

In [12]:

```
"""
=====
|      Author  of      the      Project  -      Engineer      Nicola  Lombardi      |
=====
|      MSc.    in      Telecommunications      and      Internet      of      Things  |
|      Telecommunications Engineering      |      SW  HW  ENGINEER      |

|Professional  Profile
|      it.linkedin.com/in/nicola-lombardi-09046b205

Example to demonstrate: Data Cleaning:
=====> Remove noise and correct inconsistencies in the data.
      https://www.researchgate.net/publication/382851188_A_whole_method_to_do_Data_Analysis

# https://www.geeksforgeeks.org/deque-in-python/

# Dimostrato che median_filter.py  is an invariant for length

Input DataFrame:
   timestamp      tag_id  angle  tag_height
0  1733062840000  4baf351178aa9b0e   -30         1.2
1  1733062840100  4baf351178aa9b0e   -28         1.2
2  1733062840200  4baf351178aa9b0e   -39         1.2
3  1733062840300  4baf351178aa9b0e   -27         1.2
4  1733062840400  4baf351178aa9b0e    -4         1.2  Anomalous Sample ... ==> Data Cleaning

#####

                        Test      Number 1      Preview
#####

Filtered DataFrame:
   timestamp      tag_id  angle  tag_height
0  1733062840000  4baf351178aa9b0e  -30.0         1.2
1  1733062840100  4baf351178aa9b0e  -29.0         1.2
2  1733062840200  4baf351178aa9b0e  -30.0         1.2
3  1733062840300  4baf351178aa9b0e  -29.0         1.2
4  1733062840400  4baf351178aa9b0e  -28.0         1.2

"""

import sys
import time
from constant import *
import pandas as pd
import statistics
from collections import defaultdict, deque
import datetime
from utils_subroutines import *

clock_string = f"[{datetime.datetime.now()}]"

#      =====
#      ----- DataFrame      test      -----
#      =====

def my_test():
    """
    :documentation:

    My test to completely understand the median_filter logic and the difference in terms of samples
    between input and output

    :block diagram:

    input.csv ----> |      median _ filter . py      | -----> Output.csv
                    |_____|

    :return: NONE
    """
    df = pd.read_csv("input.csv", header=None,
                    names=["timestamp", "tag_id", "angle", "tag_height"]) # Features Naming
    print(f"{clock_string}Input DataFrame PREVIEWER :")
    print(df.head(len(df)))
    count = 0
    print(header.format(count + 1))

    # filter call (median_filter.py)
    filtered_df = median_filter_df_version(df)
    print(f"\n{clock_string}Filtered DataFrame:")
```

```

print(filtered_df.head(len(filtered_df)))
print(f"{sep}\n {clock_string} Check of the Length:\n Input csv DF: {len(df)}\n Output csv DF: {len(filtered_df)}")

# Save output.csv from DataFrame
filtered_df.to_csv("output.csv", index=False, header=False)

```

```

def median_filter_df_version(df, time_window=1000):
    """

```

```

:param df: dataframe in which apply Data Cleaning phase
:param time_window: time window size of the tolerance in seconds [s]
:return: new filtered dataframe with same size in terms of cols and rows

```

```

:DOCUMENTATION: defaultdict

```

```

1. defaultdict – why use it

```

The defaultdict is a subclass of dict that allows you to avoid errors when accessing keys that are not yet present. It automatically creates a default value. It is especially useful when you need to accumulate values in lists or dictionaries.

```

from collections import defaultdict

```

```

d = defaultdict(list)
d['a'].append(1)
d['a'].append(2)
d['b'].append(3)
print(d)
# Output: defaultdict(<class 'list'>, {'a': [1, 2], 'b': [3]})

```

```

:DOCUMENTATION: deque

```

```

from collections import defaultdict, deque
from datetime import datetime, timedelta

```

```

window = timedelta(seconds=60)
tag_queue = defaultdict(deque)

```

```

for row in dataframe.itertuples():
    tag = row.tag # ad esempio
    timestamp = row.timestamp

    queue = tag_queue[tag]

    # Elimina tutti i vecchi timestamp fuori dalla finestra
    while queue and (timestamp - queue[0] > window):
        queue.popleft()

    queue.append(timestamp)

    # Ora queue contiene solo i timestamp "validi" entro la finestra per quel tag
    if len(queue) > 1:
        print(f"{len(queue)} eventi recenti per il tag {tag}")

```

Median_filter algorithm:

tag_queue is a dictionary that maps each tag to a deque of timestamps.

For each row in the DataFrame, you update the queue corresponding to the tag.

Remove timestamps that are too old (outside the tolerance window).

After cleanup, you can see how many recent events are still valid for that tag.

```

"""
data_store = defaultdict(deque)
filtered_data = []
print(f"SETUP: \n\t data_store status -> {data_store} \n\t data_filter status -> {filtered_data}\n\n\n")

```

```

print("[Data Processing]: starting the Bluetooth Beacon Acquisition...")

```

```

test_number = 1

```

```

for index, row in df.iterrows():
    print("-----")
    print(f"{clock_string}Test number {test_number} - index = {index}\n")
    timestamp, tag_id, angle, tag_height = row
    print(f"{clock_string}Received data from packet:")
    print(f"{clock_string}TIME STAMP : {timestamp}")
    print(f"{clock_string}TAG id : {tag_id}")
    print(f"{clock_string}ANGLE : {angle}")
    print(f"{clock_string}HEIGHT : {tag_height}")
    print("-----")
    # Remove outdated entries
    # Clear all old timestamps out of the window
    tag_queue = data_store[tag_id]
    print(f"{clock_string}[Data Processing]: push in the tag_queue...")
    while tag_queue and tag_queue[0][0] < timestamp - time_window:
        print(f"{clock_string}[Data Processing]: TAG is not valid and pop from tag_queue...")
        tag_queue.popleft()

```

```

# Add new entry
tag_queue.append((timestamp, angle))
print(f"{clock_string}[Data Processing]: TAG is valid and push in tag_queue...")

# Compute median
median_angle = statistics.median(pos for _, pos in tag_queue)

# Store filtered data
filtered_data.append([timestamp, tag_id, median_angle, tag_height])
test_number += 1 # upgrade the test
print(f"{sep}{clock_string}[Data Processing]: TAG queue print:\n",tag_queue)
print(f"Number of valid samples: {len(tag_queue)}")
print(f"Percentage of valid samples: {round(len(tag_queue) / test_number * 100,2)}%")

return pd.DataFrame(filtered_data, columns=df.columns)

if __name__ == "__main__":
    # call the optional test of median_filter.py
    my_test()

```

[2025-04-13 14:04:57.225806]Input DataFrame PREVIEWER :

	timestamp	tag_id	angle	tag_height
0	1733062840000	4baf351178aa9b0e	-30	1.2
1	1733062840100	4baf351178aa9b0e	-28	1.2
2	1733062840200	4baf351178aa9b0e	-39	1.2
3	1733062840300	4baf351178aa9b0e	-27	1.2
4	1733062840400	4baf351178aa9b0e	-4	1.2
5	1733062840500	4baf351178aa9b0e	-25	1.2
6	1733062840600	4baf351178aa9b0e	-20	1.2
7	1733062840700	4baf351178aa9b0e	25	1.2
8	1733062840800	4baf351178aa9b0e	15	1.2
9	1733062840900	4baf351178aa9b0e	-2	1.2
10	1733062841000	4baf351178aa9b0e	-25	1.2
11	1733062841100	4baf351178aa9b0e	5	1.2
12	1733062841200	4baf351178aa9b0e	12	1.2
13	1733062841300	4baf351178aa9b0e	21	1.2
14	1733062841400	4baf351178aa9b0e	45	1.2
15	1733062841500	4baf351178aa9b0e	28	1.2
16	1733062841600	4baf351178aa9b0e	33	1.2
17	1733062841700	4baf351178aa9b0e	-65	1.2
18	1733062841800	4baf351178aa9b0e	32	1.2
19	1733062841900	4baf351178aa9b0e	25	1.2
20	1733062842000	4baf351178aa9b0e	5	1.2
21	1733062840000	a1b2c3d4e5f6g7h8	-1	1.1
22	1733062840100	a1b2c3d4e5f6g7h8	17	1.1
23	1733062840200	a1b2c3d4e5f6g7h8	16	1.1
24	1733062840300	a1b2c3d4e5f6g7h8	-8	1.1
25	1733062840400	a1b2c3d4e5f6g7h8	19	1.1
26	1733062840500	a1b2c3d4e5f6g7h8	20	1.1
27	1733062840600	a1b2c3d4e5f6g7h8	5	1.1
28	1733062840700	a1b2c3d4e5f6g7h8	15	1.1
29	1733062840800	a1b2c3d4e5f6g7h8	17	1.1
30	1733062840900	a1b2c3d4e5f6g7h8	6	1.1
31	1733062841000	a1b2c3d4e5f6g7h8	18	1.1
32	1733062841100	a1b2c3d4e5f6g7h8	19	1.1
33	1733062841200	a1b2c3d4e5f6g7h8	20	1.1
34	1733062841300	a1b2c3d4e5f6g7h8	17	1.1
35	1733062841400	a1b2c3d4e5f6g7h8	16	1.1
36	1733062841500	a1b2c3d4e5f6g7h8	-18	1.1
37	1733062841600	a1b2c3d4e5f6g7h8	19	1.1
38	1733062841700	a1b2c3d4e5f6g7h8	20	1.1
39	1733062841800	a1b2c3d4e5f6g7h8	30	1.1
40	1733062841900	a1b2c3d4e5f6g7h8	0	1.1
41	1733062842000	a1b2c3d4e5f6g7h8	0	1.1

#####

Test	Number	1	Preview
#####			

SETUP:

```

data_store status -> defaultdict(<class 'collections.deque'>, {})
data_filter status -> []

```

[Data Processing]: starting the Bluetooth Beacon Acquisition...

[2025-04-13 14:04:57.225806]Test number 1 - index = 0

```

[2025-04-13 14:04:57.225806]Received data from packet:
[2025-04-13 14:04:57.225806]TIME STAMP : 1733062840000
[2025-04-13 14:04:57.225806]TAG id : 4baf351178aa9b0e
[2025-04-13 14:04:57.225806]ANGLE : -30
[2025-04-13 14:04:57.225806]HEIGHT : 1.2

```

[2025-04-13 14:04:57.225806][Data Processing]: push in the tag_queue...

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]


```

[2025-04-13 14:04:57.225806][Data Processing]: push in the tag_queue...
[2025-04-13 14:04:57.225806][Data Processing]: TAG is not valid and pop from tag_queue...
[2025-04-13 14:04:57.225806][Data Processing]: TAG is valid and push in tag_queue...
-----
[2025-04-13 14:04:57.225806]Test number 41 - index = 40

[2025-04-13 14:04:57.225806]Received data from packet:
[2025-04-13 14:04:57.225806]TIME STAMP : 1733062841900
[2025-04-13 14:04:57.225806]TAG id : a1b2c3d4e5f6g7h8
[2025-04-13 14:04:57.225806]ANGLE : 0
[2025-04-13 14:04:57.225806]HEIGHT : 1.1
-----
[2025-04-13 14:04:57.225806][Data Processing]: push in the tag_queue...
[2025-04-13 14:04:57.225806][Data Processing]: TAG is not valid and pop from tag_queue...
[2025-04-13 14:04:57.225806][Data Processing]: TAG is valid and push in tag_queue...
-----
[2025-04-13 14:04:57.225806]Test number 42 - index = 41

[2025-04-13 14:04:57.225806]Received data from packet:
[2025-04-13 14:04:57.225806]TIME STAMP : 1733062842000
[2025-04-13 14:04:57.225806]TAG id : a1b2c3d4e5f6g7h8
[2025-04-13 14:04:57.225806]ANGLE : 0
[2025-04-13 14:04:57.225806]HEIGHT : 1.1
-----
[2025-04-13 14:04:57.225806][Data Processing]: push in the tag_queue...
[2025-04-13 14:04:57.225806][Data Processing]: TAG is not valid and pop from tag_queue...
[2025-04-13 14:04:57.225806][Data Processing]: TAG is valid and push in tag_queue...
#####[2025-04-13 14:0
4:57.225806][Data Processing]: TAG queue print:
 deque([(1733062841000, 18), (1733062841100, 19), (1733062841200, 20), (1733062841300, 17), (1733062841400, 16),
(1733062841500, -18), (1733062841600, 19), (1733062841700, 20), (1733062841800, 30), (1733062841900, 0), (1733062
842000, 0)])
Number of valid samples: 11
Percentage of valid samples: 25.58%

[2025-04-13 14:04:57.225806]Filtered DataFrame:
   timestamp      tag_id  angle  tag_height
0  1733062840000  4baf351178aa9b0e  -30.0         1.2
1  1733062840100  4baf351178aa9b0e  -29.0         1.2
2  1733062840200  4baf351178aa9b0e  -30.0         1.2
3  1733062840300  4baf351178aa9b0e  -29.0         1.2
4  1733062840400  4baf351178aa9b0e  -28.0         1.2
5  1733062840500  4baf351178aa9b0e  -27.5         1.2
6  1733062840600  4baf351178aa9b0e  -27.0         1.2
7  1733062840700  4baf351178aa9b0e  -26.0         1.2
8  1733062840800  4baf351178aa9b0e  -25.0         1.2
9  1733062840900  4baf351178aa9b0e  -22.5         1.2
10 1733062841000  4baf351178aa9b0e  -25.0         1.2
11 1733062841100  4baf351178aa9b0e  -20.0         1.2
12 1733062841200  4baf351178aa9b0e   -4.0         1.2
13 1733062841300  4baf351178aa9b0e   -2.0         1.2
14 1733062841400  4baf351178aa9b0e    5.0         1.2
15 1733062841500  4baf351178aa9b0e   12.0         1.2
16 1733062841600  4baf351178aa9b0e   15.0         1.2
17 1733062841700  4baf351178aa9b0e   15.0         1.2
18 1733062841800  4baf351178aa9b0e   15.0         1.2
19 1733062841900  4baf351178aa9b0e   21.0         1.2
20 1733062842000  4baf351178aa9b0e   21.0         1.2
21 1733062840000  a1b2c3d4e5f6g7h8   -1.0         1.1
22 1733062840100  a1b2c3d4e5f6g7h8    8.0         1.1
23 1733062840200  a1b2c3d4e5f6g7h8   16.0         1.1
24 1733062840300  a1b2c3d4e5f6g7h8    7.5         1.1
25 1733062840400  a1b2c3d4e5f6g7h8   16.0         1.1
26 1733062840500  a1b2c3d4e5f6g7h8   16.5         1.1
27 1733062840600  a1b2c3d4e5f6g7h8   16.0         1.1
28 1733062840700  a1b2c3d4e5f6g7h8   15.5         1.1
29 1733062840800  a1b2c3d4e5f6g7h8   16.0         1.1
30 1733062840900  a1b2c3d4e5f6g7h8   15.5         1.1
31 1733062841000  a1b2c3d4e5f6g7h8   16.0         1.1
32 1733062841100  a1b2c3d4e5f6g7h8   17.0         1.1
33 1733062841200  a1b2c3d4e5f6g7h8   17.0         1.1
34 1733062841300  a1b2c3d4e5f6g7h8   17.0         1.1
35 1733062841400  a1b2c3d4e5f6g7h8   17.0         1.1
36 1733062841500  a1b2c3d4e5f6g7h8   17.0         1.1
37 1733062841600  a1b2c3d4e5f6g7h8   17.0         1.1
38 1733062841700  a1b2c3d4e5f6g7h8   17.0         1.1
39 1733062841800  a1b2c3d4e5f6g7h8   18.0         1.1
40 1733062841900  a1b2c3d4e5f6g7h8   18.0         1.1
41 1733062842000  a1b2c3d4e5f6g7h8   18.0         1.1
#####
[2025-04-13 14:04:57.225806] Check of the Length:
Input csv DF: 42
Output csv DF: 42

```

