# S1001_training Data Inspection

## 1 Boilerplate

## 2 Imports

### 2.1 prod: NVM

```
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

```
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

```
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

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

## 4 Process

### 4.1 Load synsets 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"sysnsets-data-0002-wn-text-with-antonyms"
extn0 = ".pkl"

if0 = (dir0/name0).with_suffix(extn0)
log0.info(f"loading: {if0}...")
df0 = pd.read_pickle(if0)
log0.info(f"loading: {if0}... DONE")
```

```
df0.insert(0, "TYPE", "src0")
df0["antonym_count_orig"] = df0.antonym_ids.apply(lambda x: len(x))
df0["antonym_count_uniq"] = df0.antonym_ids.apply(lambda x: len(set(x)))

log0.info(f"{df0.shape = }")
disp_df(df0.sample(n=8).sort_index())
```

```
I: loading: ../../data/d0007_synsets-selected/sysnsets-data-0002-wn-text-with-antonyms.pkl
I: loading: ../../data/d0007_synsets-selected/sysnsets-data-0002-wn-text-with-antonyms.pkl
I: df0.shape = (6914, 11)
```

| | TYPE | id0 | lemmas | |
|---|---|---|---|---|
| 1247 | src0 | oewn-01048256-s | [everyday, casual, daily] | appropriate for |
| 2627 | src0 | oewn-02402779-v | [elevate, advance, raise, upgrade, promo... | give a promotion |
| 4688 | src0 | oewn-02881464-n | [freighter, merchantman, merchant ship, ... | |
| 5109 | src0 | oewn-00942956-n | [approach, attack, plan of attack] | ideas or actions |
| 5983 | src0 | oewn-00825528-s | [honorary] | given as an hono |
| 6425 | src0 | oewn-00972247-v | [vulgarize, vulgarise, popularize, popul... | cater to popular |
| 6617 | src0 | oewn-00205951-r | [accurately] | |
| 6723 | src0 | oewn-14544431-n | [depression, low] | an air mass of l |

## 4.2 prod: Check antonym counts ORIG

```
df0.antonym_count_orig.value_counts()
```

```
antonym_count_orig
0    6056
1     844
2      14
Name: count, dtype: int64
```

## 4.3 prod: Check antonym counts UNIQ

```
df0.antonym_count_uniq.value_counts()
```

```
antonym_count_uniq
0    6056
1     844
2      14
Name: count, dtype: int64
```

## 4.4 Columns DF0

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

```
    "TYPE",
    "id0",
    "lemmas",
    "definition",
    "examples",
    "antonym_ids",
    "antonym_lemmas",
    "antonym_defs",
    "antonym_examples",
    "antonym_count_orig",
    "antonym_count_uniq",
```

### 4.5 prod: Explode antonyms

```
explode_cols = [
    "antonym_ids",
    "antonym_lemmas",
    "antonym_defs",
    "antonym_examples",
]
df2 = df0.explode(column=explode_cols)
df2.rename(columns={"antonym_ids": "antonym_id0"}, inplace=True)

log0.info(f"{df0.shape = }")
log0.info(f"{df2.shape = }")
disp_df(df2.sample(n=8).sort_index())
```

```
I: df0.shape = (6914, 11)
I: df2.shape = (6928, 11)
```

| | TYPE | id0 | lemmas | |
|---|---|---|---|---|
| 132 | src0 | oewn-00844451-v | [pick, blame, find fault] | har |
| 676 | src0 | oewn-02565025-v | [undo] | ca |
| 2370 | src0 | oewn-02465750-s | [careful] | |
| 2634 | src0 | oewn-00625102-n | [trouble, difficulty] | an |
| 3719 | src0 | oewn-00757628-a | [diligent] | characterized by |
| 4230 | src0 | oewn-00037570-a | [active] | tending to becom |
| 6368 | src0 | oewn-00543200-v | [come] | reach or enter a |
| 6628 | src0 | oewn-07211342-n | [question, interrogation, interrogative ... | a sentence of in |

### 4.6 Collapse after explosion

```
log0.info(f"{df0.shape = }")
log0.info(f"{df2.shape = }")
log0.info(f"{df2.id0.unique().shape[0] = }")
df2 = df2.drop_duplicates(subset=["id0", "antonym_id0"], ignore_index=True)

log0.info(f"{df2.shape = }")
disp_df(df2.sample(n=8).sort_index())
```

```
I: df0.shape = (6914, 11)
I: df2.shape = (6928, 11)
I: df2.id0.unique().shape[0] = 6914
I: df2.shape = (6928, 11)
```

| | TYPE | id0 | lemmas | |
|---|---|---|---|---|
| 1205 | src0 | oewn-02711906-v | [vacillate, hover, vibrate, oscillate] | be undecided abou |
| 2370 | src0 | oewn-02474747-s | [true, reliable, dependable, honest] | w |
| 2936 | src0 | oewn-13518338-n | [increment, increase, growth] | a process of beco |
| 3371 | src0 | oewn-03389381-n | [form] | a |
| 4686 | src0 | oewn-02245872-s | [aerodynamic, sleek, streamlined, flowing] | designed or arran |
| 5438 | src0 | oewn-02054955-v | [make, work] | |
| 6030 | src0 | oewn-00294242-s | [drudging, toiling, laboring, labouring] | doing |
| 6925 | src0 | oewn-01916886-v | [walk] | tr |

## 4.7 Check before explosion

```
disp_df(df0[df0.id0=="oewn-00215173-r"])
```

| | TYPE | id0 | lemmas | definition |
|---|---|---|---|---|
| 1732 | src0 | oewn-00215173-r | [fruitfully, profitably, productively] | in a productive way |

## 4.8 Check explosion

```
log0.info(f"{df2[df2.antonym_count_orig>0].shape = }")
log0.info(f"{df2.shape = }")
disp_df(df2[df2.antonym_count_orig>0].head(n=22).sort_index())
```

```
I: df2[df2.antonym_count_orig>0].shape = (872, 11)
I: df2.shape = (6928, 11)
```

| | TYPE | id0 | lemmas | |
|---|---|---|---|---|
| 7 | src0 | oewn-01525056-v | [twine, wrap, roll, wind] | |
| 17 | src0 | oewn-00117793-v | [rest, stay, remain] | stay the same; |
| 34 | src0 | oewn-00088791-r | [temporarily] | for a limited |
| 41 | src0 | oewn-02639795-v | [exclude] | |
| 47 | src0 | oewn-00752852-a | [easy] | posing no difficu |
| 55 | src0 | oewn-00018622-a | [unacceptable] | n |
| 56 | src0 | oewn-00153083-v | [increase] | |
| 60 | src0 | oewn-01244630-a | [running] | of adv |
| 68 | src0 | oewn-00244122-r | [consciously] | |
| 74 | src0 | oewn-00688760-a | [indecisive] | not def |
| 76 | src0 | oewn-00074467-r | [back, backward] | in o |
| 94 | src0 | oewn-01101986-a | [funded] | |
| 106 | src0 | oewn-00905999-v | [discharge, clear, acquit, assoil, excul... | pronounce not |
| 110 | src0 | oewn-01408930-a | [biological, natural] | (of a parent or c |
| 112 | src0 | oewn-00096162-r | [down] | away from a more |
| 130 | src0 | oewn-01084999-v | [enter, participate] | become a p |

| | | | | |
|---|---|---|---|---|
| 139 | src0 | oewn-02211988-v | [buy, purchase] | obtain by purchas |
| 149 | src0 | oewn-00687584-v | [reject] | not accept |
| 152 | src0 | oewn-02091950-a | [scientific] | conforming with t |
| 155 | src0 | oewn-00227509-r | [big] | |
| 157 | src0 | oewn-01933854-a | [intellectual, cerebral] | involving intelli |
| 160 | src0 | oewn-02337347-a | [subordinate] | subject or submis |

## 4.9 prod: Load Evaluation data

```
dir0 = "../../data/d0008_synsets-evaluated/"
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-eval-004-summary-clean-all"
extn0 = ".pkl"

if0 = (dir0/name0).with_suffix(extn0)
log0.info(f"loading: {if0}...")
df9 = pd.read_pickle(if0)
log0.info(f"loading: {if0}... DONE")

log0.info(f"{df9.shape = }")
disp_df(df9.sample(n=8).sort_index())
```

```
I: loading: ../../data/d0008_synsets-evaluated/synsets-eval-004-summary-clean-all.pkl...
I: loading: ../../data/d0008_synsets-evaluated/synsets-eval-004-summary-clean-all.pkl... D
I: df9.shape = (6914, 4)
```

| | OEWN | CNT | MEAN | STD |
|---|---|---|---|---|
| 567 | oewn-00188186-n | 8 | 1.750000 | 1.035098 |
| 1761 | oewn-00715729-n | 7 | 0.571429 | 1.272418 |
| 2009 | oewn-00807941-s | 8 | 0.875000 | 1.125992 |
| 4589 | oewn-02373016-a | 8 | -0.625000 | 1.767767 |
| 5536 | oewn-04939014-n | 7 | 0.000000 | 0.000000 |
| 5723 | oewn-05691808-n | 8 | -1.125000 | 0.834523 |
| 5980 | oewn-06649049-n | 8 | 0.125000 | 0.353553 |
| 6860 | oewn-14550556-n | 9 | 0.666667 | 1.414214 |

## 4.10 Merge SOURCE synsets with evaluations

```
df3 = pd.merge(
    df2,
    df9,
    how="left",
    left_on="id0",
    right_on="OEWN",
    sort=False,
    suffixes=("_df2", "_df9"),
    copy=True,
```

```
    validate="m:1",
).copy()

df3.insert(2, "OEWN", df3.pop("OEWN"))
df3.insert(3, "MEAN", df3.pop("MEAN"))
df3.insert(4, "STD", df3.pop("STD"))
df3.insert(5, "CNT", df3.pop("CNT"))

assert df3["id0"].equals(df3["OEWN"])
df3.drop(columns=["OEWN"], inplace=True)

log0.info(f"{df3.shape = }")
disp_df(df3.sample(n=8).sort_index())
```

```
I: df3.shape = (6928, 14)
```

| | TYPE | id0 | MEAN | STD | CNT | lemmas | |
|---|---|---|---|---|---|---|---|
| 800 | src0 | oewn-02288414-s | 0.888889 | 0.927961 | 9 | [whipping, snappy] | |
| 1895 | src0 | oewn-07438160-n | 0.444444 | 1.013794 | 9 | [development] | a rec |
| 2580 | src0 | oewn-02275064-a | 0.000000 | 0.000000 | 9 | [some] | (quan |
| 2629 | src0 | oewn-00005041-v | 0.857143 | 1.214986 | 7 | [inspire, inhale, breathe in] | |
| 3232 | src0 | oewn-00308592-s | -2.125000 | 0.834523 | 8 | [unable] | (usua |
| 3914 | src0 | oewn-06811608-n | 0.000000 | 0.000000 | 8 | [print, mark] | |
| 5082 | src0 | oewn-04888818-n | 1.375000 | 1.302470 | 8 | [discipline] | |
| 6292 | src0 | oewn-01144162-n | -0.250000 | 0.886405 | 8 | [allowance] | |

## 4.11 Merge ANTONYM synsets with evaluations

```
df4 = pd.merge(
    df3,
    df9,
    how="left",
    left_on="antonym_id0",
    right_on="OEWN",
    sort=False,
    suffixes=("", "_antonym"),
    copy=True,
    validate="m:1",
).copy()

# df4.insert(2, "OEWN", df4.pop("OEWN"))
# df4.insert(3, "MEAN", df4.pop("MEAN"))
# df4.insert(4, "MEAN", df4.pop("MEAN"))
# df4.insert(5, "STD", df4.pop("STD"))
# df4.insert(6, "CNT", df4.pop("CNT"))

# assert df4["antonym_id0"].equals(df4["OEWN"])
# df4.drop(columns=["OEWN"], inplace=True)

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

```
I: df4.shape = (6928, 18)
```

| | TYPE | id0 | MEAN | STD | CNT | lemmas | |
|---|---|---|---|---|---|---|---|
| 1136 | src0 | oewn-04925387-n | 1.600000 | 0.843274 | 10 | [directness, straightness] | |
| 2095 | src0 | oewn-00723813-v | 0.000000 | 0.000000 | 8 | [price] | |
| 3023 | src0 | oewn-15170131-n | 0.000000 | 0.000000 | 6 | [age, eld] | a |
| 3341 | src0 | oewn-01379820-a | -0.333333 | 1.000000 | 9 | [unknown] | |
| 3849 | src0 | oewn-00067913-r | 1.333333 | 1.000000 | 9 | [in the lead, out front, ahead] | |
| 4446 | src0 | oewn-02588127-v | 1.000000 | 1.195229 | 8 | [process, action, sue, litigate] | in |
| 4533 | src0 | oewn-02903285-a | 0.000000 | 0.000000 | 9 | [rental] | |
| 4968 | src0 | oewn-00156320-r | -0.428571 | 0.786796 | 7 | [uncomfortably] | |

## 4.12 DF5

```
df5 = df4[df4.antonym_count_uniq>0].copy()

log0.info(f"{df5.shape = }")
log0.info(f"{df5.MEAN_antonym.isna().sum() = } BAD COUNT")  # WARN: This count may contain duplicat
log0.info(f"{(~df5.MEAN_antonym.isna()).sum() = } BAD COUNT")  # WARN: This count may contain dupli
disp_df(df5.sample(n=8).sort_index())
```

```
I: df5.shape = (872, 18)
I: df5.MEAN_antonym.isna().sum() = 440 BAD COUNT
I: (~df5.MEAN_antonym.isna()).sum() = 432 BAD COUNT
```

| | TYPE | id0 | MEAN | STD | CNT | |
|---|---|---|---|---|---|---|
| 112 | src0 | oewn-00096162-r | -0.500000 | 1.309307 | 8 | |
| 1216 | src0 | oewn-01026905-a | -0.875000 | 0.834523 | 8 | [infl |
| 1361 | src0 | oewn-01974362-v | 1.181818 | 1.078720 | 11 | [come up, rise, ascend, |
| 2185 | src0 | oewn-02034163-a | 0.285714 | 0.487950 | 7 | |
| 5422 | src0 | oewn-01768652-v | -0.714286 | 1.603567 | 7 | [quieten, calm, lull, calm down, sti |
| 5558 | src0 | oewn-02528530-v | 2.250000 | 0.707107 | 8 | [manage, pull off, carry off, negoci |
| 6037 | src0 | oewn-02332106-a | -1.500000 | 1.195229 | 8 | |
| 6052 | src0 | oewn-01859592-v | 1.125000 | 0.991031 | 8 | |

## 4.13 Check SOURCE and ANTONYM correlations of EVALUATIONS

```
df5[["MEAN", "MEAN_antonym"]].corr()
```

| | MEAN | MEAN_antonym |
|---|---|---|
| MEAN | 1.000000 | -0.800623 |
| MEAN_antonym | -0.800623 | 1.000000 |

## 4.14 Check the actual number of unique antonyms

```
temp0 = df5.drop_duplicates(subset="antonym_id0", keep="first").copy()
log0.info(f"{df5.shape = }")
log0.info(f"{temp0.shape = }")
log0.info(f"{temp0.MEAN_antonym.isna().sum() = } GOOD COUNT")  # This count should contain NO dupli
log0.info(f"{(~temp0.MEAN_antonym.isna()).sum() = } GOOD COUNT")  # This count should contain NO du
```

```
I: df5.shape = (872, 18)
I: temp0.shape = (867, 18)
I: temp0.MEAN_antonym.isna().sum() = 436 GOOD COUNT
I: (~temp0.MEAN_antonym.isna()).sum() = 431 GOOD COUNT
```

## 4.15 Reminder

```
log0.info(f"{df5.shape = }")
disp_df(df5.head(n=22))
```

```
I: df5.shape = (872, 18)
```

| | TYPE | id0 | MEAN | STD | CNT | |
|---|---|---|---|---|---|---|
| 7 | src0 | oewn-01525056-v | 0.400000 | 0.699206 | 10 | [twine, wrap, roll, |
| 17 | src0 | oewn-00117793-v | -1.375000 | 1.302470 | 8 | [rest, stay, r |
| 34 | src0 | oewn-00088791-r | -0.727273 | 1.009050 | 11 | [tempor |
| 41 | src0 | oewn-02639795-v | -1.500000 | 1.080123 | 10 | [ex |
| 47 | src0 | oewn-00752852-a | -0.333333 | 1.556998 | 12 | |
| 55 | src0 | oewn-00018622-a | -1.200000 | 1.316561 | 10 | [unaccep |
| 56 | src0 | oewn-00153083-v | 1.222222 | 0.971825 | 9 | [inc |
| 60 | src0 | oewn-01244630-a | 0.428571 | 1.133893 | 7 | [ru |
| 68 | src0 | oewn-00244122-r | 0.700000 | 0.948683 | 10 | [consci |
| 74 | src0 | oewn-00688760-a | -2.000000 | 0.707107 | 9 | [indec |
| 76 | src0 | oewn-00074467-r | -0.400000 | 0.699206 | 10 | [back, bac |
| 94 | src0 | oewn-01101986-a | 0.222222 | 0.440959 | 9 | [f |
| 106 | src0 | oewn-00905999-v | 0.222222 | 0.440959 | 9 | [discharge, clear, acquit, assoil, ex |
| 110 | src0 | oewn-01408930-a | 0.000000 | 0.000000 | 8 | [biological, na |
| 112 | src0 | oewn-00096162-r | -0.500000 | 1.309307 | 8 | |
| 130 | src0 | oewn-01084999-v | 1.272727 | 1.103713 | 11 | [enter, partic |
| 139 | src0 | oewn-02211988-v | 0.909091 | 1.136182 | 11 | [buy, pur |
| 149 | src0 | oewn-00687584-v | -1.222222 | 1.481366 | 9 | [r |
| 152 | src0 | oewn-02091950-a | 0.444444 | 0.881917 | 9 | [scien |
| 155 | src0 | oewn-00227509-r | 1.111111 | 1.054093 | 9 | |
| 157 | src0 | oewn-01933854-a | 1.000000 | 1.154701 | 7 | [intellectual, cer |
| 160 | src0 | oewn-02337347-a | -1.666667 | 1.118034 | 9 | [subord |

## 4.16 Augument antonym evaluations when missing

```
import math
df6 = df5.copy()
df6["antonym_TYPE"] = df6.apply(lambda row: "ant2" if math.isnan(row["MEAN_antonym"]) else "ant0", a
```

```
df6["MEAN_antonym"] = df6.apply(lambda row: -row["MEAN"] if math.isnan(row["MEAN_antonym"]) else row
cols5 = [
    "TYPE",
    "id0",
    "MEAN",
    "lemmas",
    "antonym_id0",
    "antonym_lemmas",
    "MEAN_antonym",
    "antonym_TYPE",
]
disp_df(df6[cols5].head(n=22))
```

| | TYPE | id0 | MEAN | lemmas | ant |
|---|---|---|---|---|---|
| 7 | src0 | oewn-01525056-v | 0.400000 | [twine, wrap, roll, wind] | oewn-01 |
| 17 | src0 | oewn-00117793-v | -1.375000 | [rest, stay, remain] | oewn-00 |
| 34 | src0 | oewn-00088791-r | -0.727273 | [temporarily] | oewn-00 |
| 41 | src0 | oewn-02639795-v | -1.500000 | [exclude] | oewn-02 |
| 47 | src0 | oewn-00752852-a | -0.333333 | [easy] | oewn-00 |
| 55 | src0 | oewn-00018622-a | -1.200000 | [unacceptable] | oewn-00 |
| 56 | src0 | oewn-00153083-v | 1.222222 | [increase] | oewn-00 |
| 60 | src0 | oewn-01244630-a | 0.428571 | [running] | oewn-01 |
| 68 | src0 | oewn-00244122-r | 0.700000 | [consciously] | oewn-00 |
| 74 | src0 | oewn-00688760-a | -2.000000 | [indecisive] | oewn-00 |
| 76 | src0 | oewn-00074467-r | -0.400000 | [back, backward] | oewn-00 |
| 94 | src0 | oewn-01101986-a | 0.222222 | [funded] | oewn-01 |
| 106 | src0 | oewn-00905999-v | 0.222222 | [discharge, clear, acquit, assoil, excul... | oewn-00 |
| 110 | src0 | oewn-01408930-a | 0.000000 | [biological, natural] | oewn-01 |
| 112 | src0 | oewn-00096162-r | -0.500000 | [down] | oewn-00 |
| 130 | src0 | oewn-01084999-v | 1.272727 | [enter, participate] | oewn-01 |
| 139 | src0 | oewn-02211988-v | 0.909091 | [buy, purchase] | oewn-02 |
| 149 | src0 | oewn-00687584-v | -1.222222 | [reject] | oewn-00 |
| 152 | src0 | oewn-02091950-a | 0.444444 | [scientific] | oewn-02 |
| 155 | src0 | oewn-00227509-r | 1.111111 | [big] | oewn-00 |
| 157 | src0 | oewn-01933854-a | 1.000000 | [intellectual, cerebral] | oewn-01 |
| 160 | src0 | oewn-02337347-a | -1.666667 | [subordinate] | oewn-02 |

## 4.17 DF5 Cols

```
for col0 in df6.columns:
    print(f"    \"{col0}\",")
```

```
    "TYPE",
    "id0",
    "MEAN",
    "STD",
    "CNT",
    "lemmas",
    "definition",
```

```
    "examples",
    "antonym_id0",
    "antonym_lemmas",
    "antonym_defs",
    "antonym_examples",
    "antonym_count_orig",
    "antonym_count_uniq",
    "OEWN",
    "CNT_antonym",
    "MEAN_antonym",
    "STD_antonym",
    "antonym_TYPE",
```

### 4.18 Extract unique augumented and unaugumented antonyms

```
cols7 = {
    "antonym_TYPE": "TYPE",
    "antonym_id0": "id0",
    "MEAN_antonym": "MEAN",
    "STD_antonym": "STD",
    "CNT_antonym": "CNT",
    "antonym_lemmas": "lemmas",
    "antonym_defs": "definition",
    "antonym_examples": "examples",
}
df7 = df6[cols7.keys()].copy()
df7.rename(columns=cols7, inplace=True)
df7 = df7[df7.TYPE=="ant2"]
df7.drop_duplicates(subset="id0", inplace=True)

log0.info(f"{df7.shape = }")
disp_df(df7.head(n=22))
```

```
I: df7.shape = (436, 8)
```

| | TYPE | id0 | MEAN | STD | CNT | lemma |
|---|---|---|---|---|---|---|
| 7 | ant2 | oewn-01526442-v | -0.400000 | NaN | NaN | [wind off, unwind, unroll |
| 34 | ant2 | oewn-00088404-r | 0.727273 | NaN | NaN | [for good, permanently |
| 41 | ant2 | oewn-02639021-v | 1.500000 | NaN | NaN | [include |
| 56 | ant2 | oewn-00442400-v | -1.222222 | NaN | NaN | [decrease, lessen, minify |
| 68 | ant2 | oewn-00244269-r | -0.700000 | NaN | NaN | [unconsciously |
| 94 | ant2 | oewn-01102083-a | -0.222222 | NaN | NaN | [unfunded |
| 106 | ant2 | oewn-00908320-v | -0.222222 | NaN | NaN | [convict |
| 110 | ant2 | oewn-01409235-a | -0.000000 | NaN | NaN | [adoptive |
| 130 | ant2 | oewn-01085437-v | -1.272727 | NaN | NaN | [quit, give up, drop by the wayside, dro.. |
| 152 | ant2 | oewn-02092278-a | -0.444444 | NaN | NaN | [unscientific |
| 157 | ant2 | oewn-01933678-a | -1.000000 | NaN | NaN | [emotional |
| 160 | ant2 | oewn-02338328-a | 1.666667 | NaN | NaN | [insubordinate |
| 174 | ant2 | oewn-00233660-a | -0.375000 | NaN | NaN | [worsened, worse |

```
195  ant2  oewn-00115368-r -0.636364  NaN  NaN                                [unofficially
199  ant2  oewn-01574863-a  0.100000  NaN  NaN                                   [unnatural
203  ant2  oewn-02291437-a -0.000000  NaN  NaN                                      [spoken
208  ant2  oewn-14053917-n  0.250000  NaN  NaN          [drowsiness, sleepiness, somnolence
213  ant2  oewn-01618017-a -0.000000  NaN  NaN                                [disobedient
218  ant2  oewn-01475771-a  0.888889  NaN  NaN                                      [major
266  ant2  oewn-00494722-a  0.083333  NaN  NaN                                     [common
269  ant2  oewn-01109313-a -0.300000  NaN  NaN                                [nonspecific
300  ant2  oewn-01096835-a -0.363636  NaN  NaN                                   [organic
```

### 4.19 Augument with lemma negations

```python
df8 = df6.copy()

df8.insert(6, "lemmas_orig", df8.lemmas)
df8.insert(7, "lemmas_neg", df8.antonym_lemmas)

df8.insert(12, "antonym_lemmas_orig", df8.antonym_lemmas)
df8.insert(13, "antonym_lemmas_neg", df8.lemmas)

df8["lemmas_neg"] = df8.lemmas_neg.apply(lambda x: [f"not {lemma}" for lemma in x])
df8["antonym_lemmas_neg"] = df8.antonym_lemmas_neg.apply(lambda x: [f"not {lemma}" for lemma in x])

df8["lemmas"] = df8.lemmas_neg
df8["antonym_lemmas"] = df8.antonym_lemmas_neg

df8.drop(columns=["lemmas_neg", "antonym_lemmas_neg"], inplace=True)
df8.drop(columns=["lemmas_orig", "antonym_lemmas_orig"], inplace=True)

df8["TYPE"] = df8.TYPE + "_negLem"
df8["antonym_TYPE"] = df8.antonym_TYPE + "_negLem"


log0.info(f"{df8.shape = }")
disp_df(df8.head(n=22))
```

```
I: df8.shape = (872, 19)
```

```
            TYPE              id0      MEAN       STD  CNT
7    src0_negLem  oewn-01525056-v  0.400000  0.699206   10       [not wind off, not unwind
17   src0_negLem  oewn-00117793-v -1.375000  1.302470    8
34   src0_negLem  oewn-00088791-r -0.727273  1.009050   11              [not for good, not
41   src0_negLem  oewn-02639795-v -1.500000  1.080123   10
47   src0_negLem  oewn-00752852-a -0.333333  1.556998   12                  [not difficu
55   src0_negLem  oewn-00018622-a -1.200000  1.316561   10                         [no
56   src0_negLem  oewn-00153083-v  1.222222  0.971825    9       [not decrease, not lessen
60   src0_negLem  oewn-01244630-a  0.428571  1.133893    7                  [not pass,
68   src0_negLem  oewn-00244122-r  0.700000  0.948683   10                      [not u
74   src0_negLem  oewn-00688760-a -2.000000  0.707107    9                           [
76   src0_negLem  oewn-00074467-r -0.400000  0.699206   10                  [not forwar
94   src0_negLem  oewn-01101986-a  0.222222  0.440959    9                           [
```

```
106  src0_negLem  oewn-00905999-v  0.222222  0.440959    9                                    
110  src0_negLem  oewn-01408930-a  0.000000  0.000000    8                                   [
112  src0_negLem  oewn-00096162-r -0.500000  1.309307    8                                    
130  src0_negLem  oewn-01084999-v  1.272727  1.103713   11  [not quit, not give up, not dr
139  src0_negLem  oewn-02211988-v  0.909091  1.136182   11                                    
149  src0_negLem  oewn-00687584-v -1.222222  1.481366    9                                    
152  src0_negLem  oewn-02091950-a  0.444444  0.881917    9                             [not 
155  src0_negLem  oewn-00227509-r  1.111111  1.054093    9                                    
157  src0_negLem  oewn-01933854-a  1.000000  1.154701    7                                [n
160  src0_negLem  oewn-02337347-a -1.666667  1.118034    9                             [not i
```

## 4.20 Extract negated originals

```
cols8a = [
    "TYPE",
    "id0",
    "MEAN",
    "STD",
    "CNT",
    "lemmas",
    "definition",
    "examples",
    # "antonym_id0",
]
df8a = df8[cols8a].copy()
# WARN no drop here (one orig synset may get two different sets of lemmas from its antonyms)
# df8a.drop_duplicates(subset=["TYPE", "id0"], keep="first", inplace=True)

log0.info(f"{df8.shape = }")
log0.info(f"{df8a.shape = }")
disp_df(df8a.head(n=22))
```

```
I: df8.shape = (872, 19)
I: df8a.shape = (872, 8)
```

```
            TYPE              id0      MEAN       STD  CNT                                    
7    src0_negLem  oewn-01525056-v  0.400000  0.699206   10       [not wind off, not unwind
17   src0_negLem  oewn-00117793-v -1.375000  1.302470    8                                    
34   src0_negLem  oewn-00088791-r -0.727273  1.009050   11              [not for good, not
41   src0_negLem  oewn-02639795-v -1.500000  1.080123   10                                    
47   src0_negLem  oewn-00752852-a -0.333333  1.556998   12                    [not difficu
55   src0_negLem  oewn-00018622-a -1.200000  1.316561   10                              [no
56   src0_negLem  oewn-00153083-v  1.222222  0.971825    9       [not decrease, not lessen
60   src0_negLem  oewn-01244630-a  0.428571  1.133893    7                    [not pass,
68   src0_negLem  oewn-00244122-r  0.700000  0.948683   10                        [not u
74   src0_negLem  oewn-00688760-a -2.000000  0.707107    9                             [
76   src0_negLem  oewn-00074467-r -0.400000  0.699206   10                    [not forwar
94   src0_negLem  oewn-01101986-a  0.222222  0.440959    9                             [
106  src0_negLem  oewn-00905999-v  0.222222  0.440959    9                                    
110  src0_negLem  oewn-01408930-a  0.000000  0.000000    8                                   [
```

| | | | | | | |
|---|---|---|---|---|---|---|
| 112 | src0_negLem | oewn-00096162-r | -0.500000 | 1.309307 | 8 | |
| 130 | src0_negLem | oewn-01084999-v | 1.272727 | 1.103713 | 11 | [not quit, not give up, not dr |
| 139 | src0_negLem | oewn-02211988-v | 0.909091 | 1.136182 | 11 | |
| 149 | src0_negLem | oewn-00687584-v | -1.222222 | 1.481366 | 9 | |
| 152 | src0_negLem | oewn-02091950-a | 0.444444 | 0.881917 | 9 | [not |
| 155 | src0_negLem | oewn-00227509-r | 1.111111 | 1.054093 | 9 | |
| 157 | src0_negLem | oewn-01933854-a | 1.000000 | 1.154701 | 7 | [n |
| 160 | src0_negLem | oewn-02337347-a | -1.666667 | 1.118034 | 9 | [not i |

## 4.21 Extract unique augumented and unaugumented antonyms

```
cols8b = {
    "antonym_TYPE": "TYPE",
    "antonym_id0": "id0",
    "MEAN_antonym": "MEAN",
    "STD_antonym": "STD",
    "CNT_antonym": "CNT",
    "antonym_lemmas": "lemmas",
    "antonym_defs": "definition",
    "antonym_examples": "examples",
}
df8b = df8[cols8b.keys()].copy()
df8b.rename(columns=cols8b, inplace=True)
# df8b = df8b[df8b.TYPE=="ant2"]
df8b.drop_duplicates(subset="id0", inplace=True)

log0.info(f"{df8b.shape = }")
disp_df(df8b.head(n=22))
```

```
I: df8b.shape = (867, 8)
```

| | TYPE | id0 | MEAN | STD | CNT | |
|---|---|---|---|---|---|---|
| 7 | ant2_negLem | oewn-01526442-v | -0.400000 | NaN | NaN | [not twine, not wrap, not r |
| 17 | ant0_negLem | oewn-00109468-v | 0.727273 | 1.190874 | 11.0 | [not rest, not sta |
| 34 | ant2_negLem | oewn-00088404-r | 0.727273 | NaN | NaN | [no |
| 41 | ant2_negLem | oewn-02639021-v | 1.500000 | NaN | NaN | |
| 47 | ant0_negLem | oewn-00748528-a | -0.375000 | 0.744024 | 8.0 | |
| 55 | ant0_negLem | oewn-00017820-a | 0.500000 | 1.069045 | 8.0 | [not |
| 56 | ant2_negLem | oewn-00442400-v | -1.222222 | NaN | NaN | |
| 60 | ant0_negLem | oewn-01244800-a | 0.250000 | 0.462910 | 8.0 | |
| 68 | ant2_negLem | oewn-00244269-r | -0.700000 | NaN | NaN | [no |
| 74 | ant0_negLem | oewn-00687757-a | 2.375000 | 0.916125 | 8.0 | [n |
| 76 | ant0_negLem | oewn-00075708-r | 1.000000 | 0.707107 | 9.0 | [not back, |
| 94 | ant2_negLem | oewn-01102083-a | -0.222222 | NaN | NaN | |
| 106 | ant2_negLem | oewn-00908320-v | -0.222222 | NaN | NaN | [not discharge, not clear, no |
| 110 | ant2_negLem | oewn-01409235-a | -0.000000 | NaN | NaN | [not biological |
| 112 | ant0_negLem | oewn-00097310-r | 0.250000 | 0.707107 | 8.0 | |
| 130 | ant2_negLem | oewn-01085437-v | -1.272727 | NaN | NaN | [not enter, no |
| 139 | ant0_negLem | oewn-02247246-v | 0.888889 | 1.269296 | 9.0 | [not buy, |

```
149  ant0_negLem  oewn-00688348-v  0.142857  0.377964   7.0                                  
152  ant2_negLem  oewn-02092278-a -0.444444       NaN   NaN                                 [n
155  ant0_negLem  oewn-00227588-r -0.428571  0.786796   7.0                                  
157  ant2_negLem  oewn-01933678-a -1.000000       NaN   NaN             [not intellectual,   
160  ant2_negLem  oewn-02338328-a  1.666667       NaN   NaN                                 [no
```

## 4.22 Checkup before Concatenate

```
cols3a = [
    "TYPE",
    "id0",
    "MEAN",
    "STD",
    "CNT",
    "lemmas",
    "definition",
    "examples",
    # "antonym_id0",
]
df3a = df3[cols3a].copy()
df3a.drop_duplicates(subset=["TYPE", "id0"], inplace=True)
log0.info(f"{df3.shape = }")
log0.info(f"{df3a.shape = }")
disp_df(df3a.sample(n=4))
```

```
I: df3.shape = (6928, 14)
I: df3a.shape = (6914, 8)
```

```
      TYPE              id0      MEAN       STD  CNT              lemmas                
1636  src0  oewn-04210932-n  0.714286  0.755929    7          [shopping]  the commodities 
3132  src0  oewn-01709116-v  0.333333  0.866025    9   [reference, cite]                  
6408  src0  oewn-02748357-v -0.375000  0.744024    8              [mean]  have a specified
2825  src0  oewn-02274768-a  1.222222  1.201850    9  [resolved, solved]                  
```

## 4.23 Checkup before Concatenate

```
log0.info(f"{df7.shape = }")
disp_df(df7.sample(n=4))
```

```
I: df7.shape = (436, 8)
```

```
      TYPE              id0      MEAN  STD  CNT          lemmas                          
6891  ant2  oewn-00773049-a -0.571429  NaN  NaN       [inverse]  opposite in nature or eff
6630  ant2  oewn-00206071-r -1.555556  NaN  NaN  [inaccurately]                      in an
3346  ant2  oewn-02463135-a -0.125000  NaN  NaN     [trackless]                          
3113  ant2  oewn-02619707-v -0.750000  NaN  NaN          [miss]          fail to attend an
```

## 4.24 Checkup before Concatenate

```
log0.info(f"{df8a.shape = }")
disp_df(df8a.sample(n=4))
```

```
I: df8a.shape = (872, 8)
```

| | TYPE | id0 | MEAN | STD | CNT | |
|---|---|---|---|---|---|---|
| 2108 | src0_negLem | oewn-00066395-n | 0.875000 | 0.834523 | 8 | [not faili |
| 1140 | src0_negLem | oewn-01381145-a | 0.800000 | 0.788811 | 10 | [not |
| 5485 | src0_negLem | oewn-00429900-a | 0.625000 | 0.744024 | 8 | |
| 2837 | src0_negLem | oewn-02404473-v | -1.444444 | 1.013794 | 9 | [not elevate, not advance, no |

## 4.25 Checkup before Concatenate

```
log0.info(f"{df8b.shape = }")
disp_df(df8b.sample(n=4))
```

```
I: df8b.shape = (867, 8)
```

| | TYPE | id0 | MEAN | STD | CNT | |
|---|---|---|---|---|---|---|
| 668 | ant0_negLem | oewn-01871147-a | 2.000000 | 0.816497 | 7.0 | [not |
| 5822 | ant0_negLem | oewn-05104798-n | 0.500000 | 0.534522 | 8.0 | |
| 597 | ant2_negLem | oewn-01473917-a | 0.500000 | NaN | NaN | |
| 2296 | ant2_negLem | oewn-13983750-n | -0.285714 | NaN | NaN | [not truth, not trueness, not |

## 4.26 Final Checkup

```
log0.info(f"{df3a.shape = }")
log0.info(f"{df7.shape = }")
log0.info(f"{df8a.shape = }")
log0.info(f"{df8b.shape = }")
```

```
I: df3a.shape = (6914, 8)
I: df7.shape = (436, 8)
I: df8a.shape = (872, 8)
I: df8b.shape = (867, 8)
```

## 4.27 Merge

```
DF0 = pd.concat([df3a, df7, df8a, df8b], axis=0).reset_index(drop=True)
log0.info(f"{DF0.shape = }")
disp_df(DF0.sample(n=33).sort_index())
```

```
I: DF0.shape = (9089, 8)
```

| | TYPE | id0 | MEAN | STD | CNT | |
|---|---|---|---|---|---|---|
| 152 | src0 | oewn-02091950-a | 0.444444 | 0.881917 | 9.0 | |
| 364 | src0 | oewn-05925922-n | 0.625000 | 1.597990 | 8.0 | |
| 379 | src0 | oewn-01118003-v | -2.375000 | 1.060660 | 8.0 | [give |
| 415 | src0 | oewn-14508362-n | -0.272727 | 1.009050 | 11.0 | [hea |
| 960 | src0 | oewn-02722442-v | 0.500000 | 0.755929 | 8.0 | |
| 1351 | src0 | oewn-00200274-r | -1.100000 | 1.969207 | 10.0 | [obstinately, mulishly, obdu |
| 1391 | src0 | oewn-02367606-s | 0.333333 | 0.866025 | 9.0 | |
| 1397 | src0 | oewn-02488856-v | 0.666667 | 1.500000 | 9.0 | [execute |
| 2003 | src0 | oewn-01553629-s | 0.000000 | 1.732051 | 7.0 | [fast, degraded, debauched, |
| 2094 | src0 | oewn-01285124-a | 2.000000 | 0.577350 | 7.0 | |
| 2146 | src0 | oewn-01784999-v | -1.000000 | 2.000000 | 7.0 | |
| 2190 | src0 | oewn-02182456-a | -0.333333 | 0.516398 | 6.0 | |
| 2213 | src0 | oewn-00809380-s | 1.625000 | 0.916125 | 8.0 | [scintillating, effervescent |
| 2224 | src0 | oewn-01089130-s | 0.444444 | 1.013794 | 9.0 | [instinct, r |
| 2267 | src0 | oewn-05723230-n | 0.000000 | 0.000000 | 9.0 | [taste, gustatory perception |
| 2702 | src0 | oewn-01157556-v | 1.000000 | 1.290994 | 7.0 | [catch, overtake, |
| 3105 | src0 | oewn-02605633-v | 2.125000 | 1.125992 | 8.0 | |
| 3283 | src0 | oewn-01850926-s | 1.666667 | 1.000000 | 9.0 | |
| 3309 | src0 | oewn-01115009-v | 0.750000 | 1.035098 | 8.0 | |
| 4737 | src0 | oewn-01674975-s | 0.428571 | 0.534522 | 7.0 | [incorpora |
| 5448 | src0 | oewn-01340439-a | 0.500000 | 0.925820 | 8.0 | |
| 5658 | src0 | oewn-00988799-a | 0.750000 | 0.886405 | 8.0 | [exacti |
| 5721 | src0 | oewn-14322572-n | 0.500000 | 1.069045 | 8.0 | [twist |
| 5889 | src0 | oewn-05662964-n | 0.250000 | 0.462910 | 8.0 | [touch, cutaneous senses, sk |
| 6439 | src0 | oewn-01650408-v | 0.500000 | 1.414214 | 8.0 | [kick up, evoke, provo |
| 6551 | src0 | oewn-15261656-n | 0.000000 | 0.000000 | 8.0 | |
| 6727 | src0 | oewn-00191991-n | 0.111111 | 0.600925 | 9.0 | |
| 6915 | ant2 | oewn-00088404-r | 0.727273 | NaN | NaN | [for goo |
| 6936 | ant2 | oewn-00356369-a | -0.636364 | NaN | NaN | |
| 7106 | ant2 | oewn-00370277-r | -1.375000 | NaN | NaN | [imprecis |
| 8221 | src0_negLem | oewn-04756076-n | 0.111111 | 0.333333 | 9.0 | |
| 8309 | ant0_negLem | oewn-00019801-r | 0.333333 | 0.500000 | 9.0 | [not ne |
| 8572 | ant0_negLem | oewn-14498302-n | -1.500000 | 1.309307 | 8.0 | |

## 4.28 Save DF0

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

dir0 = "../../data/d0009_synsets-merged/"
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 = ["synsets-merged"]
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)

log0.info("DONE")
```

```
I: saving: ../../data/d0009_synsets-merged/synsets-merged.pkl...
I: saving: ../../data/d0009_synsets-merged/synsets-merged.csv...
I: saving: ../../data/d0009_synsets-merged/synsets-merged.xlsx...
I: DONE
```

# 5 Checkup

```
dir0 = pathlib.Path().home()/"cc/dev/c2023a/b0001_c0418_bertagen_with_prodigy_and_spacy/data/input-0
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"temp_eval_checkup_20220420T024438_temp"
extn0 = ".pkl"

if0 = (dir0/name0).with_suffix(extn0)
log0.info(f"loading: {if0}...")
df0 = pd.read_pickle(if0)
log0.info(f"loading: {if0}... DONE")

log0.info(f"{df0.shape = }")
disp_df(df0.sample(n=8).sort_index())
```

```
I: df0.shape = (132, 10)
```

| | text | m |
|---|---|---|
| 1 | not energetic, deficient in alertness or... | {'id0': 'oewn-00879020-a', 'id4': 'oewn- |
| 10 | heavy work | {'id0': 'oewn-00840083-s', 'id4': 'oewn- |
| 51 | indecisively, lacking firmness or resolu... | {'id0': 'oewn-00300164-r', 'id4': 'oewn- |

| | | |
|---|---|---|
| 78 | she fell into a rage and refused to answer | {'id0': 'oewn-14060217-n', 'id4': 'oewn- |
| 94 | prehensile employers stingy with raises ... | {'id0': 'oewn-00030046-s', 'id4': 'oewn- |
| 108 | unpersuasiveness, inability to persuade | {'id0': 'oewn-05212488-n', 'id4': 'oewn- |
| 116 | she is useless in an emergency | {'id0': 'oewn-02506940-a', 'id4': 'oewn- |
| 124 | after some hesitation he agreed | {'id0': 'oewn-04653300-n', 'id4': 'oewn- |