4.2 Download raw data (optional)

(中文版)

If a dataset is already in local disk, make sure it's in directory `raw_dir`. If one wants to run the code anywhere without bothering to download and move data to the right directory, one can do it automatically by implementing function `download()`.

If the dataset is a zip file, make `MyDataset` inherit from `dgl.data.DGLBuiltinDataset` class, which handles the zip file extraction for us. Otherwise, one needs to implement `download()` like in `QM7bDataset`:

```
import os
from dgl.data.utils import download

def download(self):
    # path to store the file
    file_path = os.path.join(self.raw_dir, self.name + '.mat')
    # download file
    download(self.url, path=file_path)
```

The above code downloads a .mat file to directory `self.raw_dir`. If the file is a .gz, .tar, .tar.gz or .tgz file, use `extract_archive()` function to extract. The following code shows how to download a .gz file in `BitcoinOTCDataset`:

```
from dgl.data.utils import download, check_sha1

def download(self):
    # path to store the file
    # make sure to use the same suffix as the original file name's
    gz_file_path = os.path.join(self.raw_dir, self.name + '.csv.gz')
    # download file
    download(self.url, path=gz_file_path)
    # check SHA-1
    if not check_sha1(gz_file_path, self._sha1_str):
        raise UserWarning('File {} is downloaded but the content hash does not match.'
                          'The repo may be outdated or download may be incomplete. '
                          'Otherwise you can create an issue for it.'.format(self.name +
'.csv.gz'))
    # extract file to directory `self.name` under `self.raw_dir`
    self._extract_gz(gz_file_path, self.raw_path)
```

The above code will extract the file into directory `self.name` under `self.raw_dir`. If the class inherits from `dgl.data.DGLBuiltinDataset` to handle zip file, it will extract the file into directory `self.name` as well.

Optionally, one can check SHA-1 string of the downloaded file as the example above does, in case the author changed the file in the remote server some day.