```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
Retrieve and store local representation of MMWR online with minimal processing,
which can include conversion to UTF-8 and basic parsing of HTML.

@author: chadheilig

Begin with journal-specific dframe, containing a complete list of files.
MMWR, PCD, EID, PHR

_dframe pandas DataFrame
   base     URL from which href values were harvested
   href     hypertext reference (bs.a['href']); + base URL -> absolute URL
   url      absolute URL constructed from href and base URL
   path     path from absolute URL
   filename name of HTML file in hypertext reference
   mirror_path path on local mirror
   string   string from anchor element content
   level    in concatentated DataFrame, volume or article

Main products: local mirror of online archive as raw copy (bytes) and lightly
formatted UTF-8 (string), as well as pickle representations for ease of
continuity. See also 2_retrieve-and-store-experiments.py for timing trials.
"""

#%% Import modules and set up environment
# import from 0_cdc-corpora-header.py

os.chdir('/Users/cmheilig/cdc-corpora/_test')

#%% Retrieve journal-specific DataFrames from pickle files
mmwr_dframe = pickle.load(open('pickle-files/mmwr_dframe.pkl', 'rb'))
# (14751, 8)
mmwr_pdf_dframe = pickle.load(open("pickle-files/mmwr_pdf_dframe.pkl", "rb"))
# (2066, 9)

pcd_dframe = pickle.load(open('pickle-files/pcd_dframe.pkl', 'rb')) # (4335, 8)
pcd_dframe.drop(pcd_dframe.index[pcd_dframe.level == 'en_es'], inplace=True)
# (3777, 8)

eid_dframe = pickle.load(open('pickle-files/eid_dframe.pkl', 'rb'))
# (11504, 8)
# phr_dframe = pickle.load(open('pickle-files/phr_dframe.pkl', 'rb'))
# (2737, 8)

#%% Set up local mirror directories for unprocessed HTML (b0)
MMWR_BASE_PATH_b0 = normpath(expanduser('~/cdc-corpora/mmwr_b0/'))
MMWR_BASE_PATH_pdf = normpath(expanduser('~/cdc-corpora/mmwr_pdf/'))
PCD_BASE_PATH_b0 = normpath(expanduser('~/cdc-corpora/pcd_b0/'))
EID_BASE_PATH_b0 = normpath(expanduser('~/cdc-corpora/eid_b0/'))
# PHR_BASE_PATH_b0 = normpath(expanduser('~/cdc-corpora/phr_b0/'))

x = create_mirror_tree(MMWR_BASE_PATH_b0, calculate_mirror_dirs(mmwr_dframe.path))
# { key: (0 if val is None else len(val)) for (key, val) in x.items() }

```

```
x = create_mirror_tree(PCD_BASE_PATH_b0, calculate_mirror_dirs(pcd_dframe.path))
# { key: (0 if val is None else len(val)) for (key, val) in x.items() }

x = create_mirror_tree(EID_BASE_PATH_b0, calculate_mirror_dirs(eid_dframe.path))
# { key: (0 if val is None else len(val)) for (key, val) in x.items() }

# x = create_mirror_tree(PHR_BASE_PATH_b0, calculate_mirror_dirs(phr_dframe.path))
# { key: (0 if val is None else len(val)) for (key, val) in x.items() }

#%% Mirror unprocessed HTML from internet to local archive (www -> b0)

# mirror_raw_html(mmwr_dframe.url[200], MMWR_BASE_PATH_b0 + mmwr_dframe.mirror_path[200])

mmwr_sizes_b0 = [mirror_raw_html(url, MMWR_BASE_PATH_b0 + path, print_url = False)
                     for url, path in tqdm(zip(mmwr_dframe.url, mmwr_dframe.mirror_path),
                                           total=14226)]

# harvest only HTML for main page and years 2021-2022 (vol 70-71)
# level in ['home', 'series'] or
#     level == 'volume' and path contains 202[12] or
#     level == 'article' and path contains volumes/7[01]

_harvest = (mmwr_dframe.level.str.fullmatch('home|series') |
            (mmwr_dframe.level.str.fullmatch('volume') &
             mmwr_dframe.mirror_path.str.contains('202[12]')) |
            (mmwr_dframe.level.str.fullmatch('article') &
             mmwr_dframe.mirror_path.str.contains('volumes/7[01]')))
# sum(_harvest) # 584

mmwr_sizes_b0 = [mirror_raw_html(url, MMWR_BASE_PATH_b0 + path, print_url = False)
                     for url, path in tqdm(zip(mmwr_dframe.url.loc[_harvest],
                                               mmwr_dframe.mirror_path.loc[_harvest]),
                                           total=584)]
# 584/584 [04:08<00:00,  2.35it/s]
# sum([x==0 for x in mmwr_sizes_b0]) # retry those with 0 length
for j in tqdm(range(584)):
   if mmwr_sizes_b0[j] == 0:
      mmwr_sizes_b0[j] = mirror_raw_html(mmwr_dframe.url.loc[_harvest][j],
         MMWR_BASE_PATH_b0 + mmwr_dframe.mirror_path.loc[_harvest][j], timeout=5)
# pickle.dump(mmwr_sizes_b0, open('mmwr_sizes_b0.pkl', 'wb'))

_harvest = mmwr_dframe.filename.str.fullmatch('mm70(23a3|34a7).htm')
mmwr_sizes_b0_ = [mirror_raw_html(url, MMWR_BASE_PATH_b0 + path, print_url = False)
                     for url, path in zip(mmwr_dframe.url.loc[_harvest],
                                          mmwr_dframe.mirror_path.loc[_harvest])]

mmwr_pdf_sizes_b0 = [mirror_raw_html(url, MMWR_BASE_PATH_pdf + '/' + flnm, print_url =
False)
                     for url, flnm in tqdm(zip(mmwr_pdf_dframe.url,
                     mmwr_pdf_dframe.filename),
                                           total=2066)]
# 2066/2066 [04:08<00:00,  2.35it/s]
# sum([x==0 for x in mmwr_pdf_sizes_b0]) # retry those with 0 length
# href for volumes 46 and 47 erroneously point to FTP
# https://www.cdc.gov/mmwr/PDF/wk/mm4601.pdf
for iss in tqdm(list(range(4601,4653)) + [4654] + list(range(4701,4752)) + [4753]):
```

```
111     mirror_raw_html(f'https://www.cdc.gov/mmwr/PDF/wk/mm{iss}.pdf',
112                     MMWR_BASE_PATH_pdf + '/mm' + f'{iss}.pdf', print_url = False)
113
114 # mirror_raw_html(pcd_dframe.url[200], PCD_BASE_PATH_b0 + pcd_dframe.mirror_path[200])
115
116 pcd_sizes_b0 = [mirror_raw_html(url, PCD_BASE_PATH_b0 + path, print_url = False)
117                     for url, path in tqdm(zip(pcd_dframe.url, pcd_dframe.mirror_path),
118                                           total=3777)]
119 # sum([x==0 for x in pcd_sizes_b0]) # retry those with 0 length
120 for j in range(3777):
121    if pcd_sizes_b0[j] == 0:
122       pcd_sizes_b0[j] = mirror_raw_html(pcd_dframe.url[j],
123          PCD_BASE_PATH_b0 + pcd_dframe.mirror_path[j], timeout=5)
124 # sum([x==0 for x in pcd_sizes_b0]) # retry those with 0 length
125
126 # pickle.dump(pcd_sizes_b0, open('pcd_sizes_b0.pkl', 'wb'))
127
128 # mirror_raw_html(eid_dframe.url[200], EID_BASE_PATH_b0 + eid_dframe.mirror_path[200])
129
130 eid_sizes_b0 = [mirror_raw_html(url, EID_BASE_PATH_b0 + path, print_url = False, timeout =
130 8)
131                     for url, path in tqdm(zip(eid_dframe.url, eid_dframe.mirror_path),
132                                           total=11504)]
133 # sum([x==0 for x in eid_sizes_b0]) # retry those with 0 length
134 for j in range(11504):
135    if eid_sizes_b0[j] == 0:
136       eid_sizes_b0[j] = mirror_raw_html(eid_dframe.url[j],
137          EID_BASE_PATH_b0 + eid_dframe.mirror_path[j], timeout=5)
138 # pickle.dump(eid_sizes_b0, open('eid_sizes_b0.pkl', 'wb'))
139
140 # phr_sizes_b0 = [mirror_raw_html(url, PHR_BASE_PATH_b0 + path, timeout = 5)
141 #                       for url, path in zip(phr_dframe.url, phr_dframe.mirror_path[:142])]
142 # sum([x==0 for x in phr_sizes_b0]) # retry those with 0 length
143 # mirroring works for /pmc/issues [:142] but not /pmc/articles [142:]
144 # pickle.dump(phr_sizes_b0, open('phr_sizes_b0.pkl', 'wb'))
145
146
147 #%% Read unprocessed HTML from local mirror; store in pickle format
148
149 mmwr_html_b0 = [read_raw_html(MMWR_BASE_PATH_b0 + path)
150                     for path in tqdm(mmwr_dframe.mirror_path)]
151 # 14751/14751 [00:04<00:00, 2954.78it/s]
152 pickle.dump(mmwr_html_b0, open('mmwr_raw_html.pkl', 'wb'))
153
154 pcd_html_b0 = [read_raw_html(PCD_BASE_PATH_b0 + path)
155                     for path in tqdm(pcd_dframe.mirror_path)]
156 ## 3627/3627 [00:08<00:00, 444.38it/s]
157 # 3777/3777 [00:01<00:00, 2547.93it/s]
158 pickle.dump(pcd_html_b0, open('pcd_raw_html.pkl', 'wb'))
159
160 # [EID_BASE_PATH_b0 + path for path in eid_dframe.mirror_path
161 #    if not os.path.exists(EID_BASE_PATH_b0 + path)]
162
163 eid_html_b0 = [read_raw_html(EID_BASE_PATH_b0 + path)
164                     for path in tqdm(eid_dframe.mirror_path)]
165 ## 10922/10922 [00:20<00:00, 521.50it/s]
```

```
166 # 11504/11504 [00:06<00:00, 1784.81it/s]
167 pickle.dump(eid_html_b0, open('eid_raw_html.pkl', 'wb'))
168
169 #%% Set up local mirror directories for lightly processed HTML (u3)
170
171 MMWR_BASE_PATH_u3 = normpath(expanduser('~/cdc-corpora/mmwr_u3/'))
172 PCD_BASE_PATH_u3 = normpath(expanduser('~/cdc-corpora/pcd_u3/'))
173 EID_BASE_PATH_u3 = normpath(expanduser('~/cdc-corpora/eid_u3/'))
174 # PHR_BASE_PATH_u3 = normpath(expanduser('~/cdc-corpora/phr_u3/'))
175
176 x = create_mirror_tree(MMWR_BASE_PATH_u3, calculate_mirror_dirs(mmwr_dframe.path))
177 # { key: (0 if val is None else len(val)) for (key, val) in x.items() }
178
179 x = create_mirror_tree(PCD_BASE_PATH_u3, calculate_mirror_dirs(pcd_dframe.path))
180 # { key: (0 if val is None else len(val)) for (key, val) in x.items() }
181
182 x = create_mirror_tree(EID_BASE_PATH_u3, calculate_mirror_dirs(eid_dframe.path))
183 # { key: (0 if val is None else len(val)) for (key, val) in x.items() }
184
185 #%% Mirror unprocessed HTML to processed HTML (b0 -> u3)
186
187 # x = read_raw_html(MMWR_BASE_PATH_b0 + mmwr_dframe.mirror_path[548])
188 # mirror_raw_to_uni(MMWR_BASE_PATH_b0 + mmwr_dframe.mirror_path[548],
189 #                   MMWR_BASE_PATH_u3 + mmwr_dframe.mirror_path[548], 548)
190
191 for path in tqdm(mmwr_dframe.mirror_path):
192    mirror_raw_to_uni(MMWR_BASE_PATH_b0 + path, MMWR_BASE_PATH_u3 + path, counter=None)
193 # 14751/14751 [22:18<00:00, 11.02it/s]
194
195 for path in tqdm(pcd_dframe.mirror_path):
196    mirror_raw_to_uni(PCD_BASE_PATH_b0 + path, PCD_BASE_PATH_u3 + path, counter=None)
197 # 3777/3777 [02:52<00:00, 21.85it/s]
198
199 for path in tqdm(eid_dframe.mirror_path):
200    mirror_raw_to_uni(EID_BASE_PATH_b0 + path, EID_BASE_PATH_u3 + path, counter=None)
201 # 13800/13800 [24:20<00:00,  9.45it/s]
202
203 # Correct the codec for 1 file, as follows:
204 # mirror_raw_to_uni(MMWR_BASE_PATH_b0, MMWR_BASE_PATH_u3, mmwr_dframe.mirror)
205 # issue with 13874: Some characters could not be decoded, and were replaced with
205 REPLACEMENT CHARACTER.
206 # code 81 in code page 437: b'\x81'.decode('cp437')
207 # https://www.cdc.gov/mmwr/preview/mmwrhtml/ss4808a2.htm
208 mmwr_dframe.iloc[14408]
209 ss4808a2_raw_html = read_raw_html(MMWR_BASE_PATH_b0 + mmwr_dframe.mirror_path[14408])
210 x = html_to_unicode_b(ss4808a2_raw_html)
211 # issue is character \x81 at ss4808a2_raw_html[51903:51904]
212 # per https://doi.org/10.1016/S0145-305X(97)00030-X, should be ü '\u00fc'
213
214 # Try adding CP437 to UnicodeDammit attempts
215 x = UnicodeDammit(ss4808a2_raw_html, ['utf-8', 'windows-1252', 'cp437']) # succeeds
216 x.tried_encodings # [('utf-8', 'strict'), ('windows-1252', 'strict'), ('cp437', 'strict')]
217 x.original_encoding # 'cp437'
218
219 # Commit this exception and write to UTF-8 mirror
220 ss4808a2_uni_html = trim_leading_space_u(
```

```
   html_prettify_u(
      html_reduce_space_u(
         UnicodeDammit(ss4808a2_raw_html, ['utf-8', 'windows-1252', 'cp437'])\
            .unicode_markup)))
with open(MMWR_BASE_PATH_u3 + mmwr_dframe.mirror_path[14408], 'w') as file_out:
   file_out.write(ss4808a2_uni_html)

#%% Read lightly processed HTML from local mirror; store in pickle format

mmwr_html_u3 = [read_uni_html(MMWR_BASE_PATH_u3 + path)
                     for path in tqdm(mmwr_dframe.mirror_path)]
# 14751/14751 [00:09<00:00, 1623.35it/s]
pickle.dump(mmwr_html_u3, open('mmwr_uni_html.pkl', 'wb'))

pcd_html_u3 = [read_uni_html(PCD_BASE_PATH_u3 + path)
                     for path in tqdm(pcd_dframe.mirror_path)]
# 3777/3777 [00:01<00:00, 3258.43it/s]
pickle.dump(pcd_html_u3, open('pcd_uni_html.pkl', 'wb'))

eid_html_u3 = [read_uni_html(EID_BASE_PATH_u3 + path)
                     for path in tqdm(eid_dframe.mirror_path)]
# 11504/11504 [00:09<00:00, 1153.86it/s]
pickle.dump(eid_html_u3, open('eid_uni_html.pkl', 'wb'))
```