Lowercase and Merge

1 Data sources

1.1 prod: DWC-Based Frameworks

- Pennebaker et al. (2015)
  - The General Inquirer (GI) data should be obtained from https://inquirer.sites.fas.harvard.edu/.
- Stone et al. (1966)
  - The Linguistic Inquiry and Word Count (LIWC) data should be obtained from https://www.liwc.app/.

1.2 prod: Full-Fledged Tools

- Pietraszkiewicz et al. (2019)
  - `PietraszkiewiczEtAl2019a_agency_A.yaml`
  - `PietraszkiewiczEtAl2019a_agency_B.yaml`
  - `PietraszkiewiczEtAl2019a_agency_C.yaml`
- Nicolas et al. (2021)
  - `NicolasEtAl2019a_full_dict_ability_negat.yaml`
  - `NicolasEtAl2019a_full_dict_ability_posit.yaml`
  - `NicolasEtAl2019a_full_dict_agency_negat.yaml`
  - `NicolasEtAl2019a_full_dict_agency_posit.yaml`

1.3 prod: Related Word Collections and Composite Dictionaries

- Abele et al. (2008a) (Abele et al., 2008b)
  - `AbeleEtAl2008a_list_agency_negat.yaml`
  - `AbeleEtAl2008a_list_agency_posit.yaml`
- Gaucher et al. (2011)

    - GaucherEtAl2011a_list_masculinity_pertaining.yaml

- Hart et al. (2011)
    - HartEtAl2011a_agency_dict_negat.yaml
    - HartEtAl2011a_agency_dict_posat.yaml
- Decter-Frain and Frimer (2016)
    - DecterFrainFrimer2016a_selected_liwc_cats_agency_negat.yaml
    - DecterFrainFrimer2016a_selected_liwc_cats_agency_posit.yaml
- Madera et al. (2009)
    - MaderaEtAl2009a_madera_agentic_adj.yaml
    - MaderaEtAl2009a_madera_agentic_orient_liwc.yaml
- Pietraszkiewicz et al. (2019)
    - PietraszkiewiczEtAl2019a_selected_liwc_cats.yaml
- Nicolas et al. (2021)
    - NicolasEtAl2019a_full_dict_status_negat.yaml
    - NicolasEtAl2019a_full_dict_status_posit.yaml
    - NicolasEtAl2019a_seed_dict_ability_negat.yaml
    - NicolasEtAl2019a_seed_dict_ability_posit.yaml
    - NicolasEtAl2019a_seed_dict_agency_negat.yaml
    - NicolasEtAl2019a_seed_dict_agency_posit.yaml
    - NicolasEtAl2019a_seed_dict_status_negat.yaml
    - NicolasEtAl2019a_seed_dict_status_posit.yaml

2 Boilerplate

3 Imports

3.1 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
```

3.2 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")
```

3.3 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")
```

4 Extra Imports

4.1 prod: More extra imports and settings

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

5 Process

5.1 prod: Load, process, merge and save

```
dir0 = "../../data/d0001_sources/"
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!"

glob0 = dir0.glob("s*/c*/*.yaml")
glob0 = sorted(list(glob0))

log0.info(f"{len(glob0) = }")

data4 = []
for if0 in glob0:
    of2 = pathlib.Path("../../data/d0002_sources-lowercased")/if0.relative_to("../../data/d0001_sou
    of2.parent.mkdir(mode=0o700, parents=True, exist_ok=True)
    data0 = srsly.read_yaml(if0)
    # log0.info(f"r: {if0} ({len(data0)})")
    data2 = [clean_str(item).lower() for item in data0]
    data2 = list(filter(re0.search, data2))
    data2 = sorted(list(set(data2)))
    data4 += data2
    log0.info(f"w: {of2} ({len(data2)}/{len(data0)} [{len(set(data0))}])")
    srsly.write_yaml(of2, data2)

of4 = pathlib.Path("../../data/d0004_sources-merged-and-deduped")/"merged.yaml"
of4.parent.mkdir(mode=0o700, parents=True, exist_ok=True)
log0.info(f"M: {of4} ({len(data4)})")
data4 = sorted(list(set(data4)))
log0.info(f"W: {of4} ({len(data4)})")
srsly.write_yaml(of4, data4)
```

```
I: len(glob0) = 25
I: w: ../../data/d0002_sources-lowercased/s0002_full-fledged-tools/c2019_PietraszkiewiczEt
I: w: ../../data/d0002_sources-lowercased/s0002_full-fledged-tools/c2019_PietraszkiewiczEt
I: w: ../../data/d0002_sources-lowercased/s0002_full-fledged-tools/c2019_PietraszkiewiczEt
I: w: ../../data/d0002_sources-lowercased/s0002_full-fledged-tools/c2021_NicolasEtAl2021a/
I: w: ../../data/d0002_sources-lowercased/s0002_full-fledged-tools/c2021_NicolasEtAl2021a/
I: w: ../../data/d0002_sources-lowercased/s0002_full-fledged-tools/c2021_NicolasEtAl2021a/
I: w: ../../data/d0002_sources-lowercased/s0002_full-fledged-tools/c2021_NicolasEtAl2021a/
I: w: ../../data/d0002_sources-lowercased/s0003_related-word-collections-and-composite-dic
```

```
I: w: ../../data/d0002_sources-lowercased/s0003_related-word-collections-and-composite-dic
I: w: ../../data/d0002_sources-lowercased/s0003_related-word-collections-and-composite-dic
I: w: ../../data/d0002_sources-lowercased/s0003_related-word-collections-and-composite-dic
I: w: ../../data/d0002_sources-lowercased/s0003_related-word-collections-and-composite-dic
I: w: ../../data/d0002_sources-lowercased/s0003_related-word-collections-and-composite-dic
I: w: ../../data/d0002_sources-lowercased/s0003_related-word-collections-and-composite-dic
I: w: ../../data/d0002_sources-lowercased/s0003_related-word-collections-and-composite-dic
I: w: ../../data/d0002_sources-lowercased/s0003_related-word-collections-and-composite-dic
I: w: ../../data/d0002_sources-lowercased/s0003_related-word-collections-and-composite-dic
I: w: ../../data/d0002_sources-lowercased/s0003_related-word-collections-and-composite-dic
I: w: ../../data/d0002_sources-lowercased/s0003_related-word-collections-and-composite-dic
I: w: ../../data/d0002_sources-lowercased/s0003_related-word-collections-and-composite-dic
I: w: ../../data/d0002_sources-lowercased/s0003_related-word-collections-and-composite-dic
I: w: ../../data/d0002_sources-lowercased/s0003_related-word-collections-and-composite-dic
I: w: ../../data/d0002_sources-lowercased/s0003_related-word-collections-and-composite-dic
I: w: ../../data/d0002_sources-lowercased/s0003_related-word-collections-and-composite-dic
I: w: ../../data/d0002_sources-lowercased/s0003_related-word-collections-and-composite-dic
I: M: ../../data/d0004_sources-merged-and-deduped/merged.yaml (7711)
I: W: ../../data/d0004_sources-merged-and-deduped/merged.yaml (4618)
```

6 Checkup

```
data0a = srsly.read_yaml("../../data/d0001_sources/s0003_related-word-collections-and-composite-dict
data0b = srsly.read_yaml("../../data/d0001_sources/s0003_related-word-collections-and-composite-dict
data0c = list(set(data0a+data0b))
log0.info(f"{len(data0c) = }")
```

```
I: len(data0c) = 1672
```

References

Abele, A. E., Uchronski, M., Suitner, C., & Wojciszke, B. (2008a). Towards an operationalization of the fundamental dimensions of agency and communion: Trait content ratings in five countries considering valence and frequency of word occurrence. *European Journal of Social Psychology*, *38*(7), 1202–1217. https://doi.org/10.1002/ejsp.575 (cit. on p. 1)

Abele, A. E., Uchronski, M., Suitner, C., & Wojciszke, B. (2008b). Towards an operationalization of the fundamental dimensions of agency and communion: Trait content ratings in five countries considering valence and frequency of word occurrence. *European Journal of Social Psychology*, *38*(7), 1202–1217. https://doi.org/10.1002/ejsp.575 (cit. on p. 1)

Decter-Frain, A., & Frimer, J. A. (2016). Impressive words: Linguistic predictors of public approval of the U.S. Congress. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00240 (cit. on p. 2)

Gaucher, D., Friesen, J., & Kay, A. C. (2011). Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of Personality and Social Psychology*, *101*(1), 109–128. https://doi.org/10.1037/a0022530 (cit. on p. 1)

Hart, C. M., Sedikides, C., Wildschut, T., Arndt, J., Routledge, C., & Vingerhoets, A. J. (2011). Nostalgic recollections of high and low narcissists. *Journal of Research in Personality*, *45*(2), 238–242. https://doi.org/10.1016/j.jrp.2011.01.002 (cit. on p. 2)

Madera, J. M., Hebl, M. R., & Martin, R. C. (2009). Gender and letters of recommendation for academia: Agentic and communal differences. *Journal of Applied Psychology*, *94*(6), 1591–1599. https://doi.org/10.1037/a0016539 (cit. on p. 2)

Nicolas, G., Bai, X., & Fiske, S. T. (2021). Comprehensive stereotype content dictionaries using a semi-automated method. *European Journal of Social Psychology*, ejsp.2724. https://doi.org/10.1002/ejsp.2724 (cit. on pp. 1–2)

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015. https://doi.org/10.15781/T29G6Z (cit. on p. 1)

Pietraszkiewicz, A., Formanowicz, M., Gustafsson Sendén, M., Boyd, R. L., Sikström, S., & Sczesny, S. (2019). The big two dictionaries: Capturing agency and communion in natural language. *European Journal of Social Psychology*, *49*(5), 871–887. https://doi.org/10.1002/ejsp.2561 (cit. on pp. 1–2)

Stone, P. J., Dunphy, D., Smith, M. S., & Ogilvie, D. M. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press. (Cit. on p. 1).