

Big Data Potential of the Global Database of Events, Language, and Tone (GDELT)

Chuck Cartledge

October 25, 2016 at 9:50am

Table of contents I

1 Introduction

2 Data sources

3 Data exploration

4 Results

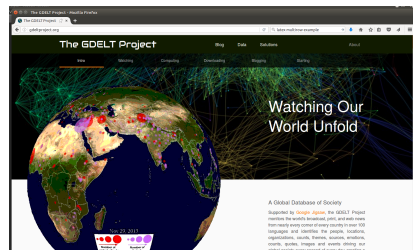
5 Conclusion

6 Files

A quick overview

Questions that were/are interesting:

- 1 What is the GDELT project
- 2 What data is available
- 3 What does the data have big data potential



GDELT claims the following:

- 1 Real-time Translation of 65 Languages.
- 2 Real-time Measurement of 2,300 Emotions and Themes.
- 3 High Resolution View of the Non-Western World.
- 4 Relevant Imagery, Videos, and Social Embeds.
- 5 Quotes, Names, and Amounts.



GDELT data is available on Google BigQuery

Basically:

- 1 You must have a Google account.
- 2 The first 1TB of data processed per month is free.
- 3 After the first 1TB, you are charged for the number of bytes processed, or stored.
 - 1 Processed: \$5 per TB
 - 2 Storage: \$0.02 per GB per month
- 4 Costs above the free tier are charged to your credit card on file with Google.

The screenshot shows the Google BigQuery web interface. On the left, there is a navigation pane with a tree view showing the project structure, including folders for 'gdeelt', 'gdeelt_public', and 'gdeelt_public'. The main area displays a 'Table Details' view for a table named 'events'. The table has columns for 'event_id', 'event_date', and 'event_text'. The 'event_text' column contains various news snippets, such as 'United to be southeast', 'Duke the westward push in YYYYMM format', and 'Alternative branding of the network, in YYYY format'. The interface includes a search bar at the top, a 'New Query' button, and a 'Table Details' tab selected.

It is easy to blow through 1TB

```
SELECT DATE, coord, cnt from (
SELECT DATE, coord, COUNT(*) as cnt
FROM (
select date, REGEXP_REPLACE(REGEXP_EXTRACT(
SPLIT(V2Locations,','))
,r'^[2-5]#.*?#.*?#.*?#.*?#(.*?#.*?)#')
, '^(.*)#(.*?)', '\\1;\\2') as coord
from [gdelt-bq:gdeltv2.gkg]
where DATE >= 20150322000000 and DATE <= 20150328999999 )
where coord is not null group by date, coord ORDER BY 3 DESC
where cnt >= 3;
```

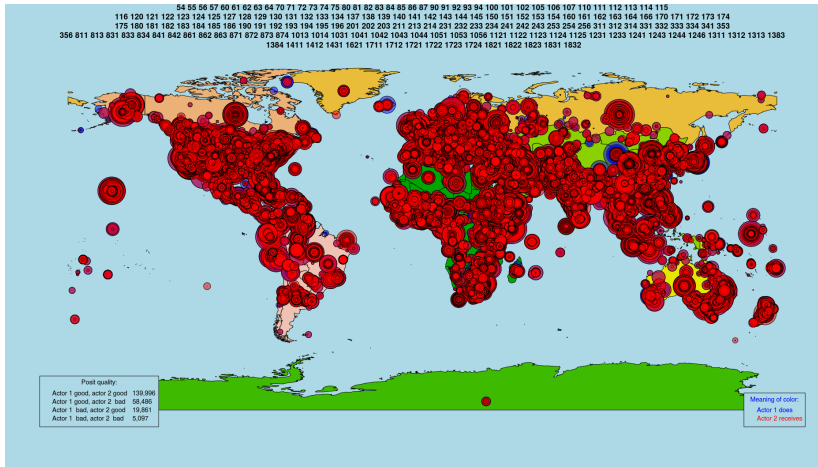
The script takes about 15 seconds to run, and processes 159GB of data.

About the GDELT data

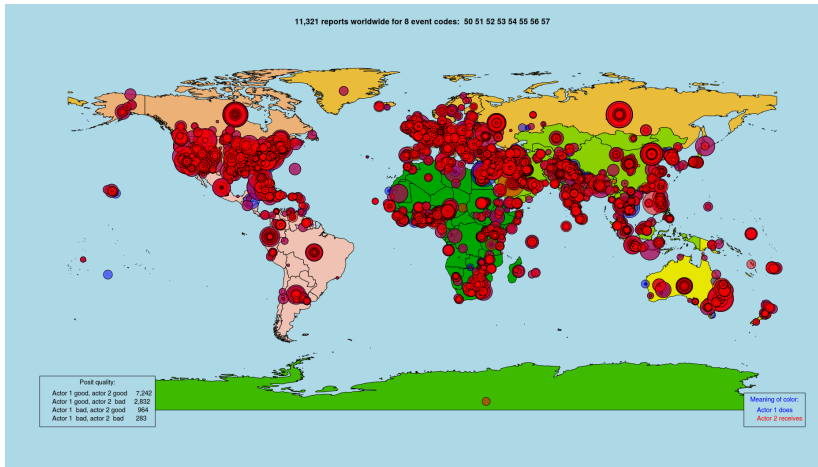
- 1 Actors 1 and 2. Actor 1 does something to Actor 2. Each actor is identified by lots of fields.
- 2 Every event has action attributes. Every event has lots of fields.
- 3 Geographic information about actor 1 and actor 2. Data beyond position.

Our explorations will focus on a single day 17 October 2017.

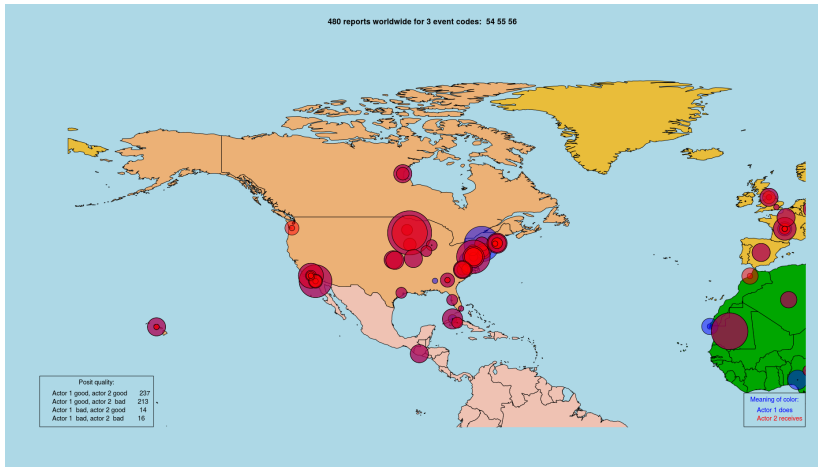
All the events in one day



Event codes 50 - 57



Event codes 54 - 56 for the North West quadrasphere



The six most frequently reported event codes.

Table: The six most frequently reported event codes.

Code	Explanation	Reports
10	Make public statement	17,551
42	Consult, make a visit	16,373
43	Consult, host a visit	15,039
40	Consult	13,606
30	Express intent to cooperate	12,691
51	Praise or endorse	12,500

Completeness of actor codes.

Table: Completeness of actor codes.

		Actor 1	
		Good	Bad
Actor2	Good	143,294	20,272
	Bad	59,874	0

Most common source domains.

Table: Most common source domains.

Source domain	Count
www.yahoo.com	1,502
allafrica.com	1,328
www.newsnw.in	923
www.dailymail.co.uk	900
www.4-traders.com	891
www.thehindu.com	547

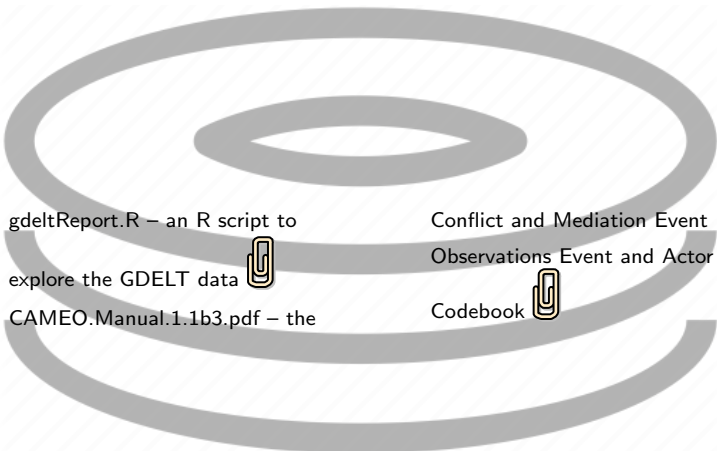

What have we covered?

- GDEL data is available on Google BigQuery site
- GDEL data is noisy
- GDEL data useful for long term trend analysis
- GDEL data can grow to Big Data size

Next time: see where the wind blows



Files of interest

- 
- 1 `gdeltReport.R` – an R script to explore the GDELT data  Conflict and Mediation Event Observations Event and Actor
 - 2 `CAMEO.Manual.1.1b3.pdf` – the Codebook 