Documentation: text_to_CAMEO.py

Contact:
Philip Schrodt, Parus Analytics (schrodt735@gmail.com)

Repository for code: `https://github.com/philip-schrodt/text_to_CAMEO`

Last update: 30 March 2015

This Python 2.6 program `text_to_CAMEO.py` takes the text-oriented format of the ICEWS *.tab* files released in the DataVerse Study 28075 and converts these to a more conventional data format using the CAMEO codes. The conversion process is described below.

To run: python text_to_CAMEO.py
Requires:

- CAMEO_codefile.txt
- countrynames.txt
- agentnames.txt
- filenames.txt

The program processes the files listed one per line in *filelist.txt* and produces tab-delimited output files with the name *reduced.ICEWS.events.[year].txt*. The data fields are in the following order

Date in YYYY-MM-DD format
Source country ISO-3166-alpha-3
Source country COW numeric
Source agent
Target country ISO-3166-alpha-3
Target country COW numeric
Target agent
Event CAMEO code
Event Goldstein score[1]
Event Quad score

Sample output

1996-01-01	CHN	710	GOV	CHN	710	PTY	013	0.4	1
1996-01-01	TWN	713	GOV	USA	002	OTH	042	1.9	1
1996-01-01	USA	002	OTH	TWN	713	GOV	043	2.8	1
1996-01-01	BGD	771	OPP	BGD	771	GOV	1121	-2	3
1996-01-01	ISR	666	CVL	PSE	000	GOV	112	-2	3

[1] These actually aren't Goldstein's scores for WEIS: they are a modification ca. 2002 by Uwe Reising for the CAMEO system, but they are generally known as Goldstein scores

1996-01-01 PSE 000 MED PSE 000 OPP 112 -2 3

Conversion notes

Country codes:

There are two fields, with ISO-3166-alpha-3 codes and COW numerical codes. See the file *countrynames.txt* for the conversion. '000' indicates there is no COW code. I think I got all of these: please let me know if you see errors. There were a very small number of obscure cases such as the Swedo-Finnish—or is it Finno-Swedish?—Åland Islands[2] which were converted to the code '---' and thus dropped.

CAMEO event codes:

The text fields turn out to all correspond to the existing CAMEO framework with a few spelling and phrasing modifications (these are noted in the file), so the codes from the CAMEO manual were used. The actual coding is based on the BBN dialect of CAMEO, which I've been calling CAMEO-B, rather than the original: this is extensively documented on the DataVerse files.

Goldstein scores:

These were copied from field 7 of the data

Quad score category:

1: Verbal cooperation, CAMEO 01-05
2: Material cooperation, CAMEO 06-09
3: Verbal conflict, CAMEO 10-14
4: Material conflict, CAMEO 15-20

Agents:

This was by far the most problematic conversion. The original data has a list of textual sectors which are almost but not quite comma delimited (as a small number of fields themselves contain commas, which is not cool). These generally correspond to CAMEO "agent" codes, and are sort of documented in the file *sectors.xml* as provided in early June 2014. This file, however, appears to be a work-in-progress, with a substantial number of places where information has not been completely filled in. I extracted almost all of the texts (there are a very small number of cases where this is not possible due to the aforementioned issue with commas: these go to the null code) into the file *agentnames.txt*[3], then I looked at every case that occurred with >0.01% frequency and made sure these generally made sense, then extracted the primary agent code as it would typically show up in a CAMEO-coded data set. These are—in order of priority—

---

[2] okay, they are not obscure if you live there. Which I don't, though I have visited. Putin is probably also considering annexing them based on longstanding historical Russian claims.

[3] The third tab-delimited field in that file attempts to extract the secondary fields from *sectors.xml*—these are typically numerical but quite a few refer to *tmp* files—and could be incorporated, but those fields are not used consistently and these probably need a lot of editing.

'GOV' : government
'MIL': military
'REB': militarized opposition, including the ICEWS-specific INS and SEP codes
'OPP': political opposition
'PTY': political party
'COP': police and security
'JUD': judiciary
'SPY': intelligence agencies
'IGO': IGOs and NGOs
'MED': Media (not medicine: that is 'HLH')
'EDU': Education
'BUS': Business [also add MNCs here?]
'CRM': criminals
'CVL': civilians

If you want more detail—and there definitely is more detail available in those sector lists at least for some of the cases—the code where this reduction is done is the procedure `reduce_sectors()` and it could easily be changed.