Notebook

September 15, 2014

1 Registered User Contributions

We are interested in obtaining country information from registered users. Unfortunately, it is not possible to get country information for registered users directly. However there users can use location categories to provide information regarding their location.

The "Wikipedians by location" category provides a listing of all existing location categories. From these, we can select the categories for countries we are interested in.

NOTE: functionality for resolving user countries has been wrapped up in *wikidat.utils.userresolver* package

```
In [1]: import btb.utils.userresolver as usolve
        from __future__ import division
```

1.1 Contributing country categories

We are interested in identifying country contributions from users who are located in countries which make significant number of contributions. Wikimedia provides statistics on the countries which contribute more to wikipedia. We will try to identify users from countries in the following list:

Country	Contribution (%)
United States	38.30%
United Kingdom	13.20%
India	6.90%
Canada	5.40%
Australia	3.60%
Philippines	2.60%
Germany	1.50%
Brazil	1.10%
Italy	1.00%
Ireland	1.00%
Pakistan	0.90%
France	0.80%
Malaysia	0.80%
Netherlands	0.80%
Indonesia	0.80%
China	0.70%

Country	Contribution (%)
New Zealand	0.70%
Spain	0.70%
Iran	0.70%
Mexico	0.50%
Sweden	0.50%
Russia	0.50%
Greece	0.50%
Turkey	0.50%

Users in these countries will be indicated in the following categories:

```
In [3]: countrySeeds = usolve.__getCountrySeeds__()

        for country in countrySeeds:
            seeds = countrySeeds[country]
            print country,'contains seeds: '
            for seed in seeds:
                print '\t',seed
```

```
FR contains seeds:
        Wikipedians in France
DE contains seeds:
        Wikipedians in Germany
BR contains seeds:
        Wikipedians in Brazil
GR contains seeds:
        Wikipedians in Greece
RU contains seeds:
        Wikipedians in Russia
NL contains seeds:
        Wikipedians in Netherlands
        Wikipedians in the Netherlands
TR contains seeds:
        Wikipedians in Turkey
NZ contains seeds:
        Wikipedians in New Zealand
PK contains seeds:
        Wikipedians in Pakistan
        Wikipedians in the Pakistan
PH contains seeds:
        Wikipedians in the Philippines
        Wikipedians in Philippines
CN contains seeds:
        Wikipedians in China
        Wikipedians in Mainland China
        Wikipedians in the People's Republic of China
        Wikipedians in the People's Republic of China/
        Wikipedians in the Republic of China
CA contains seeds:
        Wikipedians in Canada
IR contains seeds:
        Wikipedians in the Iran
        Wikipedians in Iran
IT contains seeds:
        Wikipedians in Italy
AU contains seeds:
        Wikipedians in the Australia
        Wikipedians in AUSTRALIA
        Wikipedians in Australia
IN contains seeds:
        Wikipedians in India
        Wikipedians in the India
        Wikipedians in the Republic of INDIA
IE contains seeds:
        Wikipedians in Ireland
        Wikipedians in the Republic of Ireland
ID contains seeds:
        Wikipedians in Indonesia
ES contains seeds:
        Wikipedians in Spain
        Wikipedians in in Spain
US contains seeds:
        Wikipedians in U.S.A.
        Wikipedians in United States
        Wikipedians in United States Of America
        Wikipedians in United States of America
        Wikipedians in US
        Wikipedians in USA
        Wikipedians in the United States
        Wikipedians in the United States of America
UK contains seeds:
        Wikipedians in UK
        Wikipedians in United Kingdom
        Wikipedians in the United Kingdom
        Wikipedians in the UK
MY contains seeds:
        Wikipedians in Malaysia
MX contains seeds:
        Wikipedians in Mexico
SE contains seeds:
        Wikipedians in Sweden
```

```
In [4]: import mwclient
        wiki = mwclient.Site('en.wikipedia.org')
```

For any given category, we can get the users registered in this category, as well as any sub-category listed in this category:

```
In [5]: seedCountry = 'NL'
        seedCats = countrySeeds[seedCountry]
        cat = wiki.Categories[seedCats[1]]

        users, cats = usolve.__getAllUsers__(cat)
        print 'Using seed cat: ', cat.name
        print '  Category contains: {:,} users'.format(len(users))
        print '  Category contains: {:,} categories'.format(len(cats))
        for cat_i in cats:
            print '    > ',cat_i.name
```

```
Using seed cat:  Category:Wikipedians in the Netherlands
  Category contains: 332 users
  Category contains: 5 categories
    >  Category:Wikipedians in North Holland
    >  Category:Wikipedians in Aruba
    >  Category:Wikipedians in Leiden
    >  Category:Wikipedians in Amsterdam
    >  Category:Wikipedians in Tilburg
```

We can use this process recursively to get all the users from one category and all its sub-categories. In this way, we can generate a list of users in a country.

```
In [6]: seedCountry = 'NL'
        seedCats = countrySeeds[seedCountry]
        print 'Using seed categories: '
        for seed in seedCats:
            print '  > ',seed
        print ''

        # wmclient.Category object required as seeds
        wikiCats = [ wiki.Categories[seed] for seed in seedCats ]
        # we keep a record of
        log = []
        visited = set()
        allUsers = usolve.__fetchUsersCategory__(wikiCats, log, visited)

        print ''
        print 'Used seed categories yielded {:,} users in {:}'.format(len(allUsers),seedCountry)
```

```
Using seed categories:
  >  Wikipedians in Netherlands
  >  Wikipedians in the Netherlands

Fetching  Category:Wikipedians in the Netherlands ...
Fetching  Category:Wikipedians in Tilburg ...
Fetching  Category:Wikipedians in Amsterdam ...
Fetching  Category:Wikipedians in Leiden ...
Fetching  Category:Wikipedians in Aruba ...
Fetching  Category:Wikipedians in North Holland ...
Fetching  Category:Wikipedians in Netherlands ...

Used seed categories yielded 348 users in NL
```

In this way the country seed categories can be used to generate a full list of users for each country. The ***fetchData()*** method performs this task. However, because this process can be slow (a few minutes), a list of user names is kept in a cache file (a python pickle file).

The ***fetchData()*** method produces a list of user names for each country. Additionally, this method keeps record of the subcategories listed for each category. This record can be used to visualize the tree structure of categories visited for each country. Visualization of these trees is explained in the Visualize Category Trees notebook.

The *wikidat.utils.userresolver* module keeps track of the user names associated with each country. At the same time this module provides functionality to determine the country of origin of a given user.

The following is a user count of the known users for each country:

```
In [8]: usernameSets,log = usolve.__getUserMap__(lang='en')

        tot = 0
        for country in sorted(usernameSets.keys()):
            print '{:<10}\t{:>6,} users'.format(country,len(usernameSets[country]))
            tot += len(usernameSets[country])
        print '============================'
        print '{:<10}\t{:>6,} users'.format('TOTAL',tot)
```

```
AU                 2,300 users
BR                   461 users
CA                 2,675 users
CN                   349 users
DE                   733 users
ES                   257 users
FR                   194 users
GR                   228 users
ID                   322 users
IE                   622 users
IN                 1,463 users
IR                    32 users
IT                   168 users
MX                   255 users
MY                   219 users
NL                   340 users
NZ                   653 users
PH                   111 users
PK                   164 users
RU                   464 users
SE                   364 users
TR                   226 users
UK                 4,094 users
US                13,585 users
============================
TOTAL             30,279 users
```

The following exampl show how the *wikidat.utils.userresolver* module can be used to find out the country of a given user.

```
In [9]: aUser = 'Alaney2k'  # This is a known canadian user
        usolve.getUserCountry('Alaney2k')

Out[9]: 'CA'
```