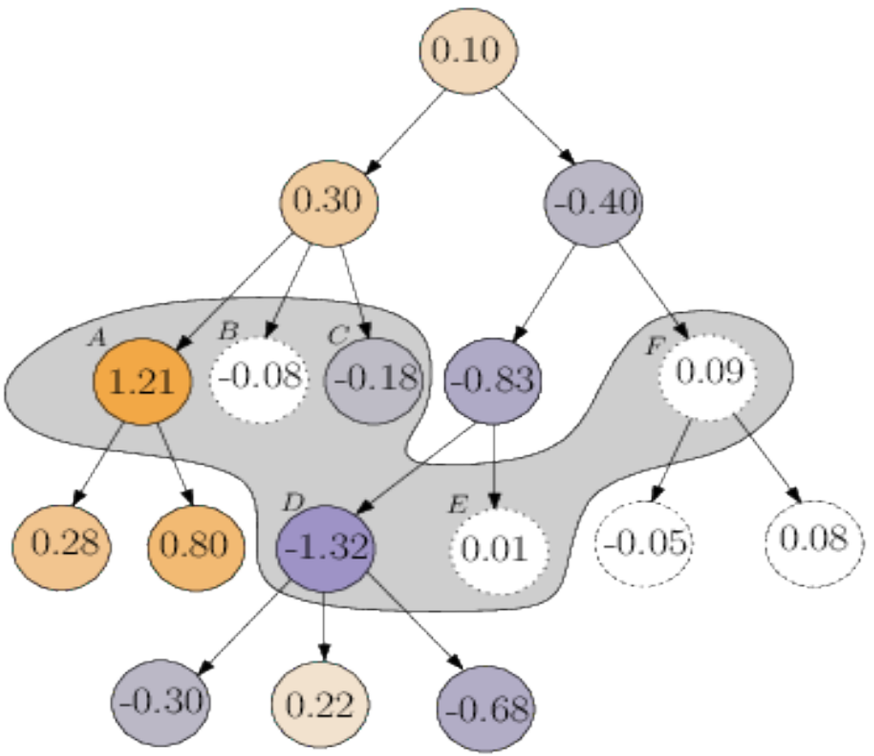


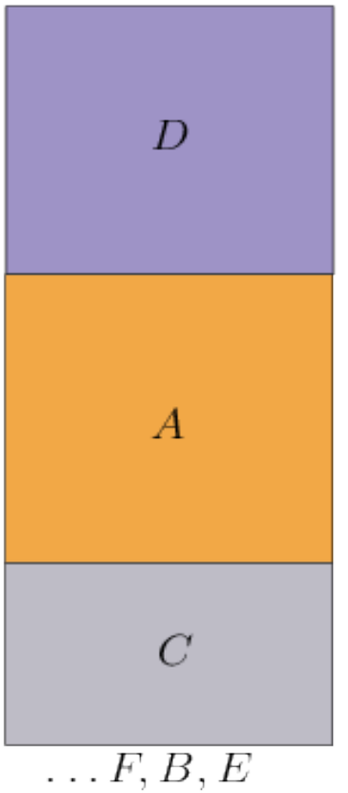
SentireCrowds



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

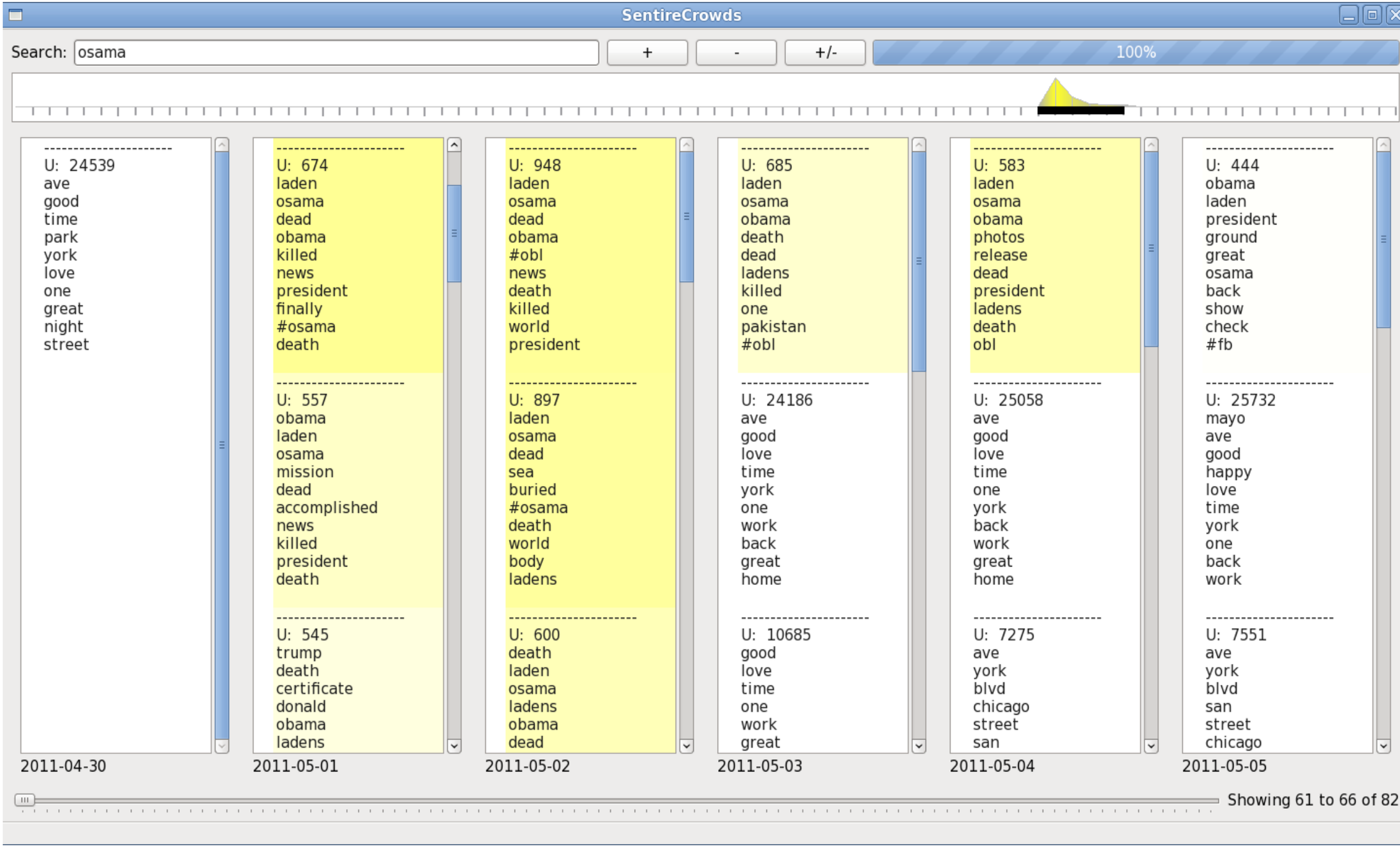
Strong Topics or Sentiment



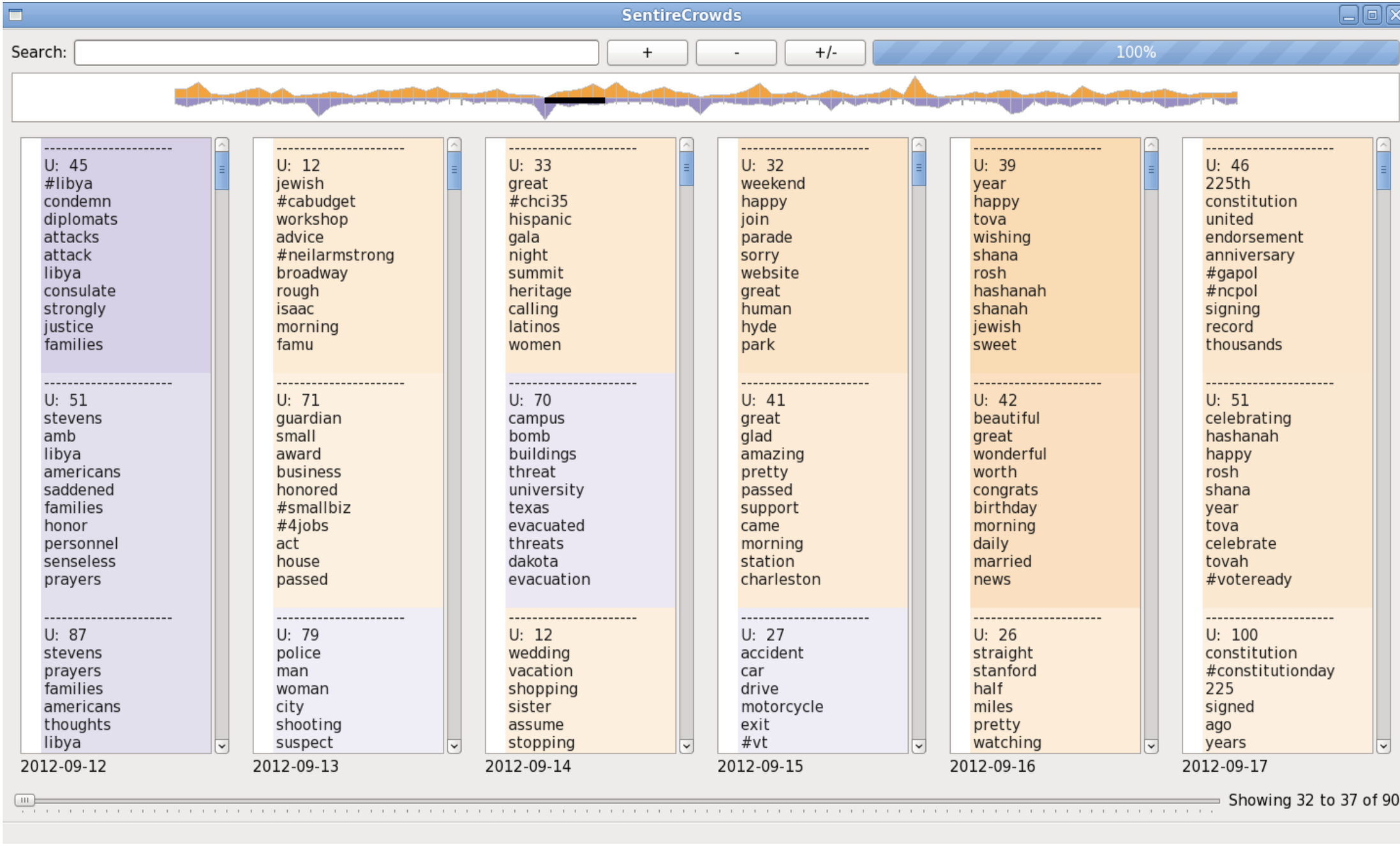


- If hierarchy does not matter, the topics can be represented using lists

ThemeCrowds



SentireCrowds



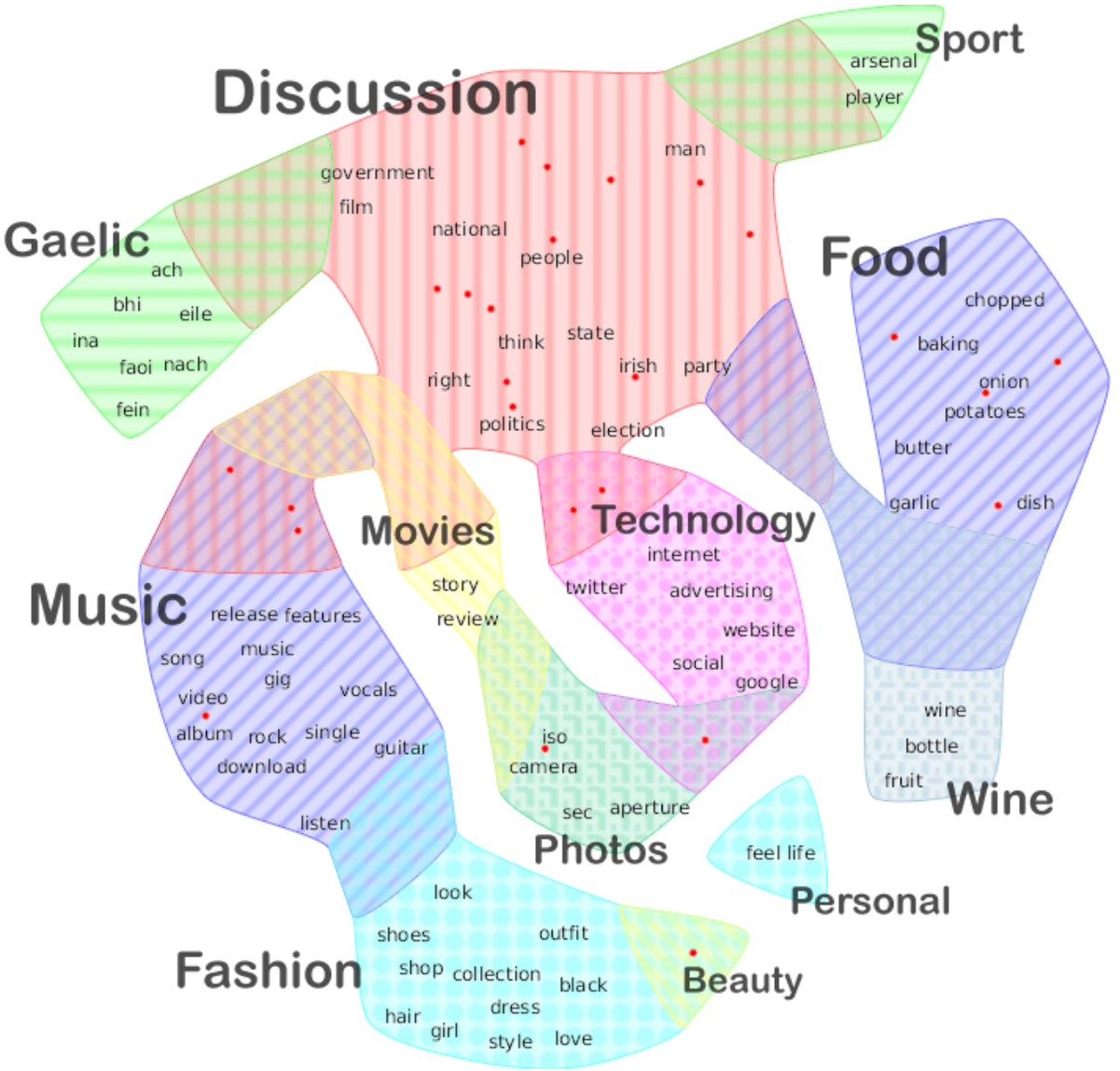
Irish Blog Study

- 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?

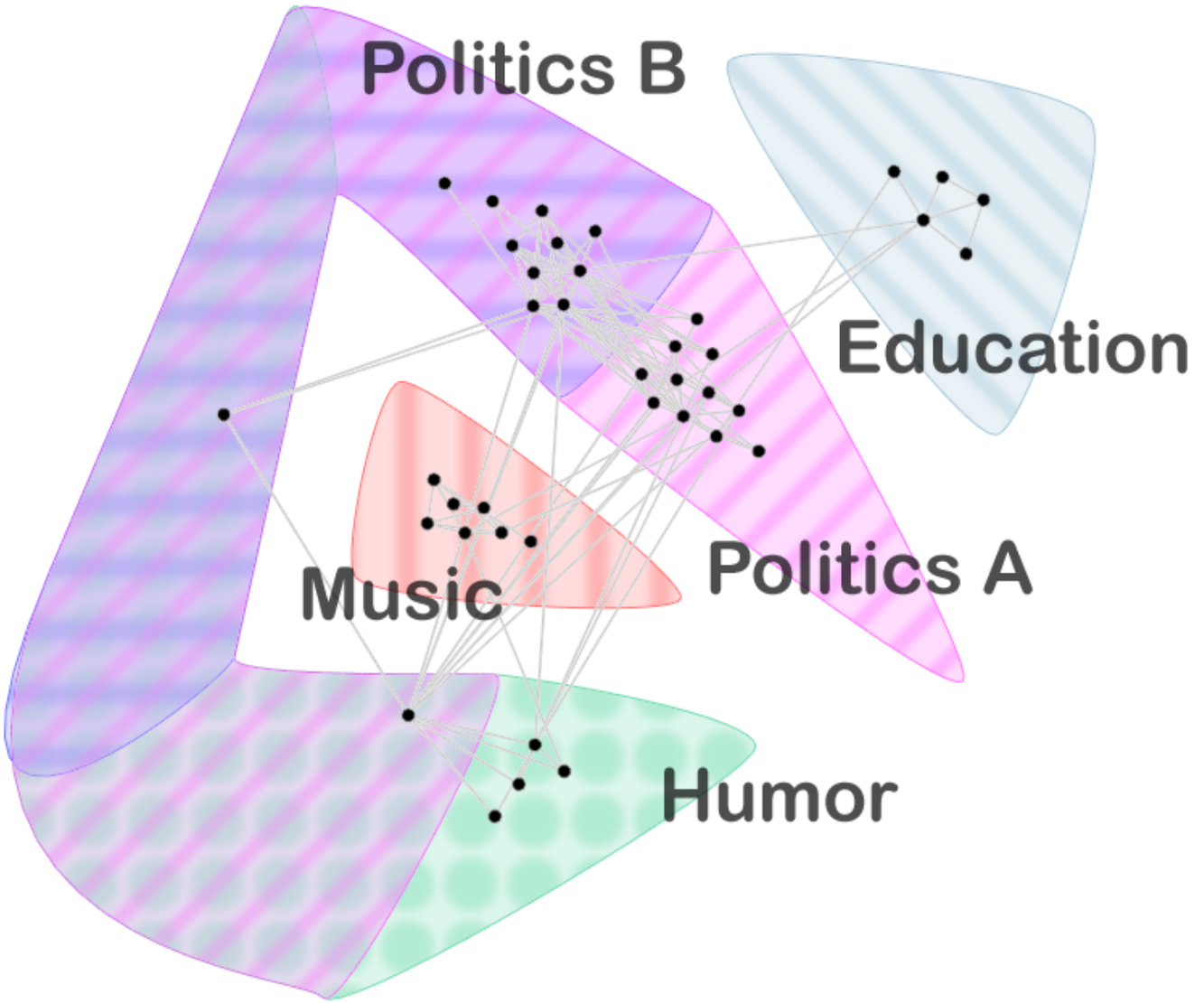
Karen Wade, Derek Greene, Conrad Lee, Daniel Archambault, and Pádraig Cunningham. Identifying Representative Textual Sources in Blog Networks. International AAAI Conference on Weblogs and Social Media (ICWSM 2011), 393-400.

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