# Stat480_Homework 5

Bin Feng

## 0. Data Preparation (Command line commands)

```
# Download dataset
cd ~/Stat480/hb-workspace/input/ncdc
chmod u+x ncdc_data.sh
./ncdc_data.sh 1915 1924
# Copy data to Hadoop
hadoop fs -put ~/Stat480/hb-workspace/input/ncdc/all/* input/ncdc/HW5
# Go to python files stored folder
cd ~/Stat480/hb-workspace/ch02-mr-intro/src/main/python
```

## 1. Exercise 1

Using Hadoop and MapReduce, find the minimum monthly recorded air temperature from 1915 to 1924 and return those minimum values in degrees Celsius. (You should have 12 values total, one for each month).

**Map script:** min_temperature_map.py
**Reduce script:** min_temperature_reduce.py
**Command line commands:**

```
# Create new map/reduce file by copying and editing the exists
cp max_temperature_map.py min_temperature_map.py
cp max_temperature_reduce.py min_temperature_reduce.py

# Open and edit python map/reduce files
vi min_temperature_map.py
vi min_temperature_reduce.py

# Run MapReduce
hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
-files /home/binfeng2/Stat480/hb-workspace/ch02-mr-intro/src/main/python/
min_temperature_map.py, /home/binfeng2/Stat480/hb-workspace/ch02-mr-
intro/src/main/python/min_temperature_reduce.py \
-input input/ncdc/HW5/ \
-output outputpy \
-mapper "/home/binfeng2/Stat480/hb-workspace/ch02-mr-intro/src/main/python/
min_temperature_map.py" \
-reducer "/home/binfeng2/Stat480/hb-workspace/ch02-mr-intro/src/main/python/
min_temperature_reduce.py"

# Show results
hadoop fs -cat outputpy/part*
```

*# Delete directory*
hadoop fs -rm -r -f outputpy

**Results:**

| Month | Min Temperature (Celsius) |
|---|---|
| 01 | -45.6 |
| 02 | -47.8 |
| 03 | -39.4 |
| 04 | -31.1 |
| 05 | -7.8 |
| 06 | -2.8 |
| 07 | 1.1 |
| 08 | 0.0 |
| 09 | -13.9 |
| 10 | -28.9 |
| 11 | -36.1 |
| 12 | -42.8 |

The table above shows results queried for the minimum monthly recorded air temperature in degrees Celsius from 1915 and 1924. Note that all months' minimum temperatures are below or equal 0 degree Celsius except July. The minimum temperature among all months is -47.8 degree Celsius recorded in February.

## 2. Exercise 2

Using Hadoop and MapReduce, obtain the number of trusted temperature observations and the minimum and maximum monthly temperatures in degrees Fahrenheit over the period of 1915 to 1924. Make sure your code only goes through the data once to get these results (to do this you will need to update the minimum, maximum, and count at the same step in the code).

**Map script:** trust_max_min_temperature_map.py
**Reduce script:** trust_max_min_temperature_reduce.py
**Command line commands:**

*# Create new map/reduce file by copying and editing the exists*
cp max_temperature_map.py trust_max_min_temperature_map.py
cp max_temperature_reduce.py trust_max_min_temperature_reduce.py

*# Open and edit python map/reduce files*
vi trust_max_min_temperature_map.py
vi trust_max_min_temperature_reduce.py

*# Run MapReduce*
hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
-files /home/binfeng2/Stat480/hb-workspace/ch02-mr-intro/src/main/python/
trust_max_min_temperature_map.py,\

```
/home/binfeng2/Stat480/hb-workspace/ch02-mr-intro/src/main/python/
trust_max_min_temperature_reduce.py \
-input input/ncdc/HW5/ \
-output outputpy \
-mapper "/home/binfeng2/Stat480/hb-workspace/ch02-mr-intro/src/main/python/
trust_max_min_temperature_map.py" \
-reducer "/home/binfeng2/Stat480/hb-workspace/ch02-mr-intro/src/main/python/t
rust_max_min_temperature_reduce.py"

# Show results
hadoop fs -cat outputpy/part*

# Delete directory
hadoop fs -rm -r -f outputpy
```

**Results:**

| Month | Max Temperature (Fahrenheit) | Min Temperature (Fahrenheit) | Observations |
|---|---|---|---|
| 01 | 44.96 | -50.08 | 6500 |
| 02 | 44.96 | -54.04 | 6003 |
| 03 | 62.96 | -38.92 | 6572 |
| 04 | 77.0 | -23.98 | 6285 |
| 05 | 82.94 | 17.96 | 6571 |
| 06 | 89.06 | 26.96 | 6342 |
| 07 | 100.04 | 33.98 | 6581 |
| 08 | 86.0 | 32.0 | 6505 |
| 09 | 75.92 | 6.98 | 6183 |
| 10 | 62.06 | -20.02 | 6502 |
| 11 | 50.0 | -32.98 | 6294 |
| 12 | 50.0 | -45.04 | 6504 |

The table above shows results queried for the number of trusted temperature observations and the minimum and maximum monthly temperatures in degrees Fahrenheit over the period of 1915 to 1924. Note that all months' maximum temperatures are above 0 degree Fahrenheit with the maximum of all at 100.04 degree Fahrenheit in July. And month 1, 2, 3, 4, 10, 11, and 12 have minimum temperatures below 0 degree Fahrenheit. The minimum temperature among all months is -54.04 degree Fahrenheit recorded in February. The number of observations are quite equal among all months.

## 3. Exercise 3

Using Hadoop and MapReduce, obtain the total number of air temperature observations that are not missing for each month during the period from 1915 to 1924 and the total number of observations with acceptable quality codes for each month during that period. Make sure your code only goes through the data once to get these results (to do this, you could have the mapper return (month, tempcount, validqcount) for each observation, and have the reducer aggregate).

**Map script:** count_temperature_map.py
**Reduce script:** count_temperature_reduce.py
**Command line commands:**

```
# Create new map/reduce file by copying and editing the exists
cp max_temperature_map.py count_temperature_map.py
cp max_temperature_reduce.py count_temperature_reduce.py

# Open and edit python map/reduce files
vi count_temperature_map.py
vi count_temperature_reduce.py

# Run MapReduce
hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
-files /home/binfeng2/Stat480/hb-workspace/ch02-mr-intro/src/main/python/
count_temperature_map.py,\
/home/binfeng2/Stat480/hb-workspace/ch02-mr-intro/src/main/python/
count_temperature_reduce.py \
-input input/ncdc/HW5/ \
-output outputpy \
-mapper "/home/binfeng2/Stat480/hb-workspace/ch02-mr-intro/src/main/python/
count_temperature_map.py" \
-reducer "/home/binfeng2/Stat480/hb-workspace/ch02-mr-intro/src/main/python/
count_temperature_reduce.py"

# Show results
hadoop fs -cat outputpy/part*

# Delete directory
hadoop fs -rm -r -f outputpy
```

**Results:**

| Month | Air Temperature Observations | Quality Code Observations |
|---|---|---|
| 01 | 6500 | 6500 |
| 02 | 6003 | 6005 |
| 03 | 6572 | 6595 |
| 04 | 6285 | 6286 |
| 05 | 6571 | 6578 |
| 06 | 6342 | 6365 |
| 07 | 6581 | 6595 |
| 08 | 6505 | 6508 |
| 09 | 6183 | 6202 |
| 10 | 6502 | 6502 |
| 11 | 6294 | 6294 |
| 12 | 6504 | 6504 |

The table above shows results queried for the total number of air temperature observations that are not missing for each month during the period from 1915 to 1924 and the total number of observations with acceptable quality codes for each month during that period. Note that these two set of numbers quite similar with only a few differences. Also note that the number of observations are quite equal among all months.

## 4. Exercise 4

Using Hadoop and MapReduce, obtain the monthly mean air temperature in degrees Celsius for the period from 1915 to 1924. If you use a combiner, make sure your code will work when data needs to be recombined from samples of different sizes.

**Map script:** mean_temperature_map.py
**Reduce script:** mean_temperature_reduce.py
**Command line commands:**

```
# Create new map/reduce file by copying and editing the exists
cp min_temperature_map.py mean_temperature_map.py
cp min_temperature_reduce.py mean_temperature_reduce.py

# Open and edit python map/reduce files
vi mean_temperature_map.py
vi mena_temperature_reduce.py

# Run MapReduce
hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
-files /home/binfeng2/Stat480/hb-workspace/ch02-mr-intro/src/main/python/
mean_temperature_map.py,\
/home/binfeng2/Stat480/hb-workspace/ch02-mr-intro/src/main/python/
mean_temperature_reduce.py \
-input input/ncdc/HW5/ \
-output outputpy \
-mapper "/home/binfeng2/Stat480/hb-workspace/ch02-mr-intro/src/main/python/
mean_temperature_map.py" \
-reducer "/home/binfeng2/Stat480/hb-workspace/ch02-mr-intro/src/main/python/
mean_temperature_reduce.py"

# Show results
hadoop fs -cat outputpy/part*

# Delete directory
hadoop fs -rm -r -f outputpy
```

**Results:**

| Month | Mean Air Temperature (Celsius) |
|---|---|
| 01 | -8.13858461538 |
| 02 | -8.37404631018 |
| 03 | -4.85763846622 |
| 04 | 1.01907716786 |
| 05 | 7.09373002587 |
| 06 | 12.0541784926 |
| 07 | 15.80205136 |
| 08 | 13.4259800154 |
| 09 | 9.11060973637 |
| 10 | 3.68974161796 |
| 11 | -1.7658087067 |
| 12 | -5.91194649446 |

The table above shows results queried for the monthly mean air temperature in degrees Celsius for the period from 1915 to 1924. Note that month 1, 2, 3, 11, 12 have mean air temperature below 0 with minimum mean temperature of all achieved in February for -8.37404631018 degree Celsius. The maximum mean temperature of all achieved in August for 13.4259800154 degree Celsius.