OpenML in Python

OpenML is an online collaboration platform for machine learning:

- Share/reuse machine learning datasets, algorithms, models, experiments
- Well documented/annotated datasets, uniform access
- APIs in Java, R, Python*,... to download/upload everything
- Better reproducibility of experiments, reuse of machine learning models
- Works well with machine learning libraries such as scikit-learn
- Large scale benchmarking, compare to state of the art

```
<IPython.core.display.HTML object>
```

Authentication

- Create an OpenML account (free) on http://www.openml.org.
- After logging in, open your account page (avatar on the top right)
- Open 'Account Settings', then 'API authentication' to find your API key.

There are two ways to authenticate:

- Create a plain text file `~/.openml/config` with the line 'apikey=MYKEY', replacing MYKEY with your API key.
- Run the code below, replacing 'MYKEY' with your API key.

[2]:
```
# Uncomment and run this to authenticate. Don't share your API key!
# oml.config.apikey = os.environ.get('OPENMLKEY','MYKEY')
```

Data sets

We can list, select, and download all OpenML datasets

List datasets

[3]:
```
datalist = oml.datasets.list_datasets() # Returns a dict
datalist = pd.DataFrame.from_dict(datalist, orient='index') # Create a DataFrame
print("First 10 of %s datasets..." % len(datalist))
datalist[:10][['did','name','NumberOfInstances',
               'NumberOfFeatures','NumberOfClasses']]
```

```
First 10 of 19492 datasets...
```

	did	name	NumberOfInstances	NumberOfFeatures	NumberOfClasses
1	1	anneal	898	39	6
2	2	anneal	898	39	6
3	3	kr-vs-kp	3196	37	2
4	4	labor	57	17	2
5	5	arrhythmia	452	280	16
6	6	letter	20000	17	26
7	7	audiology	226	70	24
8	8	liver-disorders	345	7	-1

```
9      9            autos                205                26                 7
10    10            lymph                148                19                 4
```

There are many properties that we can query

```
[4]: list(datalist)
     datalist = datalist[['did','name','NumberOfInstances',
                          'NumberOfFeatures','NumberOfClasses']]
```

```
['MinorityClassSize',
 'NumberOfFeatures',
 'MajorityClassSize',
 'NumberOfSymbolicFeatures',
 'NumberOfClasses',
 'NumberOfNumericFeatures',
 'status',
 'name',
 'NumberOfInstances',
 'NumberOfInstancesWithMissingValues',
 'NumberOfMissingValues',
 'did',
 'format',
 'MaxNominalAttDistinctValues']
```

and we can filter or sort on all of them

```
[5]: datalist[datalist.NumberOfInstances>10000
              ].sort(['NumberOfInstances'])[:20]
```

```
         did                                    name  NumberOfInstances  \
23515  23515                                 sulfur              10081
372      372                       internet_usage              10108
981      981                    kdd_internet_usage              10108
1536    1536                         volcanoes-b6              10130
4562    4562                        InternetUsage              10168
1531    1531                         volcanoes-b1              10176
1534    1534                         volcanoes-b4              10190
1459    1459                artificial-characters              10218
1478    1478                                  har              10299
1533    1533                         volcanoes-b3              10386
1532    1532                         volcanoes-b2              10668
1053    1053                                  jm1              10885
1414    1414  Kaggle_bike_sharing_demand_challange              10886
1044    1044                        eye_movements              10936
1019    1019                            pendigits              10992
32        32                            pendigits              10992
4534    4534                    PhishingWebsites              11055
399      399                           ohscal.wc              11162
310      310                         mammography              11183
1568    1568                             nursery              12958
```

```
       NumberOfFeatures  NumberOfClasses
23515                 7               -1
372                  72               46
981                  69                2
1536                  4                5
4562                 72               -1
1531                  4                5
1534                  4                5
1459                  8               10
1478                562                6
1533                  4                5
1532                  4                5
1053                 22                2
1414                 12               -1
1044                 28                3
1019                 17                2
32                   17               10
4534                 31                2
399               11466               10
310                   7                2
1568                  9                4
```

or find specific ones

```
[6]: datalist.query('name == "eeg-eye-state"')
```

```
       did           name  NumberOfInstances  NumberOfFeatures  \
1471  1471  eeg-eye-state              14980                15

      NumberOfClasses
1471                2
```

```
[7]: datalist.query('NumberOfClasses > 50')
```

```
       did                        name  NumberOfInstances  NumberOfFeatures  \
1491  1491   one-hundred-plants-margin               1600                65
1492  1492    one-hundred-plants-shape               1600                65
1493  1493  one-hundred-plants-texture               1599                65
4546  4546                      Plants              44940                16
4552  4552          BachChoralHarmony               5665                17

      NumberOfClasses
1491              100
1492              100
1493              100
4546               57
4552              102
```

Download a specific dataset. This is done based on the dataset ID (called 'did').