Module 4: GATE and Social Media
Part 1: Introduction
Social Media: a digital sample of all human discourse?

What could we do with that?

What are we already doing with it?
Socioscope (Xu 2012) builds realtime maps of roadkill
• Treats tweets as observations, roadkill events as latent variables
• Normalisation for spatio-temporal reporting rates, human activity, animal activity

Evaluated against government cleanup figures

Fig. 2. Human population intensity $\hat{g}$.

Fig. 3. Socioscope estimates match animal habits well. (Left) range map from NatureServe, (Middle) Socioscope $\hat{f}$ aggregated spatially, (Right) $\hat{f}$ aggregated temporally.
Later research led to improved earthquake alerting systems

“We consider each Twitter user as a sensor and apply Kalman filtering and particle filtering, which are widely used for location estimation in ubiquitous/pervasive computing. The particle filter works better than other comparable methods for estimating the centers of earthquakes and the trajectories of typhoons.” - Sakaki 2010

“these feeds represent a hybrid form of a sensor system that allows for the identification and localization of the impact area of the event” (USGS) – Crooks 2012
Social media analysis

Retrospective analyses into cause and effect

"There's a dead crow in my garden"

Social media mentions of dead crows predict West Nile Virus in humans *

Epidemic prediction (flu)
Sadilek (2012) monitored geolocated tweets in greater NY area
• Built classifier for detecting whether a twitterer is unwell
• Monitor friends and collocated twitterers
• See if people become ill based on their social network and movement path

@mari: i think im sick ugh..

Result: predict whether an individual will become ill in the next week with 80% accuracy
Disaster response (fires)
Bushfires regular, dangerous occurrence in Australia
Large region makes it difficult to collect data
Further, difficult problem of distinguishing reports of fires from other fire mentions

Filtering false reports most useful outside of peak season
Uses transductive learning to bypass problem of generalising from noisy data
First deployed in Nov ‘14

Figure 2: Daily ‘fire’ Tweet counts.
Social media analysis

Ability to extract sequences of events

Retrieve information on:

- Lifecycle of socially connected groups
- Analyse precursors to events, post-hoc

* Weikum et al. 2011: "Longitudinal analytics on web archive data: It's about time", Proc. CIDR
Gartner "3V" definition:

1. **Volume**
2. **Velocity**
3. **Variety**

High volume & velocity of messages:

- Twitter has ~330 000 000 users per month
- They write ~500 000 000 messages per day

Massive variety:
- Stock markets;
- Earthquakes;
- Social arrangements;
Social media sites

Twitter, LinkedIn, Facebook

Twitter has varied uptake per country:
- Low in Denmark, Germany (Facebook is preferred)
- Medium in UK, though often complementary to Facebook
- High in USA

Networks have common themes:
- Individuals as nodes in a common graph
- Relations between people
- Sharing and privacy restrictions
- No curation of content
- Multimedia posting and re-posting

Other features: topics, liking, media, groups, person discovery ..

How can we get information out of these discussions, into a discrete machine-readable format?
Multiple sources & definitions of “social media” and “social network site”
Which to choose?

Twitter as the D. Melanogaster of social media

Newswire: regulated
• “our most frequently-used corpora […] written and edited predominantly by working-age white men”

Twitter: wild; many styles
• Headlines
• Conversations
• Colloquial
• Just “noise” (hashtags, URLs, mentions)
General challenges

Common complaints we have about social media text:
- Documents are short;
- Language is different to standard prose;
- Words are ambiguous;
- Nonstandard / new lexical items;
- New syntactic patterns;
- Memes

The impact (or the cause?) of these complaints: Low performance of existing systems. Maybe we need to re-train?
- Shortage of training data;
- Many different sub-domains
- Low-performance of existing techniques.

How can we characterise social media text? What new techniques can help us process it?

Let's start at the deep end: Twitter text.*

* also – it's public and plentiful
Hands-on Example

Let's compare ANNIE's ability to process news with processing tweets

- In GATE, create a new corpus called “News” or similar
- Create a datastore somewhere and save the corpus there
- Load the XML news articles from the hands-on folder into this corpus (in corpora/news-texts)
- Load ANNIE and run it over the corpus
- Look at the Token annotations, and the Persons, Locations and Organisations
- Create another new corpus, called “Tweets” or similar, in the DS
- Load the documents from the corpora/tweet-texts subdirectory
- Run ANNIE on this corpus

How are the annotations in the tweets? Text.category, entities, names