The Collaborative Online Social Media Observatory

Rob Procter
Warwick University
www.wisc.warwick.ac.uk
www.cosmosproject.net
rob.procter@warwick.ac.uk
@robnprocter

Social Research Association
May 16th, 2014
BIG CITIES, BIG DATA
Analysing cities today, to plan for the cities of tomorrow
Open Platform for Social Data Analysis

- Forging interdisciplinary working between social and computing scientists.
- Provide open, sustainable platform for interoperable social media analysis tools.
- Refine and evolve capabilities, provide service models compatible with needs of diverse user communities.
Development Team

- William Housley
- Matt Williams
- Adam Edwards
- Pete Burnap
- Luke Sloan
- Omer Rana
- Jeffrey Morgan
- Alex Voss
- Chris Choi
- Mike Thelwall
- Ewan Klein
- Miranda Taylor
- Andy Hudson-Smith
- Panos Mavros
- Mark Birkin
- Kirk Harland
- Anita Greenhill
- Rob Procter
- Maria Liakata
COSMOS Tools

**User profiling**
- Gender
- Language
- Geo-location

**Tweet analytics**
- Sentiment
- Tension

**Corpus analytics**
- Re-tweet metrics
- Frequency metrics
- Network analysis
- Geo-clustering
- Topic discovery

**Data and Analytic Support**
- Data import
- Data mashing
- Data export
- Workflow composition
Digging into Data, Deriving Metrics & Visualisation

80 Tweets/second  80 Filter Matches/second

Gender Stats
- Male: 14.44%
- Female: 24.88%
- Unsure: 7.53%
- Unknown: 53.15%

Language Stats
- English: 42.11%
- Spanish: 13.20%
- Portuguese: 7.31%
- Arabic: 5.09%
- French: 3.97%
- Japanese: 3.34%
- Indonesian: 2.54%
- Dutch: 2.54%
- Russian: 2.07%
- Italian: 1.91%
- Somali: 1.91%
- Turkish: 1.75%
- Lithuanian: 1.27%
- All: 1.27%
- Romanian: 0.95%

Hashtags
- #MakingStopPedophilebeat (1)
- #rt (@ 1)
- #ecater (1)
- #emergens (1)
- #kingshilewinyagondor (1)
- #ask (1)
- #psocototackers (1)
- #emergens (1)
- #emergens (1)
- #emergens (1)
- #emergens (1)
- #emergens (1)
- #emergens (1)
- #emergens (1)

Tension Topics
- Streaming real-time data from 127:01
Inspecting Spikes and the Qualitative Element
Social Network Analysis
Current Research

- A study of community resilience as a mundane achievement:
  - Explore everyday threats to social resilience as perceived by communities and how are they expressed through social media.
  - Examine how do community members use social media as a tool for information gathering, dissemination and interacting with representatives of public agencies and services.

- Social Media and Prediction: Crime Sensing, Data Integration and Statistical Modelling (NCRM methods innovation programme):
  - Develop new curated data extraction and mashing tools to facilitate interoperability of social media, curated ONS and administrative police crime data sources.
  - Incorporate these data sources into predictive statistical models to reveal associations between various independent indicators and the dependent variable of daily crime rates.

- ‘Hate’ Speech and Social Media: Understanding Users, Networks and Information Flows (ESRC-Google):
  - Profile hateful and antagonistic social media networks to develop a typology of users (antagonists, influencers, propagators, reactors).
  - Use machine learning to forecast emergence and evolution of information flows within social media networks through which hate-related content is transmitted.

- Suicide Clusters and Social Media.
Community Resilience

Table 1: TWEETS

<table>
<thead>
<tr>
<th>Reference centrality</th>
<th>In-degree</th>
<th>Out-degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CARDIFFCOUNCIL</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>SWPEAST</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>SWPOLICE</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>HINTOFSCARASM</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>THEETHICALCHEF</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>JOHNFISHER</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>GETTGOOD</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>ASHLEYGOVER</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>DAVHYMYP</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>JENNYTHE</td>
<td>10</td>
</tr>
<tr>
<td>11</td>
<td>VAUGHANGETTHING</td>
<td>11</td>
</tr>
<tr>
<td>12</td>
<td>DOUGTHYMP</td>
<td>12</td>
</tr>
<tr>
<td>13</td>
<td>VAUGHANGETTHING</td>
<td>13</td>
</tr>
<tr>
<td>14</td>
<td>kevinbrennenrap</td>
<td>14</td>
</tr>
<tr>
<td>15</td>
<td>VAUGHANGETTHING</td>
<td>15</td>
</tr>
<tr>
<td>16</td>
<td>DOUGTHYMP</td>
<td>16</td>
</tr>
<tr>
<td>17</td>
<td>DAVHYMYP</td>
<td>17</td>
</tr>
<tr>
<td>18</td>
<td>LERKESP</td>
<td>18</td>
</tr>
<tr>
<td>19</td>
<td>CSC</td>
<td>19</td>
</tr>
<tr>
<td>20</td>
<td>peoplevspeople</td>
<td>20</td>
</tr>
<tr>
<td>21</td>
<td>peoplevspeople</td>
<td>21</td>
</tr>
<tr>
<td>22</td>
<td>peoplevspeople</td>
<td>22</td>
</tr>
<tr>
<td>23</td>
<td>peoplevspeople</td>
<td>23</td>
</tr>
<tr>
<td>24</td>
<td>peoplevspeople</td>
<td>24</td>
</tr>
<tr>
<td>25</td>
<td>peoplevspeople</td>
<td>25</td>
</tr>
<tr>
<td>26</td>
<td>peoplevspeople</td>
<td>26</td>
</tr>
<tr>
<td>27</td>
<td>peoplevspeople</td>
<td>27</td>
</tr>
<tr>
<td>28</td>
<td>peoplevspeople</td>
<td>28</td>
</tr>
<tr>
<td>29</td>
<td>peoplevspeople</td>
<td>29</td>
</tr>
<tr>
<td>30</td>
<td>peoplevspeople</td>
<td>30</td>
</tr>
</tbody>
</table>

Social network of Twitter communications in West Cardiff
Mashing Curated and Administrative Data with Twitter
Future Plans
Distribution of geo-located tweets around Boston area of United States from sample of Twitter data covering terrorist bombing of 2013 Boston marathon
Three co-ordinated views of a temporal selection of 2013 Boston marathon terrorist incident data.

Word cloud visualises frequency of most commonly used words in selected tweets between 8pm and 9pm on Monday 15th April.
Geographic distribution of tweets inside Olympic stadium.

Table view has user-selectable visual properties that map onto user-selectable data values. Positive sentiment mapped onto shade of red.
Topic Clustering

Topic Discovery 4227 total tweets

Cluster id: cl_30185666029285376
Cluster Features: 98

- oscar
- pistorius
- girlfriend
- dead
- shot
- allegedly
- shooting
- intruder
- mistaking
- appeared

Cluster tweets: 64

- "@itnews: Report: Oscar Pistorius to appear in court this morning after allegedly shooting dead his girlfriend http://t.co/WJqPktXn"
- Key News: Olympic champ Greg Rutherford tells us of his shock that Oscar Pistorius is arrested for shooting his girlfriend
- "@guardian: Latest: Oscar Pistorius arrested after girlfriend shot dead http://t.co/UQkgeph4 via @guardiannews @glover_david"
- Breaking news: Paralympic athlete Oscar Pistorius has allegedly shot killed his girlfriend after apparently thinking she was a burglar
- Report: Oscar Pistorius is being held at Pretoria Central Prison http://t.co/R3aV33a

Timeline

The linechart inside the timeline shows how many clusters there are at each step.

Cluster Development

The linecharts below show how the selected cluster develops in time through the entire dataset. Each chart displays one statistical property of the cluster. For more information, go to the documentation.

Cluster Comparison

This chart shows how the clusters compare to each other.
Deployment Plans

**COSMOS I**
- Limited storage and compute capacity
- Limited scalability

**COSMOS II**
- Scalable storage and compute
- New tools
- Workflow management
- Collaboration support

Cloud Hosted Environment:
- User Datasets
- Shared Workflow
- On-Demand Compute Nodes
Methodological Issues
“Huge potential exists to harness the power of crowdsourcing for the study of society and human behaviors… but it’s just not happening as well as it could… Social science issues are human issues, after all - they are about how we relate to each and organise our society and economy - so there seems to be a natural fit with crowdsourcing that’s largely being overlooked. This raises some obvious and legitimate concerns - from representation to research ethics and integrity - but none of these seem insurmountable. Indeed, social scientists would surely benefit from greater public engagement with their work. The prize is surely quicker, cheaper and more imaginative research - the findings from which could benefit us all.”

Harris, M. (2012). Oh man, the crowd is getting an F in social science, Daily Crowdsourse, June.
Barack Obama was elected as president. AND THE TWEETS COME IN.

is this text offensive or antagonistic in terms of race or ethnicity?

- Yes
- No
- Undecided

RT @SoYuSavage: Obama won that shit mane "webble voice"

is this text offensive or antagonistic in terms of race or ethnicity?

- Yes
- No
- Undecided

CNN: 249 Obama!!!

is this text offensive or antagonistic in terms of race or ethnicity?

- Yes
- No
- Undecided
Citizen Social Science
Twitter Sampling

- Identifying target data:
  - Matching against hashtags, keywords, accounts, language, location etc
  - Snowballing from seed keywords and accounts

- Interaction of sampling strategy and APIs:
  - The firehose, garden hose, 1% sample
  - Streaming API
    - Real-time search by accounts, keywords, etc
    - Rate limiting if results exceed 1% of firehose
  - Search API
    - Search by accounts, keywords
    - Rate limiting – results, requests
Summary

• Scaling coding effort through citizen social science:
  • Crowdflower, Mechanical Turk

• Sampling:
  “[..] social sciences have not developed an approach to robustly sample social media data sets.”


• The ‘computational turn’ in social sciences:
  “[..] use of automated tools [..] raises threat that researchers may treat them as black box technologies whose inner workings need not be understood as long as they produce outcomes that ‘look right’.”