S5001_checkups

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 prod: Load data

```
dir0 = "../../data/d0010_training-data/"
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!"

extn0 = ".pkl"

name0 = f"ft0x"
if0 = (dir0/name0).with_suffix(extn0)
log0.info(f"loading: {if0}...")
ft0x = pd.read_pickle(if0)

name0 = f"ft1x"
```

```
if0 = (dir0/name0).with_suffix(extn0)
log0.info(f"loading: {if0}...")
ft1x = pd.read_pickle(if0)

name0 = f"ft2x"
if0 = (dir0/name0).with_suffix(extn0)
log0.info(f"loading: {if0}...")
ft2x = pd.read_pickle(if0)

name0 = f"ft3x"
if0 = (dir0/name0).with_suffix(extn0)
log0.info(f"loading: {if0}...")
ft3x = pd.read_pickle(if0)

name0 = f"ft4x"
if0 = (dir0/name0).with_suffix(extn0)
log0.info(f"loading: {if0}...")
ft4x = pd.read_pickle(if0)

name0 = f"gs0x"
if0 = (dir0/name0).with_suffix(extn0)
log0.info(f"loading: {if0}...")
gs0x = pd.read_pickle(if0)

for df in [ft0x, ft1x, ft2x, ft3x, ft4x, gs0x]:
    assert df.target.max() <= 1
    assert df.target.min() >= -1

log0.info(f"{ft0x.columns = }")
log0.info(f"{ft1x.columns = }")
log0.info(f"{ft2x.columns = }")
log0.info(f"{ft3x.columns = }")
log0.info(f"{ft4x.columns = }")
log0.info(f"{gs0x.columns = }")

log0.info(f"{ft0x.shape = }")
log0.info(f"{ft1x.shape = }")
log0.info(f"{ft2x.shape = }")
log0.info(f"{ft3x.shape = }")
log0.info(f"{ft4x.shape = }")
log0.info(f"{gs0x.shape = }")
disp_df(ft2x.sample(n=8).sort_index())
```

```
I: loading: ../../data/d0010_training-data/ft0x.pkl...
I: loading: ../../data/d0010_training-data/ft1x.pkl...
I: loading: ../../data/d0010_training-data/ft2x.pkl...
I: loading: ../../data/d0010_training-data/ft3x.pkl...
I: loading: ../../data/d0010_training-data/ft4x.pkl...
I: loading: ../../data/d0010_training-data/gs0x.pkl...
I: ft0x.columns = Index(['text', 'target'], dtype='object')
I: ft1x.columns = Index(['text', 'target'], dtype='object')
I: ft2x.columns = Index(['text', 'target'], dtype='object')
```

```
I: ft3x.columns = Index(['text', 'target'], dtype='object')
I: ft4x.columns = Index(['text', 'target'], dtype='object')
I: gs0x.columns = Index(['text', 'target'], dtype='object')
I: ft0x.shape = (12812, 2)
I: ft1x.shape = (27625, 2)
I: ft2x.shape = (35801, 2)
I: ft3x.shape = (59803, 2)
I: ft4x.shape = (4310, 2)
I: gs0x.shape = (300, 2)
```

	text	target
328	avaricious, immoderately desirous of acq...	-0.083333
637	line up, get something or somebody for a...	0.333333
8887	pressurize, increase the pressure in or of	0.208333
10624	intellectually, of or relating to the in...	0.166667
14343	altogether, with everything included or ...	0.000000
21133	a complex set of variations based on a s...	-0.047619
24629	The race car driver lived through severa...	0.407407
30987	a moderate grade of intelligence	0.250000

4.2 Checkups

```
log0.info(f"{ft0x.target.mean() = }")
log0.info(f"{ft0x.target.std() = }")
log0.info(f"{ft0x.target.min() = }")
log0.info(f"{ft0x.target.max() = }")

log0.info(f"{ft1x.target.mean() = }")
log0.info(f"{ft1x.target.std() = }")
log0.info(f"{ft1x.target.min() = }")
log0.info(f"{ft1x.target.max() = }")

log0.info(f"{ft2x.target.mean() = }")
log0.info(f"{ft2x.target.std() = }")
log0.info(f"{ft2x.target.min() = }")
log0.info(f"{ft2x.target.max() = }")

log0.info(f"{ft3x.target.mean() = }")
log0.info(f"{ft3x.target.std() = }")
log0.info(f"{ft3x.target.min() = }")
log0.info(f"{ft3x.target.max() = }")

log0.info(f"{ft4x.target.mean() = }")
log0.info(f"{ft4x.target.std() = }")
log0.info(f"{ft4x.target.min() = }")
log0.info(f"{ft4x.target.max() = }")

log0.info(f"{gs0x.target.mean() = }")
log0.info(f"{gs0x.target.std() = }")
log0.info(f"{gs0x.target.min() = }")
log0.info(f"{gs0x.target.max() = }")
```

```
for df in [ft0x, ft1x, ft2x, ft3x, ft4x]:
    assert df.target.max() <= 1
    assert df.target.min() >= -1
```

```
I: ft0x.target.mean() = 0.08603202685313208
I: ft0x.target.std() = 0.3206837913472018
I: ft0x.target.min() = -1.0
I: ft0x.target.max() = 0.9259259259259259
I: ft1x.target.mean() = 0.08615285368226545
I: ft1x.target.std() = 0.3224558686569525
I: ft1x.target.min() = -1.0
I: ft1x.target.max() = 0.9259259259259259
I: ft2x.target.mean() = 0.06165959521661878
I: ft2x.target.std() = 0.33469121140295915
I: ft2x.target.min() = -1.0
I: ft2x.target.max() = 0.9583333333333334
I: ft3x.target.mean() = 0.059020742241901715
I: ft3x.target.std() = 0.34168407560227754
I: ft3x.target.min() = -1.0
I: ft3x.target.max() = 0.9583333333333334
I: ft4x.target.mean() = 0.1367782180439041
I: ft4x.target.std() = 0.5234643011136165
I: ft4x.target.min() = -1.0
I: ft4x.target.max() = 0.9851851851866668
I: gs0x.target.mean() = 0.041393770856507243
I: gs0x.target.std() = 0.512430411303803
I: gs0x.target.min() = -0.8444444444444444
I: gs0x.target.max() = 0.9333333333333332
```

4.3 Check tokenization 1

```
import torch
from transformers import AutoModelForSequenceClassification
from transformers import AutoTokenizer
from transformers import Trainer
from transformers import TrainingArguments
from transformers import EarlyStoppingCallback

df0_temp = ft3x.copy()
df0_temp = ft2x.copy()
df0_temp = pd.concat([ft0x, ft1x, ft2x, ft3x, ft4x, gs0x])

log0.info(f"{df0_temp.shape = }")

df0_temp = df0_temp.drop_duplicates(subset=["text"], keep="first")
```

```
log0.info(f"{df0_temp.shape = }")
disp_df(df0_temp)
```

```
I: df0_temp.shape = (140651, 2)
I: df0_temp.shape = (72116, 2)
```

	text	target
0	he waited impatiently in the blind	-0.090909
1	The convicted murderer escaped from a hi...	0.583333
2	a weak mind	-0.875000
3	Murdoch owns many newspapers	0.000000
4	the coach told his players that defeat w...	-0.333333
5	Common sense is not so common	0.185185
5	he hasn't got the sense God gave little ...	0.185185
5	fortunately she had the good sense to ru...	0.185185
6	the reconciliation of his checkbook and ...	0.100000
7	roll your hair around your finger	0.133333
7	Twine the thread around the spool	0.133333
7	She wrapped her arms around the child	0.133333
8	This beggars description!	-0.416667
9	logically, you should now do the same to...	0.166667
10	instructional designers are trained in s...	0.222222
10	the CIA chief of station accepted respon...	0.222222
..	...	...
284	It could be a confidence thing, or the f...	-0.483871
285	and we'd sometimes suggest things to eac...	0.388889
286	The Judo team trained 2 days a week on M...	0.288889
287	She is lazy.	-0.724138
288	I have failed to find success, or a nich...	-0.666667
289	I was so proud of myself, and it helped ...	0.784946
290	For example, his unsuccessful streak app...	-0.611111
291	I continued to work for the company and ...	0.133333
292	I lost touch with him as I moved away fr...	0.053763
293	All this was now 20 years ago and i am p...	0.622222
294	I started to become more active during t...	0.688889
295	Samara started working at her personal t...	0.377778
296	I knew they were around the house somewh...	-0.678161
297	This individual would always sit in our ...	-0.777778
298	I eventually dropped out of college.	-0.711111
299	I have lost friends, partners and countl...	-0.744444

[72116 rows x 2 columns]

4.4 Check tokenization 2

```
log0.debug(f"{df0_temp.shape = }")
df0_temp.drop_duplicates(keep="first", inplace=True, ignore_index=True)
log0.debug(f"{df0_temp.shape = }")
```

```
tokenizer = AutoTokenizer.from_pretrained("bert-base-uncased")
tokenized = df0_temp["text"].apply(
    (lambda x: tokenizer.encode(x, add_special_tokens=True))
)
```

4.5 Check tokenization 3

```
type(tokenized)
df0_temp["tok"] = tokenized
df0_temp["num"] = df0_temp.tok.apply(len)
df0_temp.num.hist()
log0.info(f"{df0_temp.num.max()}")
disp_df(df0_temp)
```

```
I: 87
```

	text	target	
0	he waited impatiently in the blind	-0.090909	[101, 2002, 4741, 19951, 199
1	The convicted murderer escaped from a hi...	0.583333	[101, 1996, 7979, 13422, 637
2	a weak mind	-0.875000	[101, 1037, 5
3	Murdoch owns many newspapers	0.000000	[101, 19954, 8617, 2
4	the coach told his players that defeat w...	-0.333333	[101, 1996, 2873, 2409, 2010
5	Common sense is not so common	0.185185	[101, 2691, 3168, 2003, 2025
6	he hasn't got the sense God gave little ...	0.185185	[101, 2002, 8440, 1005, 1056
7	fortunately she had the good sense to ru...	0.185185	[101, 14599, 2016, 2018, 199
8	the reconciliation of his checkbook and ...	0.100000	[101, 1996, 16088, 1997, 201
9	roll your hair around your finger	0.133333	[101, 4897, 2115, 2606, 2105
10	Twine the thread around the spool	0.133333	[101, 5519, 2063, 1996, 1168
11	She wrapped her arms around the child	0.133333	[101, 2016, 5058, 2014, 2608
12	This beggars description!	-0.416667	[101, 2023, 11693, 6843, 201
13	logically, you should now do the same to...	0.166667	[101, 11177, 2135, 1010, 201
14	instructional designers are trained in s...	0.222222	[101, 23219, 11216, 2024, 47
15	the CIA chief of station accepted respon...	0.222222	[101, 1996, 9915, 2708, 1997
...	...	...	
72100	It could be a confidence thing, or the f...	-0.483871	[101, 2009, 2071, 2022, 1037
72101	and we'd sometimes suggest things to eac...	0.388889	[101, 1998, 2057, 1005, 1040
72102	The Judo team trained 2 days a week on M...	0.288889	[101, 1996, 19083, 2136, 473
72103	She is lazy.	-0.724138	[101, 2016, 2003, 13
72104	I have failed to find success, or a nich...	-0.666667	[101, 1045, 2031, 3478, 2000
72105	I was so proud of myself, and it helped ...	0.784946	[101, 1045, 2001, 2061, 7098
72106	For example, his unsuccessful streak app...	-0.611111	[101, 2005, 2742, 1010, 2010
72107	I continued to work for the company and ...	0.133333	[101, 1045, 2506, 2000, 2147
72108	I lost touch with him as I moved away fr...	0.053763	[101, 1045, 2439, 3543, 2007
72109	All this was now 20 years ago and i am p...	0.622222	[101, 2035, 2023, 2001, 2085
72110	I started to become more active during t...	0.688889	[101, 1045, 2318, 2000, 2468
72111	Samara started working at her personal t...	0.377778	[101, 3520, 5400, 2318, 2551
72112	I knew they were around the house somewh...	-0.678161	[101, 1045, 2354, 2027, 2020
72113	This individual would always sit in our ...	-0.777778	[101, 2023, 3265, 2052, 2467

```
72114         I eventually dropped out of college. -0.711111  [101, 1045, 2776, 3333, 2041
72115  I have lost friends, partners and countl... -0.744444  [101, 1045, 2031, 2439, 2814

[72116 rows x 4 columns]
```