Get Synsets Data From OEWN

1 Boilerplate

2 Imports

2.1 prod: NVM

```python
from nvm import disp_df
from nvm import clean_str
from nvm.aux_str import CLEAN_STR_MAPPINGS_LARGE as maps0
from nvm.aux_str import REGEX_ABC_DASH_XYZ_ASTERISK as re0
from nvm.aux_pandas import fix_column_names
```

2.2 prod: Basics

```python
import os
import pathlib
import numpy as np
import pandas as pd
import re
import json
import yaml
import srsly
import uuid
import random
import numbers
from collections import OrderedDict
from contextlib import ExitStack
import warnings
# warnings.warn("\nwarning")
from hashlib import md5
import humanfriendly as hf
import time
import datetime as dt
from pytz import timezone as tz
tz0 = tz("Europe/Berlin")
from glob import glob
from tqdm import tqdm
import logging
log0.info("DONE: basic imports")
```

2.3 prod: Extra imports and settings

```python
from contexttimer import Timer
import textwrap

HOME = pathlib.Path.home()

tqdm.pandas()

import matplotlib
from matplotlib import pyplot as plt
# import seaborn as sns
# import plotly.graph_objects as go
# import plotly.express as px

# get_ipython().run_line_magic("matplotlib", "qt")
# get_ipython().run_line_magic("matplotlib", "inline")

with Timer() as elapsed:
    time.sleep(0.001)

log0.info(hf.format_timespan(elapsed.elapsed))

log0.info("DONE: extra imports and settings")
```

3 Extra Imports

3.1 prod: More extra imports and settings

```python
import wn
from wn.morphy import Morphy
ANTONYM_SENSE_RELATIONS = [
    "antonym",
    "anto_gradable",
    "anto_simple",
    "anto_converse",
]

log0.info(f"{wn.__file__}")
log0.info("DONE: more extra imports and settings")
```

4 Process

4.1 Prepare WordNet

```python
wn0 = wn.Wordnet(
    "oewn:2021",
    lang="en",
    lemmatizer=Morphy(),
    search_all_forms=True,
)
```

4.2 prod: Load data

```
dir0 = "../../data/d0007_synsets-selected/"
dir0 = pathlib.Path(dir0)
# dir0.mkdir(mode=0o700, parents=True, exist_ok=True)
assert dir0.exists(), f"The data directory dir0={str(dir0)} not found!"

name0 = f"synsets"
extn0 = ".yaml"

if0 = (dir0/name0).with_suffix(extn0)
log0.info(f"loading: {if0}...")
ss_list = srsly.read_yaml(if0)
log0.info(f"loading: {if0}... DONE")

log0.info(f"{len(ss_list) = }")
print(srsly.yaml_dumps(ss_list[:12]))
```

```
I: loading: ../../data/d0007_synsets-selected/synsets.yaml...
I: loading: ../../data/d0007_synsets-selected/synsets.yaml... DONE
I: len(ss_list) = 6914
  - oewn-02854643-n
  - oewn-02078906-v
  - oewn-00443540-s
  - oewn-08079806-n
  - oewn-02090199-s
  - oewn-05622440-n
  - oewn-00185953-n
  - oewn-01525056-v
  - oewn-02713541-v
  - oewn-00365540-r
  - oewn-05651861-n
  - oewn-01896097-v
```

4.3 Synsets data dict

```
ss_data0 = []
for id0 in ss_list:
    ss0 = wn0.synset(id0)
    if not any(item["id0"]==ss0.id for item in ss_data0):
        antonym_ids = []
        antonym_defs = []
        antonym_lemmas = []
        antonym_examples = []
        for sense0 in ss0.senses():
            rels = sense0.relations(*ANTONYM_SENSE_RELATIONS)
            for relname, antonym_senses_list in rels.items():
                for sense2 in antonym_senses_list:
                    if sense2.synset().id not in antonym_ids:
                        if len(sense2.synset().examples())>0:
                            antonym_ids.append(sense2.synset().id)
                            antonym_defs.append(sense2.synset().definition())
                            antonym_lemmas.append(sense2.synset().lemmas())
                            antonym_examples.append(sense2.synset().examples())

        ss_data0.append(dict(
            id0=ss0.id,
            lemmas=ss0.lemmas(),
            definition=ss0.definition(),
            examples=ss0.examples(),
            antonym_ids=antonym_ids,
            antonym_lemmas=antonym_lemmas,
            antonym_defs=antonym_defs,
            antonym_examples=antonym_examples,
        ))

log0.info(f"{len(ss_data0) = }")
```

```
I: len(ss_data0) = 6914
```

4.4 Synsets DataFrame

```
df0 = pd.DataFrame.from_records(ss_data0)
df2 = df0[[col0 for col0 in df0.columns if not col0.startswith("antonym_") ]]
log0.info(f"{df0.shape = }")
disp_df(df0.sample(n=8).sort_index())
```

```
I: df0.shape = (6914, 8)

                  id0                              lemmas
1138  oewn-02684248-v  [worry, concern, occupy, interest]                            be on
1774  oewn-02456941-v                           [inhibit]  limit, block, or decrease the a
1799  oewn-00498547-n                  [draw, draw poker]  poker in which a player can dis
2345  oewn-02353009-s                           [supreme]         highest in excellence or
4163  oewn-02630209-v                              [head]        form a head or come or gr
4174  oewn-05846174-n                              [idea]                              a p
5052  oewn-10641415-n                           [soldier]  an enlisted man or woman who se
6242  oewn-00592037-v                             [touch]
```

4.5 Cols DF0

```python
for col0 in df0.columns:
    print(f"    \"{col0}\",")
```

```
    "id0",
    "lemmas",
    "definition",
    "examples",
    "antonym_ids",
    "antonym_lemmas",
    "antonym_defs",
    "antonym_examples",
```

4.6 Cols DF2

```python
for col0 in df2.columns:
    print(f"    \"{col0}\",")
```

```
    "id0",
    "lemmas",
    "definition",
    "examples",
```

4.7 Antonyms exclusive

```python
df4 = df0[df0.antonym_ids.apply(lambda x: x != [])]

log0.info(f"{df4.shape = }")
disp_df(df4.sample(n=8).sort_index())
```

```
I: df4.shape = (858, 8)

                  id0                                       lemmas
1918  oewn-07556441-n                                       [hope]  the general feeling th
2166  oewn-02217607-v                               [refuse, deny]
2823  oewn-00370083-r             [precisely, exactly, incisively]
2938  oewn-01762851-a                        [lasting, persistent]
4190  oewn-00338302-a               [incertain, uncertain, unsure]  lacking or indicating
5215  oewn-02605525-v                                       [fail]
5410  oewn-01768652-v  [quieten, calm, lull, calm down, still, ...
6857  oewn-14498478-n                                    [success]                a state
```

4.8 Save DF2

```
import pathlib
import csv
import datetime as dt
from pytz import timezone as tz
tz0 = tz("Europe/Berlin")

dir0 = "../../data/d0007_synsets-selected/"
dir0 = pathlib.Path(dir0)
dir0.mkdir(mode=0o700, parents=True, exist_ok=True)
assert dir0.exists(), f"The data directory dir0={str(dir0)} was not found!"

now0 = [dt.datetime.now(tz0).strftime("%Y%m%dT%H%M%S")]
now0 = []
pfx0 = ["sysnsets-data-0001-wn-text"]
sfx0 = []

bfn0 = dir0/"_".join(pfx0+now0+sfx0).replace(".", "_")

xtn0 = ".pkl"
ofn0 = bfn0.with_suffix(xtn0)
log0.info(f"saving: {ofn0}...")
df2.to_pickle(ofn0)

xtn0 = ".csv"
ofn0 = bfn0.with_suffix(xtn0)
log0.info(f"saving: {ofn0}...")
df2.to_csv(ofn0, index=False, quoting=csv.QUOTE_NONNUMERIC)

xtn0 = ".xlsx"
ofn0 = bfn0.with_suffix(xtn0)
log0.info(f"saving: {ofn0}...")
df2.to_excel(ofn0)

xtn0 = ".jsonl"
ofn0 = bfn0.with_suffix(xtn0)
log0.info(f"saving: {ofn0}...")
with open(ofn0, "w") as fh: pass
srsly.write_jsonl(ofn0, df2.to_dict(orient="records"))

log0.info("DONE")
```

```
I: saving: ../../data/d0007_synsets-selected/sysnsets-data-0001-wn-text.pkl...
I: saving: ../../data/d0007_synsets-selected/sysnsets-data-0001-wn-text.csv...
I: saving: ../../data/d0007_synsets-selected/sysnsets-data-0001-wn-text.xlsx...
I: saving: ../../data/d0007_synsets-selected/sysnsets-data-0001-wn-text.jsonl...
I: DONE
```

4.9 Save DF0

```
import pathlib
import csv
import datetime as dt
from pytz import timezone as tz
tz0 = tz("Europe/Berlin")

dir0 = "../../data/d0007_synsets-selected/"
dir0 = pathlib.Path(dir0)
dir0.mkdir(mode=0o700, parents=True, exist_ok=True)
assert dir0.exists(), f"The data directory dir0={str(dir0)} was not found!"

now0 = [dt.datetime.now(tz0).strftime("%Y%m%dT%H%M%S")]
now0 = []
pfx0 = ["sysnsets-data-0002-wn-text-with-antonyms"]
sfx0 = []

bfn0 = dir0/"_".join(pfx0+now0+sfx0).replace(".", "_")

xtn0 = ".pkl"
ofn0 = bfn0.with_suffix(xtn0)
log0.info(f"saving: {ofn0}...")
df0.to_pickle(ofn0)

xtn0 = ".csv"
ofn0 = bfn0.with_suffix(xtn0)
log0.info(f"saving: {ofn0}...")
df0.to_csv(ofn0, index=False, quoting=csv.QUOTE_NONNUMERIC)

xtn0 = ".xlsx"
ofn0 = bfn0.with_suffix(xtn0)
log0.info(f"saving: {ofn0}...")
df0.to_excel(ofn0)

xtn0 = ".jsonl"
ofn0 = bfn0.with_suffix(xtn0)
log0.info(f"saving: {ofn0}...")
with open(ofn0, "w") as fh: pass
srsly.write_jsonl(ofn0, df0.to_dict(orient="records"))

log0.info("DONE")
```

```
I: saving: ../../data/d0007_synsets-selected/sysnsets-data-0002-wn-text-with-antonyms.pkl.
I: saving: ../../data/d0007_synsets-selected/sysnsets-data-0002-wn-text-with-antonyms.csv.
I: saving: ../../data/d0007_synsets-selected/sysnsets-data-0002-wn-text-with-antonyms.xlsx
I: saving: ../../data/d0007_synsets-selected/sysnsets-data-0002-wn-text-with-antonyms.json
I: DONE
```

4.10 Save DF4

```
import pathlib
import csv
import datetime as dt
from pytz import timezone as tz
```

```
tz0 = tz("Europe/Berlin")

dir0 = "../../data/d0007_synsets-selected/"
dir0 = pathlib.Path(dir0)
dir0.mkdir(mode=0o700, parents=True, exist_ok=True)
assert dir0.exists(), f"The data directory dir0={str(dir0)} was not found!"

now0 = [dt.datetime.now(tz0).strftime("%Y%m%dT%H%M%S")]
now0 = []
pfx0 = ["sysnsets-data-0004-wn-text-only-antonyms"]
sfx0 = []

bfn0 = dir0/"_".join(pfx0+now0+sfx0).replace(".", "_")

xtn0 = ".pkl"
ofn0 = bfn0.with_suffix(xtn0)
log0.info(f"saving: {ofn0}...")
df4.to_pickle(ofn0)

xtn0 = ".csv"
ofn0 = bfn0.with_suffix(xtn0)
log0.info(f"saving: {ofn0}...")
df4.to_csv(ofn0, index=False, quoting=csv.QUOTE_NONNUMERIC)

xtn0 = ".xlsx"
ofn0 = bfn0.with_suffix(xtn0)
log0.info(f"saving: {ofn0}...")
df4.to_excel(ofn0)

xtn0 = ".jsonl"
ofn0 = bfn0.with_suffix(xtn0)
log0.info(f"saving: {ofn0}...")
with open(ofn0, "w") as fh: pass
srsly.write_jsonl(ofn0, df4.to_dict(orient="records"))

log0.info("DONE")
```

```
I: saving: ../../data/d0007_synsets-selected/sysnsets-data-0004-wn-text-only-antonyms.pkl.
I: saving: ../../data/d0007_synsets-selected/sysnsets-data-0004-wn-text-only-antonyms.csv.
I: saving: ../../data/d0007_synsets-selected/sysnsets-data-0004-wn-text-only-antonyms.xlsx
I: saving: ../../data/d0007_synsets-selected/sysnsets-data-0004-wn-text-only-antonyms.json
I: DONE
```