APACHE PIG

Cleaner.py (Create this in VSCode / gedit / nano):

```
import re
import sys

for i in sys.stdin:
        print(re.sub(r'\s+',' ',i))
```

Command: cat weather.txt | python3 cleaner.py > weather_cleaned.txt

Program:

```
records = LOAD 'weather_cleaned.txt' USING PigStorage(' ') AS (number:int,
temperature:int, time:Chararray, a:float);
dump records;
```

Output:

```
(690190,13910,20060201_0,51.75)
(690190,13910,20060201_1,54.74)
(690190,13910,20060201_2,50.59)
(690190,13910,20060201_3,51.67)
(690190,13910,20060201_4,65.67)
(690190,13910,20060201_5,55.37)
(690190,13910,20060201_6,49.26)
(690190,13910,20060201_7,55.44)
(690190,13910,20060201_8,64.05)
(690190,13910,20060201_9,68.77)
(690190,13910,20060201_10,48.93)
(690190,13910,20060201_11,65.37)
(690190,13910,20060201_12,69.45)
(690190,13910,20060201_13,52.91)
(690190,13910,20060201_14,53.69)
(690190,13910,20060201_15,53.3)
(690190,13910,20060201_16,66.17)
(690190,13910,20060201_17,53.83)
(690190,13910,20060201_18,50.54)
(690190,13910,20060201_19,50.27)
(690190,13910,20060201_20,59.08)
(690190,13910,20060201_21,53.05)
(690190,13910,20060201_22,57.97)
(690190,13910,20060201_23,48.23)
(690190,13910,20060202_0,47.16)
(690190,13910,20060202_1,69.72)
(690190,13910,20060202_2,62.71)
(690190,13910,20060202_3,46.34)
(690190,13910,20060202_4,53.15)
(690190,13910,20060202_5,64.59)
(690190,13910,20060202_6,58.26)
(690190,13910,20060202_7,53.27)
(690190,13910,20060202_8,43.68)
(690190,13910,20060202_9,65.7)
```

(690190,13910,20060202_10,66.27)
(690190,13910,20060202_11,53.05)
(690190,13910,20060202_12,68.45)
(690190,13910,20060202_13,49.03)
(690190,13910,20060202_14,66.59)
(690190,13910,20060202_15,63.12)
(690190,13910,20060202_16,49.13)
(690190,13910,20060202_17,62.85)
(690190,13910,20060202_18,64.67)
(690190,13910,20060202_19,55.73)
(690190,13910,20060202_20,56.42)
(690190,13910,20060202_21,53.83)
(690190,13910,20060202_22,45.14)
(690190,13910,20060202_23,68.18)
(690190,13910,20060203_0,48.41)
(690190,13910,20060203_1,55.12)
(690190,13910,20060203_2,46.48)
(690190,13910,20060203_3,54.99)
(690190,13910,20060203_4,50.62)
(690190,13910,20060203_5,55.81)
(690190,13910,20060203_6,59.28)
(690190,13910,20060203_7,60.55)
(690190,13910,20060203_8,52.62)
(690190,13910,20060203_9,62.27)
(690190,13910,20060203_10,49.9)
(690190,13910,20060203_11,45.12)
(690190,13910,20060203_12,62.85)
(690190,13910,20060203_13,50.9)
(690190,13910,20060203_14,49.73)
(690190,13910,20060203_15,47.28)
(690190,13910,20060203_16,46.67)
(690190,13910,20060203_17,49.48)
(690190,13910,20060203_18,59.53)
(690190,13910,20060203_19,59.49)
(690190,13910,20060203_20,52.25)
(690190,13910,20060203_21,57.67)
(690190,13910,20060203_22,54.04)
(690190,13910,20060203_23,58.76)
(690190,13910,20060204_0,48.6)
(690190,13910,20060204_1,50.54)
(690190,13910,20060204_2,38.55)
(690190,13910,20060204_3,50.14)
(690190,13910,20060204_4,34.86)
(690190,13910,20060204_5,35.38)
(690190,13910,20060204_6,41.52)
(690190,13910,20060204_7,42.28)
(690190,13910,20060204_8,45.11)
(690190,13910,20060204_9,51.01)
(690190,13910,20060204_10,43.6)
(690190,13910,20060204_11,50.12)

(690190,13910,20060204_12,55.07)
(690190,13910,20060204_13,43.24)
(690190,13910,20060204_14,49.84)
(690190,13910,20060204_15,44.42)
(690190,13910,20060204_16,58.77)
(690190,13910,20060204_17,35.32)
(690190,13910,20060204_18,34.83)
(690190,13910,20060204_19,51.72)
(690190,13910,20060204_20,52.39)
(690190,13910,20060204_21,57.54)
(690190,13910,20060204_22,50.37)
(690190,13910,20060204_23,51.75)

filtered_records = FILTER records BY temperature == 13910 AND (a <= 35 OR a >= 50);
dump filtered_records;

Output:

(690190,13910,20060201_0,51.75)
(690190,13910,20060201_1,54.74)
(690190,13910,20060201_2,50.59)
(690190,13910,20060201_3,51.67)
(690190,13910,20060201_4,65.67)
(690190,13910,20060201_5,55.37)
(690190,13910,20060201_7,55.44)
(690190,13910,20060201_8,64.05)
(690190,13910,20060201_9,68.77)
(690190,13910,20060201_11,65.37)
(690190,13910,20060201_12,69.45)
(690190,13910,20060201_13,52.91)
(690190,13910,20060201_14,53.69)
(690190,13910,20060201_15,53.3)
(690190,13910,20060201_16,66.17)
(690190,13910,20060201_17,53.83)
(690190,13910,20060201_18,50.54)
(690190,13910,20060201_19,50.27)
(690190,13910,20060201_20,59.08)
(690190,13910,20060201_21,53.05)
(690190,13910,20060201_22,57.97)
(690190,13910,20060202_1,69.72)
(690190,13910,20060202_2,62.71)
(690190,13910,20060202_4,53.15)
(690190,13910,20060202_5,64.59)
(690190,13910,20060202_6,58.26)
(690190,13910,20060202_7,53.27)
(690190,13910,20060202_9,65.7)
(690190,13910,20060202_10,66.27)
(690190,13910,20060202_11,53.05)
(690190,13910,20060202_12,68.45)
(690190,13910,20060202_14,66.59)
(690190,13910,20060202_15,63.12)
(690190,13910,20060202_17,62.85)

(690190,13910,20060202_18,64.67)
(690190,13910,20060202_19,55.73)
(690190,13910,20060202_20,56.42)
(690190,13910,20060202_21,53.83)
(690190,13910,20060202_23,68.18)
(690190,13910,20060203_1,55.12)
(690190,13910,20060203_3,54.99)
(690190,13910,20060203_4,50.62)
(690190,13910,20060203_5,55.81)
(690190,13910,20060203_6,59.28)
(690190,13910,20060203_7,60.55)
(690190,13910,20060203_8,52.62)
(690190,13910,20060203_9,62.27)
(690190,13910,20060203_12,62.85)
(690190,13910,20060203_13,50.9)
(690190,13910,20060203_18,59.53)
(690190,13910,20060203_19,59.49)
(690190,13910,20060203_20,52.25)
(690190,13910,20060203_21,57.67)
(690190,13910,20060203_22,54.04)
(690190,13910,20060203_23,58.76)
(690190,13910,20060204_1,50.54)
(690190,13910,20060204_3,50.14)
(690190,13910,20060204_4,34.86)
(690190,13910,20060204_9,51.01)
(690190,13910,20060204_11,50.12)
(690190,13910,20060204_12,55.07)
(690190,13910,20060204_16,58.77)
(690190,13910,20060204_18,34.83)
(690190,13910,20060204_19,51.72)
(690190,13910,20060204_20,52.39)
(690190,13910,20060204_21,57.54)
(690190,13910,20060204_22,50.37)
(690190,13910,20060204_23,51.75)

grouped_records = GROUP records BY time;
dump grouped_records;
Output:
(20060201_0,{(690190,13910,20060201_0,51.75)})
(20060201_1,{(690190,13910,20060201_1,54.74)})
(20060201_2,{(690190,13910,20060201_2,50.59)})
(20060201_3,{(690190,13910,20060201_3,51.67)})
(20060201_4,{(690190,13910,20060201_4,65.67)})
(20060201_5,{(690190,13910,20060201_5,55.37)})
(20060201_6,{(690190,13910,20060201_6,49.26)})
(20060201_7,{(690190,13910,20060201_7,55.44)})
(20060201_8,{(690190,13910,20060201_8,64.05)})
(20060201_9,{(690190,13910,20060201_9,68.77)})
(20060202_0,{(690190,13910,20060202_0,47.16)})
(20060202_1,{(690190,13910,20060202_1,69.72)})

(20060202_2,{(690190,13910,20060202_2,62.71)})
(20060202_3,{(690190,13910,20060202_3,46.34)})
(20060202_4,{(690190,13910,20060202_4,53.15)})
(20060202_5,{(690190,13910,20060202_5,64.59)})
(20060202_6,{(690190,13910,20060202_6,58.26)})
(20060202_7,{(690190,13910,20060202_7,53.27)})
(20060202_8,{(690190,13910,20060202_8,43.68)})
(20060202_9,{(690190,13910,20060202_9,65.7)})
(20060203_0,{(690190,13910,20060203_0,48.41)})
(20060203_1,{(690190,13910,20060203_1,55.12)})
(20060203_2,{(690190,13910,20060203_2,46.48)})
(20060203_3,{(690190,13910,20060203_3,54.99)})
(20060203_4,{(690190,13910,20060203_4,50.62)})
(20060203_5,{(690190,13910,20060203_5,55.81)})
(20060203_6,{(690190,13910,20060203_6,59.28)})
(20060203_7,{(690190,13910,20060203_7,60.55)})
(20060203_8,{(690190,13910,20060203_8,52.62)})
(20060203_9,{(690190,13910,20060203_9,62.27)})
(20060204_0,{(690190,13910,20060204_0,48.6)})
(20060204_1,{(690190,13910,20060204_1,50.54)})
(20060204_2,{(690190,13910,20060204_2,38.55)})
(20060204_3,{(690190,13910,20060204_3,50.14)})
(20060204_4,{(690190,13910,20060204_4,34.86)})
(20060204_5,{(690190,13910,20060204_5,35.38)})
(20060204_6,{(690190,13910,20060204_6,41.52)})
(20060204_7,{(690190,13910,20060204_7,42.28)})
(20060204_8,{(690190,13910,20060204_8,45.11)})
(20060204_9,{(690190,13910,20060204_9,51.01)})
(20060201_10,{(690190,13910,20060201_10,48.93)})
(20060201_11,{(690190,13910,20060201_11,65.37)})
(20060201_12,{(690190,13910,20060201_12,69.45)})
(20060201_13,{(690190,13910,20060201_13,52.91)})
(20060201_14,{(690190,13910,20060201_14,53.69)})
(20060201_15,{(690190,13910,20060201_15,53.3)})
(20060201_16,{(690190,13910,20060201_16,66.17)})
(20060201_17,{(690190,13910,20060201_17,53.83)})
(20060201_18,{(690190,13910,20060201_18,50.54)})
(20060201_19,{(690190,13910,20060201_19,50.27)})
(20060201_20,{(690190,13910,20060201_20,59.08)})
(20060201_21,{(690190,13910,20060201_21,53.05)})
(20060201_22,{(690190,13910,20060201_22,57.97)})
(20060201_23,{(690190,13910,20060201_23,48.23)})
(20060202_10,{(690190,13910,20060202_10,66.27)})
(20060202_11,{(690190,13910,20060202_11,53.05)})
(20060202_12,{(690190,13910,20060202_12,68.45)})
(20060202_13,{(690190,13910,20060202_13,49.03)})
(20060202_14,{(690190,13910,20060202_14,66.59)})
(20060202_15,{(690190,13910,20060202_15,63.12)})
(20060202_16,{(690190,13910,20060202_16,49.13)})
(20060202_17,{(690190,13910,20060202_17,62.85)})

(20060202_18,{(690190,13910,20060202_18,64.67)})
(20060202_19,{(690190,13910,20060202_19,55.73)})
(20060202_20,{(690190,13910,20060202_20,56.42)})
(20060202_21,{(690190,13910,20060202_21,53.83)})
(20060202_22,{(690190,13910,20060202_22,45.14)})
(20060202_23,{(690190,13910,20060202_23,68.18)})
(20060203_10,{(690190,13910,20060203_10,49.9)})
(20060203_11,{(690190,13910,20060203_11,45.12)})
(20060203_12,{(690190,13910,20060203_12,62.85)})
(20060203_13,{(690190,13910,20060203_13,50.9)})
(20060203_14,{(690190,13910,20060203_14,49.73)})
(20060203_15,{(690190,13910,20060203_15,47.28)})
(20060203_16,{(690190,13910,20060203_16,46.67)})
(20060203_17,{(690190,13910,20060203_17,49.48)})
(20060203_18,{(690190,13910,20060203_18,59.53)})
(20060203_19,{(690190,13910,20060203_19,59.49)})
(20060203_20,{(690190,13910,20060203_20,52.25)})
(20060203_21,{(690190,13910,20060203_21,57.67)})
(20060203_22,{(690190,13910,20060203_22,54.04)})
(20060203_23,{(690190,13910,20060203_23,58.76)})
(20060204_10,{(690190,13910,20060204_10,43.6)})
(20060204_11,{(690190,13910,20060204_11,50.12)})
(20060204_12,{(690190,13910,20060204_12,55.07)})
(20060204_13,{(690190,13910,20060204_13,43.24)})
(20060204_14,{(690190,13910,20060204_14,49.84)})
(20060204_15,{(690190,13910,20060204_15,44.42)})
(20060204_16,{(690190,13910,20060204_16,58.77)})
(20060204_17,{(690190,13910,20060204_17,35.32)})
(20060204_18,{(690190,13910,20060204_18,34.83)})
(20060204_19,{(690190,13910,20060204_19,51.72)})
(20060204_20,{(690190,13910,20060204_20,52.39)})
(20060204_21,{(690190,13910,20060204_21,57.54)})
(20060204_22,{(690190,13910,20060204_22,50.37)})
(20060204_23,{(690190,13910,20060204_23,51.75)})

```
max_temp = FOREACH grouped_records GENERATE group, MAX(records.temperature);
dump max_temp;
```

Output:

(20060201_0,13910)
(20060201_1,13910)
(20060201_2,13910)
(20060201_3,13910)
(20060201_4,13910)
(20060201_5,13910)
(20060201_6,13910)
(20060201_7,13910)
(20060201_8,13910)
(20060201_9,13910)
(20060202_0,13910)
(20060202_1,13910)

(20060202_2,13910)
(20060202_3,13910)
(20060202_4,13910)
(20060202_5,13910)
(20060202_6,13910)
(20060202_7,13910)
(20060202_8,13910)
(20060202_9,13910)
(20060203_0,13910)
(20060203_1,13910)
(20060203_2,13910)
(20060203_3,13910)
(20060203_4,13910)
(20060203_5,13910)
(20060203_6,13910)
(20060203_7,13910)
(20060203_8,13910)
(20060203_9,13910)
(20060204_0,13910)
(20060204_1,13910)
(20060204_2,13910)
(20060204_3,13910)
(20060204_4,13910)
(20060204_5,13910)
(20060204_6,13910)
(20060204_7,13910)
(20060204_8,13910)
(20060204_9,13910)
(20060201_10,13910)
(20060201_11,13910)
(20060201_12,13910)
(20060201_13,13910)
(20060201_14,13910)
(20060201_15,13910)
(20060201_16,13910)
(20060201_17,13910)
(20060201_18,13910)
(20060201_19,13910)
(20060201_20,13910)
(20060201_21,13910)
(20060201_22,13910)
(20060201_23,13910)
(20060202_10,13910)
(20060202_11,13910)
(20060202_12,13910)
(20060202_13,13910)
(20060202_14,13910)
(20060202_15,13910)
(20060202_16,13910)
(20060202_17,13910)

(20060202_18,13910)
(20060202_19,13910)
(20060202_20,13910)
(20060202_21,13910)
(20060202_22,13910)
(20060202_23,13910)
(20060203_10,13910)
(20060203_11,13910)
(20060203_12,13910)
(20060203_13,13910)
(20060203_14,13910)
(20060203_15,13910)
(20060203_16,13910)
(20060203_17,13910)
(20060203_18,13910)
(20060203_19,13910)
(20060203_20,13910)
(20060203_21,13910)
(20060203_22,13910)
(20060203_23,13910)
(20060204_10,13910)
(20060204_11,13910)
(20060204_12,13910)
(20060204_13,13910)
(20060204_14,13910)
(20060204_15,13910)
(20060204_16,13910)
(20060204_17,13910)
(20060204_18,13910)
(20060204_19,13910)
(20060204_20,13910)
(20060204_21,13910)
(20060204_22,13910)
(20060204_23,13910)

describe records;
Output:
records: {number: int,temperature: int,time: chararray,a: float}

describe filtered_records;
Output:
filtered_records: {number: int,temperature: int,time: chararray,a: float}

describe grouped_records;
Output:
grouped_records: {group: chararray,records: {(number: int,temperature: int,time: chararray,a: float)}}

describe max_temp;
Output:

max_temp: {group: chararray,int}