pprobber manual

Michal Kvasnička, Rostislav Staněk, and Ondřej Krčál

January 16, 2017

This document describes how to use **pprobber**, a software that downloads data from *Věstník veřejných zakázek* (the Czech official public procurement registry) and converts them do a CSV file.

pprobber consists of two small command-line utilities written in Python 2.7.

- `pprob.py` downloads data from *Věstník veřejných zakázek* and stores them in intermediary format
- `ppsteal.py` converts the intermediary format to CSV file

The whole software has been tested only on Linux so far.

`pprob.py`

There are two ways to run `pprob.py`. The first way is to download data in some range:

```
./pprob.py -s <since> -t <to> -f <form> -o <outputkeydict> \
           -m <mincedmeat> -e <errorfile>
```

When used this way, `pprob.py` downloads all forms of type (i.e. code name) `<form>` from the date `<since>` (in a format `dd.mm.yyyy`) to the date `<to>` (in the same format). `pprob.py` then produces three output files: it writes the intermediary data to `<mincedmeat>` file, error logs to `<errorfile>` file, and a list of tags used in the forms to `<outputkeydict>` file.

Example: to download all forms 3 in the whole year 2015:

```
./pprob.py -s 1.1.2015 -t 31.12.2015 -f 3 -o dict2015.txt \
           -m data2015.txt -e errfile2015.txt
```

Sometimes some forms are not downloaded. In such a case, the list of failed downloads is written into `<errorfile>`. These forms can be downloaded later this way:

```
./pprob.py --process_failed_forms -r <former errorfile> -o <new outputkeydict> \
           -m <new mincedmeat> -e <new errorfile>
```

Example:

```
./pprob.py --process_failed_forms -r errfile2015.txt -o adddict2015.txt \
           -m adddata2015.txt -e adderrfile2015.txt
```

`ppsteal.py`

When all data are downloaded, they can be converted into a CSV file with `ppsteal.py`. It is run with three arguments:

```
./ppsteal.py -i <mincedmeat> -o <csvfile> -d <definition_file>
```

where `<mincedmeat>` is the intermediary data file created by `pprobber.py`, `<csvfile>` is the name of the output CSV file, and `<definition_file>` is the name of a file defining what parts of the intermediary data should be placed into the output CSV file. This way, you can produce various CSV files from one intermediary data file.

The definition file has a following format:

- each line describes one element in the intermediary data and one column in the output CSV file
- the definition of each element has tree parts separated by a pipe “`|`”
  1. the name of the element in the original web page
  2. type of the the element in the original web page
  3. the name of the variable (i.e. column) in the output CSV
- there are some special quasi-elements; there are not part of the web page, rather, they are created by the `pprob.py`; these quasi-elements’ names begin with `ROBBER_`
- public procurement projects can have many parts; if it is the case, then some elements can be present in the web page many times; these elements’ names should include “`[*]`” in the definition file; then `ppsteal.py` would produce one record per part including the common parts of the form

Example of the definition file is `PAdef.txt`:

```
ROBBER_HEADER_ID_Formulare | robberheader | ID_formulare
ROBBER_VARIANTS_NUMBER | any | pocet_variant
ROBBER_VARIANT[*] | any | varianta_cislo
ROBBER_HEADER_Evidencni_cislo_zakazky | robberheader | ID_zakazky
ROBBER_HEADER_ICO_dodavatele | robberheader | ico_dodavatele
FormItems.IdentifikacniCislo_1 | text | ico_zadavatele
FormItems.DruhVerejnehoZadavatele_I_2 | radio | druh_zadavatele
FormItems.NUTS1_II_1_2 | text | NUTS
FormItems.HlavniSlovnikHp_II_1_5 | text | CPV
FormItems.Hodnota_II_2_1 | text | cena
FormItems.Mena_II_2_1 | select | mena
FormItems.DruhRizeni_IV_1_1 | radio | druh_rizeni
AttModels[*].AttItems.Datum_V_1 | text | datum
AttModels[*].AttItems.PocetNabidek_V_2 | text | pocet_nabidek
AttModels[*].AttItems.NazevDodavatele_V_3 | text | dodavatel
AttModels[*].AttItems.Hodnota1_V_4 | text | dilci_odhad
AttModels[*].AttItems.Hodnota2_V_4 | text | dilci_cena
AttModels[*].AttItems.PocetNabidek_V_2 | text | dilci_pocet_nabidek
AttModels[*].AttItems.RocniHodnota_V_4 | text |   roky
AttModels[*].AttItems.MesicniHodnota_V_4 | text | mesice
AttModels[*].AttItems.ZakazkaCast_V | text | cast_zakazky
```

License and citation

You can use **pprobber** software free of charge if you are one of

- *Úřad pro ochranu hospodářské soutěže*
- a university or a research institute

In other cases, please contact the authors.

If you use the data downloaded by this software in any publication, please cite this manual.

Contact

If you want to use this software, please contact Michal Kvasnička via e-mail `michal.kvasnicka` at `econ.muni.cz`.

Acknowledgement

pprobber was created for *Úřad pro ochranu hospodářské soutěže* (the Czech antitrust authority). Its develpment was supported by TA ČR grant TD03000121 “Tvorba systému automatické detekce jednání podezřelého z koluze ve veřejných zakázkách”.