Current Developments in Event Data

Philip A. Schrodt

Parus Analytical Systems
schrodt735@gmail.com

March 29, 2014

Overview

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us.

Charles Dickens. *A Tale of Two Cities*

Paradox

There is now tremendous interest in event data

- 100 participants in GDELT Gallup "hackathon" in December 2013

Researchers are reluctant to use ICEWS because it is too expensive

- $50-million in US government investment
- Privatized version is being marketed at $150,000 per country

Researchers are reluctant to use GDELT because it is free

WTF???

Welcome to the new normal

- Rifkin [NYT March 2014]: The most disruptive technologies in the current environment combine network effects with zero marginal cost
- Key: zero *marginal* costs: open source software is "free-as-in-puppy"
- Examples
  - Operating systems: Linux
  - Statistical software: *R*
  - Encyclopedia: Wikipedia
  - Commercial photography: Shutterstock (55K photographers; 30K new images per day) vs Getty Images in $5B/year market

Open Source Software

Enterprise models

Twentieth-century: control as much of the environment as possible

IBM, Microsoft, Apple, Oracle, Dell

"You will be assimilated"

Twenty-first-century: build the structure, users provide the content

Google, eBay, Twitter, Facebook, post-2000 Amazon

YouTube, Pintrest, Reddit, Yelp, TripAdvisor, AirBNB

Open source models

Charismatic leadership

Emacs, GNU (Stallman), perl (Wall), early Linux (Torvalds) TEX(Knuth)

Organizational (profit or not-for-profit) development

FireFox, Apache, Android, mySQL,SourceForge, later Linux (IBM, RedHat, Ubuntu) OpenOffice/LibreOffice (also motivated by loathing of Microsoft)

Community development

LATEX, R, Python, Arduino micro controller, Raspberry Pi single-board computer, 3D printing

EL:DIABLO
Event Location: Dataset in a Box, Linux Option

- Full modular open-source pipeline to produce daily event data from web sources
- Scraper from white-list of RSS feeds and web pages
- Event coding from any of several coders: TABARI, PETRARCH, others
- Geolocation: Penn State "GeoVista" project coder, UT/Dallas coder
- Conventional reduplication keeping URLs of all duplicates
- Additional feature detectors are easily added

PETRARCH

- Python
- Full parsing using the Penn Treebank format and Stanford Core NLP
- Synonym sets from WordNet
- Identifies actors even if they are not in the dictionaries
- Extendible through program "hooks": "issues" facility

Sources for historical texts

- LDC Gigaword 2000-2010; easily licensed
- SPEED (Cline Center, University of Illinois at Urbana-Champaign)
- Usual proprietary sources, but these are awkward and expensive
- Collective resources: in the US, coded data on facts does not inherit the IP constraints of the source
- Discussion of legal issues for US: http://asecondmouse.wordpress.com/2014/02/14/the-legal-status-of-event-data/

Open Event Data Alliance

- Institutionalize per CRAN and many other groups
- 24/7/365 data reliability
- Common standards
- Open source, open access

What is to be done?

Goldilocks solution:
There is a lot of space between $50-million and free

Try to generate network effects comparable to those of LaTeX, *R*, Python and numerous other scientific communities

Parallel but compatible efforts:

- the era of "one data set to rule them all" has ended.
- At some point alternative coding decisions are *trade-offs*, not right/wrong.
- Survey research matured in this fashion: this enables ensemble "poll of polls" methods

Thank you

Email: schrodt735@gmail.com

Software: `https://github.com/openeventdata`

Software: `https://openeventdata.github.io/`

Papers: `http://eventdata.parusanalytics.com/papers.dir`

Slides:
`http://eventdata.parusanalytics.com/presentations.html`