SEC Financial Statement Data I

Our standard libraries:

In [1]:

```
import pandas as pd
import requests
```

Main page: https://www.sec.gov/dera/data/financial-statement-and-notes-data-set.html

Zip files

Link to most recent file:

In [2]:

```
url = 'https://www.sec.gov/files/dera/data/financial-statement-and-notes-data-se
```

We need these libraries to open the zip file:

In [3]:

```
import zipfile
import io
```

In [4]:

```
r = requests.get(url)
r
```

Out[4]:

```
<Response [200]>
```

Did download work?

In [5]:

```
if r.ok:
    print('good')
```

```
good
```

Unzip file:

In [7]:

```
z = zipfile.ZipFile(io.BytesIO(r.content))
```

Get names of all the files in the zip folder:

In [8]:

```
z.namelist()
```

Out[8]:

```
['sub.tsv',
 'tag.tsv',
 'dim.tsv',
 'ren.tsv',
 'cal.tsv',
 'pre.tsv',
 'num.tsv',
 'txt.tsv',
 'readme.htm',
 'notes-metadata.json']
```

Now we can open any of these files.

For example, open the numbers file:

```
In [9]:
num = z.open( 'num.tsv' )
num
```

```
Out[9]: <zipfile.ZipExtFile name='num.tsv' mode='r' compress_type=deflate>
```

Read this file:

```
In [10]:
pd.read_table(num)
```

```
/Users/janschneider/opt/anaconda3/lib/python3.7/site-packages/IPython/core/inter
activeshell.py:3147: DtypeWarning: Columns (12) have mixed types.Specify dtype o
ption on import or set low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

Out[10]:

	adsh	tag	version	ddate	qtrs	
0	0001640334-20-002990	AccountsPayableAndAccruedLiabilitiesCurrent	us-gaap/2019	20180131	0	
1	0001640334-20-002990	AccountsPayableAndAccruedLiabilitiesCurrent	us-gaap/2019	20170131	0	
2	0001640334-20-002995	AccountsPayableAndAccruedLiabilitiesCurrent	us-gaap/2019	20190831	0	
3	0001640334-20-002995	AccountsPayableAndAccruedLiabilitiesCurrent	us-gaap/2019	20200831	0	
4	0001640334-20-003002	AccountsPayableAndAccruedLiabilitiesCurrent	us-gaap/2019	20191231	0	
...	...	...	...	...	...	
567315	0001477932-20-007602	WarrantsExercisedValue	0001477932-20-007602	20190630	0	
567316	0001477932-20-007602	WarrantsExercisedValue	0001477932-20-007602	20200331	4	
567317	0001477932-20-007602	WarrantsGranted	0001477932-20-007602	20190630	2	s
567318	0001477932-20-007602	WarrantsGranted	0001477932-20-007602	20190331	4	s
567319	0001477932-20-007602	WarrantsIssuedShares	0001477932-20-007602	20190331	4	s

567320 rows × 16 columns

Which companies are these? → Open the submissions file:

```
In [11]:
sub = z.open( 'sub.tsv' )

pd.read_table(sub)
```

Out[11]:

adsh	cik	name	sic	countryba	stprba	cityba	zipba

	adsh	cik	name	sic	countryba	stprba	cityba	zipba
0	0000021510-20-000049	21510	COHERENT INC	3826.0	US	CA	SANTA CLARA	95054
1	0000034088-20-000095	34088	EXXON MOBIL CORP	2911.0	US	TX	IRVING	75039-2298
2	0000045012-20-000111	45012	HALLIBURTON CO	1389.0	US	TX	HOUSTON	77032
3	0000049071-20-000153	49071	HUMANA INC	6324.0	US	KY	LOUISVILLE	40202
4	0000066756-20-000085	66756	ALLETE INC	4931.0	US	MN	DULUTH	55802-2093
...	...	...	...	...	...	...	...	...
3349	0001640334-20-003182	1623360	MIRAGE ENERGY CORP	3089.0	US	TX	SAN ANTONIO	78216
3350	0001654954-20-014038	725363	CEL SCI CORP	2836.0	US	VA	VIENNA	22182
3351	0001683168-20-004497	725929	B2DIGITAL, INC.	7997.0	US	FL	TAMPA	33624
3352	0001721868-20-000664	1087022	ALR TECHNOLOGIES INC.	3669.0	US	VA	RICHMOND	23225
3353	0001721868-20-000670	1530746	KAYA HOLDINGS, INC.	2834.0	US	FL	FORT LAUDERDALE	33304

3354 rows × 40 columns

Now let's save these files. You need to generate the following directory (folder) structure:

- current directory (where you run this notebook): folder "data"
- inside data folder: folder "sec"
- inside sec folder: folder "downloads

This is where we save all file for 2020-12:

In [12]:

```
period = '2020_12'
```

```
unzip_folder_name = 'data/sec/downloads/'  + period
unzip_folder_name
```

Out[12]:

```
'data/sec/downloads/2020_12'
```

Generate the directory '2020_12' inside 'data/sec/downloads/':

In [13]:

```
import os


if not os.path.exists(unzip_folder_name):
    os.mkdir(unzip_folder_name)                 # Create directory for unzipped f
```

Save to this directory:

In [14]:

```
z.extractall(unzip_folder_name)              # Unzip file into new directory
```

Structure of the URLs:

In [15]:

```
'https://www.sec.gov/files/dera/data/financial-statement-and-notes-data-sets/202
'https://www.sec.gov/files/dera/data/financial-statement-and-notes-data-sets/202
'https://www.sec.gov/files/dera/data/financial-statement-and-notes-data-sets/202
'https://www.sec.gov/files/dera/data/financial-statement-and-notes-data-sets/202
```

Out[15]:

```
'https://www.sec.gov/files/dera/data/financial-statement-and-notes-data-sets/202
0q2_notes.zip'
```

Construct specific URL:

In [16]:

```
period = '2020q1'

'https://www.sec.gov/files/dera/data/financial-statement-and-notes-data-sets/' +
```

Out[16]:

```
'https://www.sec.gov/files/dera/data/financial-statement-and-notes-data-sets/202
0q1_notes.zip'
```

Put all this into a function:

In [17]:

```
def download_file(period):
    url = 'https://www.sec.gov/files/dera/data/financial-statement-and-notes-dat

    unzip_folder_name = 'data/sec/downloads/' + period

    r = requests.get(url)
    if r.ok:
        print('Downloaded:', url, 'to:', unzip_folder_name)
        if not os.path.exists(unzip_folder_name): os.mkdir(unzip_folder_name)
        z = zipfile.ZipFile(io.BytesIO(r.content))
        z.extractall(unzip_folder_name)
```

Now use the function like this (check your directory to make sure that the files got saved correctly:

In [18]:

```
download_file('2020_11')
```

```
Downloaded: https://www.sec.gov/files/dera/data/financial-statement-and-notes-da
ta-sets/2020_11_notes.zip to: data/sec/downloads/2020_11
```

In [19]:

```
download_file('2009q2')
```

```
Downloaded: https://www.sec.gov/files/dera/data/financial-statement-and-notes-da
ta-sets/2009q2_notes.zip to: data/sec/downloads/2009q2
```

To download all files, we loop over all periods.

For example, all months in 2020:

In [59]:

```
for year in range(2020,2021):
    for month in range(1,13):
        period = str(year)+'_'+str(month)
        print(period)
```

```
2020_1
2020_2
2020_3
2020_4
2020_5
2020_6
2020_7
2020_8
2020_9
2020_10
2020_11
2020_12
```

Now run this cell to download all available files:

In [ ]:

```
for year in range(2010,2021):
    for quarter in [1,2,3,4]:
        period = str(year)+'q'+str(quarter)
        download_file(period)

for year in range(2020,2021):
    for month in range(1,13):
        period = str(year)+'_'+str(month)
        download_file(period)
```