I. Installing the Unified Medical Language System (UMLS)

In [1]:
```
%load_ext autoreload
%autoreload 2

import sys
sys.path.insert(0,'../../trove')
```

A. Download the UMLS Release Files

Trove requires access to the Unified Medical Language System (UMLS) which is freely available after signing up for an account with the National Library of Medicine (NLM).

Visit the link below and download the latest "UMLS Metathesaurus Files" release 2020AB. This file is quite large (5.3 GB compressed), so it may take some time to download. **Please note, "full" release zip files are currently not supported.**

Alternatively, if you have an existing API KEY you can use the following script to download the zip file from the command line. See https://documentation.uts.nlm.nih.gov/automating-downloads.html for technical details on NLM authentication.

```
python download_umls.py \
   --apikey XXXXXXXX-XXXX-XXXX-XXXX-XXXXXXXXXXXX \
   --url https://download.nlm.nih.gov/umls/kss/2020AB/umls-2020AB-
metathesaurus.zip
```

B. Installation

Currently there are 2 ways to initalize the UMLS:

- From the source zip file (e.g., `umls-2020AB-metathesaurus.zip`)
- From the Rich Release Format (RRF) files generated by MetamorphoSys
- **TBD** An existing database instance

Depending on your machine, this should take 2-5 minutes.

Option 1: Install from NLM Zip File

In [8]:
```
%%time
from trove.labelers.umls import UMLS

# setup defaults
UMLS.config(
    cache_root = "~/.trove/umls2020AB",
```

```
    backend = 'pandas'
)

NLM_ZIPFILE_PATH = "umls-2020AB-metathesaurus.zip"
if not UMLS.is_initalized():
    print("Initializing the UMLS from zip file...")
    UMLS.init_from_nlm_zip(NLM_ZIPFILE_PATH, use_checksum=True)
```

```
Initializing the UMLS from zip file...
```

```
---------------------------------------------------------------------------
AssertionError                            Traceback (most recent call last)
<timed exec> in <module>

~\Hubs\trove\trove\labelers\umls.py in init_from_nlm_zip(fpath, outdir, backend, use_
checksum, keep_original_rrfs)
    302
    303         path = Path(fpath)
--> 304         assert os.path.exists(path)
    305         # Checksum test
    306         if use_checksum:

AssertionError:
```

Option 2: Install from RRF Files

If you have installed the UMLS before using MetamorphoSys to create custom vocabulary subsets you can directly use the generated RRF files.

In [3]:
```
%%time

RRF_FILE_PATH = ""
if not UMLS.is_initalized():
    print("Initializing the UMLS from RRFs...")
    UMLS.init_from_rrfs(RRF_FILE_PATH)
```

```
Initializing the UMLS from RRFs...
```

```
---------------------------------------------------------------------------
FileNotFoundError                         Traceback (most recent call last)
<timed exec> in <module>

~\Hubs\trove\trove\labelers\umls.py in init_from_rrfs(indir, outdir, backend)
    362         for fname in ["MRCONSO.RRF", "MRSTY.RRF", "MRSAB.RRF"]:
    363             if not os.path.exists(f"{indir}/{fname}"):
--> 364                 raise FileNotFoundError(errno.ENOENT, os.strerror(errno.ENOEN
T), fname)
    365
    366         # Source terminologies - MRSAB.RRF

FileNotFoundError: [Errno 2] No such file or directory: 'MRCONSO.RRF'
```

Option 3: Install from an Existing Database Instance

TBD: If you have a live UMLS database instance running, you can initialize Trove as follows.

In [4]:
```
# if not UMLS.is_initalized():
#     UMLS.init_from_dbconn(engine='mysql', dbname='UMLS2020AB')
```

3. Test the Installation

Here we apply some common term transformations. This should run in 2-4 minutes.

In [5]:
```
%%time
from trove.labelers.umls import UMLS
from trove.transforms import SmartLowercase
from trove.contrib.datasets.stopwords import stopwords

# english stopwords
stopwords = stopwords.union(set([t[0].upper() + t[1:] for t in stopwords]))

# options for filtering terms
config = {
    "type_mapping"  : "TUI",  # TUI = semantic types, CUI = concept ids
    'min_char_len'  : 2,
    'max_tok_len'   : 8,
    'min_dict_size' : 500,
    'stopwords'     : stopwords,
    'transforms'    : [SmartLowercase()],
    'languages'     : {"ENG"},
    'filter_sabs'   : {"SNOMEDCT_VET"},
    'filter_rgx'    : r'''^[-+]*[0-9]+([.][0-9]+)*$'''  # filter numbers
}

umls = UMLS(**config)
```

```
---------------------------------------------------------------------------
ModuleNotFoundError                       Traceback (most recent call last)
<timed exec> in <module>

ModuleNotFoundError: No module named 'trove.contrib.datasets.stopwords'
```

Look at semantic type assignments for an example term `acetaminophen` from the Medical Subject Headings (MeSH®) terminology.

In [6]:
```
from trove.labelers.umls import SemanticGroups

semgroups = SemanticGroups()
stys = umls.terminologies['MSH']['acetaminophen']
print(stys)
print([semgroups.types[sty] for sty in stys])
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
~\AppData\Local\Temp\ipykernel_30464\3226213350.py in <module>
      2
      3 semgroups = SemanticGroups()
----> 4 stys = umls.terminologies['MSH']['acetaminophen']
      5 print(stys)
      6 print([semgroups.types[sty] for sty in stys])

NameError: name 'umls' is not defined
```

In [ ]:
```

```