Documentation: text_to_CAMEO.py

Contact:
Philip Schrodt, Parus Analytics LLC (schrodt735@gmail.com)[1]

Repository for code: `https://github.com/openeventdata/text_to_CAMEO`

Last update: 8 February 2021
Version: 1.1B1

This Python 3.7 program `text_to_CAMEO.py` takes the text-oriented format of the ICEWS files released in slightly different formats in DataVerse Study 28075 or the government FOUO version 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 [-F] [-c] [-t <filename>] [-m]`

Requires:

CAMEO_codefile.txt [FOUO only]
countrynames.txt
agentnames.txt

The program processes the files and produces tab-delimited output files with the name *reduced.<file-name>.txt*.

Options:

`-F:` Files are in FOUO format. Default: Files are in Dataverse format

`-c:` Include COW numerical country codes in addition to ISO-3166 code. Default: Include only the ISO codes

`-t: <file-name>`: Process the files listed one per line in the text file <file-name>. Default: process all of the files in the working directory that end in ".csv" (Dataverse format) or ".tab" (FOUO format)

`-m:` Output all of the substate agents in a concatenated string. Default: only a single agent is used, with the priority determined by the list `agentcodes`.

[1] Thanks to Rob Boswell for identifying some quirks and issues in the April-2017 version. Happy for additional suggestions from the community.

Output (tab-delimited):
Date in YYYY-MM-DD format
Source country ISO-3166-alpha-3
[optional] Source country COW numeric
Source agent
Target country ISO-3166-alpha-3
[optional] Target country COW numeric
Target agent
Event CAMEO code
Event Goldstein score[2]
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
1996-01-01    PSE    000    MED    PSE    000    OPP    112     -2     3
```

Conversion notes

New formats

In 2020, ICEWS transitioned from Lockheed-Martin, where it has lost funding and was unavailable for several month, to the Political Instability Task Force (PITF), where funding should now be stable.[3] PITF has embarked in a number of modifications, including a number of long-overdue changes to the search strategy which should gradually reduce the Asian—and specifically South Asian—bias of the original and the incorporation of a number of additional non-English-language sources with machine translation. This has also resulted in some small (but uncorrected, fatal) changes to the format, specifically the addition of a CAMEO code starting in June-2020 (file: `events.20200601073501.Release468.csv`), which shifts the target fields. The program detects this change automatically but may break if there are additional changes. So, if your output doesn't look like the above, check whether that has occurred.

Country codes:

There are one or two fields, with ISO-3166-alpha-3 codes and [optionally] 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

[2]. 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, Though as noted below, some of the values in the data don't correspond to the values provided in the Dataverse documentation.

[3] These newer files have not made it into Dataverse as of the time of this update but I believe the intention is that they eventually will.

obscure cases such as the Swedo-Finnish—or is it Finno-Swedish?—Åland Islands[4] which were converted to the code '---'.

CAMEO event codes:

The text fields 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: these codes is extensively documented on the DataVerse files. The Dataverse format provides the actual code; in the FOUO version this has not been provided so the translation table is used.

“Goldstein” scores:

These were copied from field 7 of the data. However, note that the file *CAMEO.SCALE.txt* (which is simply a copy, with attribution, of a file I created in early 2007 for the original CAMEO, not CAMEO-B) provided on Dataverse differs in some cases from the scaled values in the data: for example code 016 has both changed meanings (from “Make empathetic comment” to “Deny responsibility”) and dramatically changed values (from +3.4 to -5).[5] It is obviously relatively straightforward to go through the data to figure out the codes and associated values, but don't depend on the *CAMEO.SCALE.txt* file.

Quad score category:

The general translation is the framework used in most academic research with CAMEO:

1: Verbal cooperation, CAMEO 01-05

2: Material cooperation, CAMEO 06-08

3: Verbal conflict, CAMEO 09-14

4: Material conflict, CAMEO 15-20

However, in CAMEO-B some of the 1 and 2 categories contain negative Goldstein scores—otherwise associated with conflict—so I've shifted the various '1' cases that have negative scores to '3'.

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

[4]. Okay, these places 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.

[5]This new value is arguably consistent with the magnitude of scores for '10x' (“Demand”) and '12x' (“Reject”), but is dramatically larger than most of the scaled values in the '01x' category.

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*[6], 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—

'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

By default, the single code—note that in some cases this is a six-character combined code such as GOVMIL or GOVPTY—with the highest priority in this list used. If only six-character codes are available, the first one found is returned.[7]

If the command line option `-m` is used, all of the codes are concatenated with “-”, e.g.

```
GOVPTY-GOV-LEG-CVL-PTY-MIL
```

and, presumably, this is used in subsequent customized processing, for example using a substring function (e.g. `substr()` in R; `in` or `find()` in Python) to detect whether MIL is present.

[6] 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.

[7] This is a lazy choice...it can be modified in the function `reduce_sectors()`.

Comments on initial April-2017 GitHub version

1. This program merges two earlier versions I'd used separately for the two formats. Both of those were used successfully in large-scale projects so I'm confident they were working, but I've not done that sort of testing on this merged version (I have done *basic* testing on it...really...). So definitely check and make sure the output makes sense.

2. The fact that there are at least two incompatible formats of the ICEWS data suggests the possibility that there might be others: again, check to make sure your output makes sense.

3. There's an assortment of “commented out” code in the program that was used earlier to do some basic marginals on the code: this could be reactivate but is in Python 2.6 and will also need a bit of updating to Python 3.5.

Comments on the March-2020 modifications

Dataverse files for 2015 through 2019 have options for three different file formats: the one compatible with this program is `Original File Format (Tab-Delimited)` which downloads as a file with a`.tsv` suffix. The new files under the `Tab-Delimited` option (`.tab` file suffix) have the string data fields in quotes, which is cool but that format change causes the actor names not to match, so these all produce "---". I've modified the program so that the default (no command line options) codes all files with either `.tab` or `.tsv` as the suffix.

ICEWS has been off-line for about nine months and will be produced by Leidos rather than Lockheed when/if it returns, so I'm waiting to see if there are additional changes under the new sponsorship; in the meantime this fix should work as long as you download the `.tsv` file and not the `.tab` file for those years.

Revision history

March-2020

- See comments above

February 2021

- -m option for output of all substate actors
- -F now detects the format change in FOUO version ca. May-2020. If the CAMEO code is already in the data, this is used.
- 1951 and 1952 codes added to CAMEO_codefile.txt
- `os.path.basename()` used for finding filename, generalizing beyond Unix