CIE 427 Project Report

Weather Data Analytics

Name: Ahmed Muhammad Tarek
ID: 201500885
E-mail: s-ahmed.tarek@zewailcity.edu.eg

Contents

Problem Description

In this project, we will use weather data collected by land-based stations from all over the world. The aim is to undergo exploratory analysis of the data to investigate the weather trends in different periods of time and analyze the impact of global warming. NASA claims that there is a global rise in temperature due to global warming. Eighteen of the nineteen warmest years in history have occurred since 2001. The temperature anomaly is illustrated in figure 1 where anomaly means the departure from a reference or an average value.

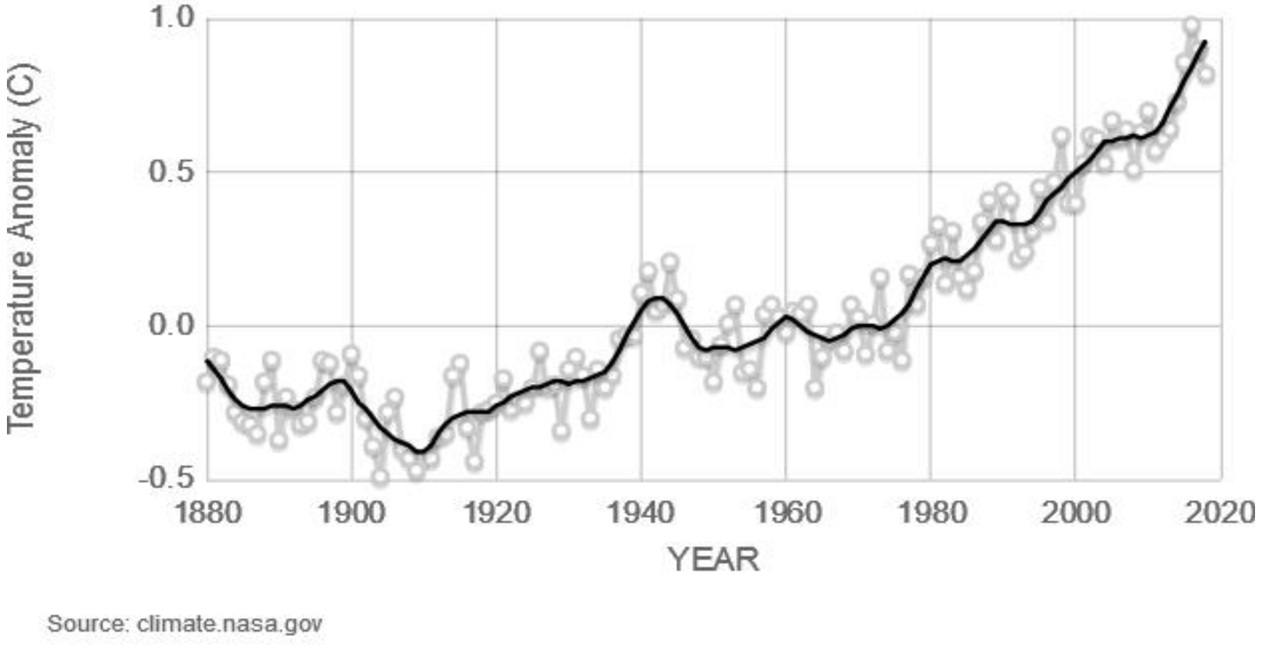


Figure 1. Temperature anomaly over the years

We aim to use the Global Historical Climatology Network Daily (GHCN-Daily) dataset to verify the graph above.

Description of the dataset

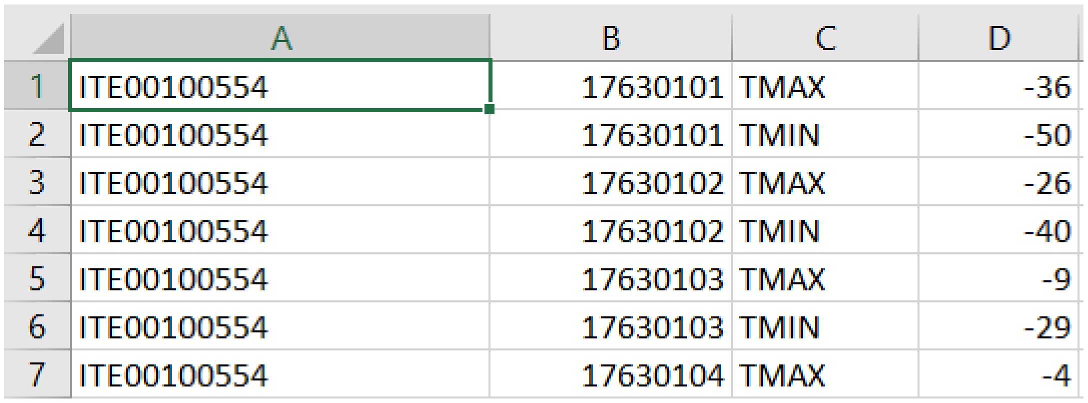
The dataset (GHCN-Daily) is collected and maintained by the National Oceanic and Atmospheric Administration (NOAA). It can be accessed through queries to an FTP API and different formats are available. We are using the CSV format where the observations of each year are stored in a file 'year.csv' The total uncompressed size of the data is 94 gigabytes

The structure of the dataset is as follows:

Station ID	Date	Feature Name	Feature Value

Examples of features: maximum temperature, minimum temperature, snow depth, precipitation and snowfall

There are other columns; however, they are not of interest to us; they are mainly flags related to quality assurance. An example of a record in the dataset is shown in figure 2.



	A	B	C	D
1	ITE00100554	17630101	TMAX	-36
2	ITE00100554	17630101	TMIN	-50
3	ITE00100554	17630102	TMAX	-26
4	ITE00100554	17630102	TMIN	-40
5	ITE00100554	17630103	TMAX	-9
6	ITE00100554	17630103	TMIN	-29
7	ITE00100554	17630104	TMAX	-4

Figure 2. Part of the 1763.csv record in the dataset

Analysis

In order to freely experiment with the relevant data to our problem in the dataset, we aim to run an operation on the dataset where the output is a dictionary that is accessed using three keys: year, month-day and country where the element is the average temperature on the date “day-month-year” in the specified country. Accessing the dictionary with a key $year = "1997"$, $month - day = "0201"$ and $country = "AE"$ is illustrated in figure 3.

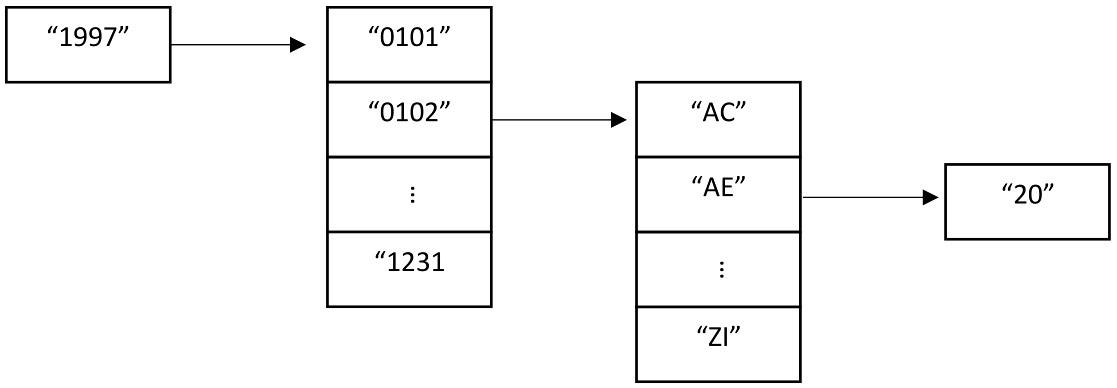


Figure 3. Accessing the temperature dictionary

Analysis Pipeline

Pseudocodes of initial attempts

First pseudocode

```
Generate a list of all possible dates and a list of all countries.
Loop over all files and then loop over dates list and countries list filtering the RDD by a date instance, a country instance and the required feature.
Compute the average temperature and store the output.
```

Estimated local runtime: 4 days

Second Pseudocode

```
Generate a list of all possible dates.
Loop over all files and then loop over dates filtering the RDD by a date instance and the required feature
Compute the average temperature using the country as a key in the reducing operation and store the output.
```

Estimated local runtime: 2 days

Third Pseudocode

```
Loop over all files
Filter the RDD by the required feature
Compute the average temperature using the country and the date as a key in the reducing operation and store the output.
End Loop
```

Estimated local runtime: 3 hours

Detailed description of the final pseudocode

Final Pseudocode

```
Loop over all files and merge them in a single RDD
End Loop
Filter the RDD to keep only the correct temperature features
Extract the key-value pair (date-country, temperature) using a map operation
Country the number of occurrences for each key using a word count step and store it in a separate RDD called counts
```

Compute the sum of the temperatures for each key using a reduce by key step and store it in an RDD called sums.

Join the two RDDs sums and counts by key such that the key-value pairs have the form (date-country, (count, sum)) where count is the number of occurrences of each key and sum is the sum of temperatures for each key.

Calculate the average for each key by dividing the first value by the second value and then divide by 2 (as temperatures are reported as minimum and maximum values for each day) using a map operation and store the result in a temperature dictionary.

Estimated Runtime on the cluster: 1 hour

Flowchart of the final pipeline

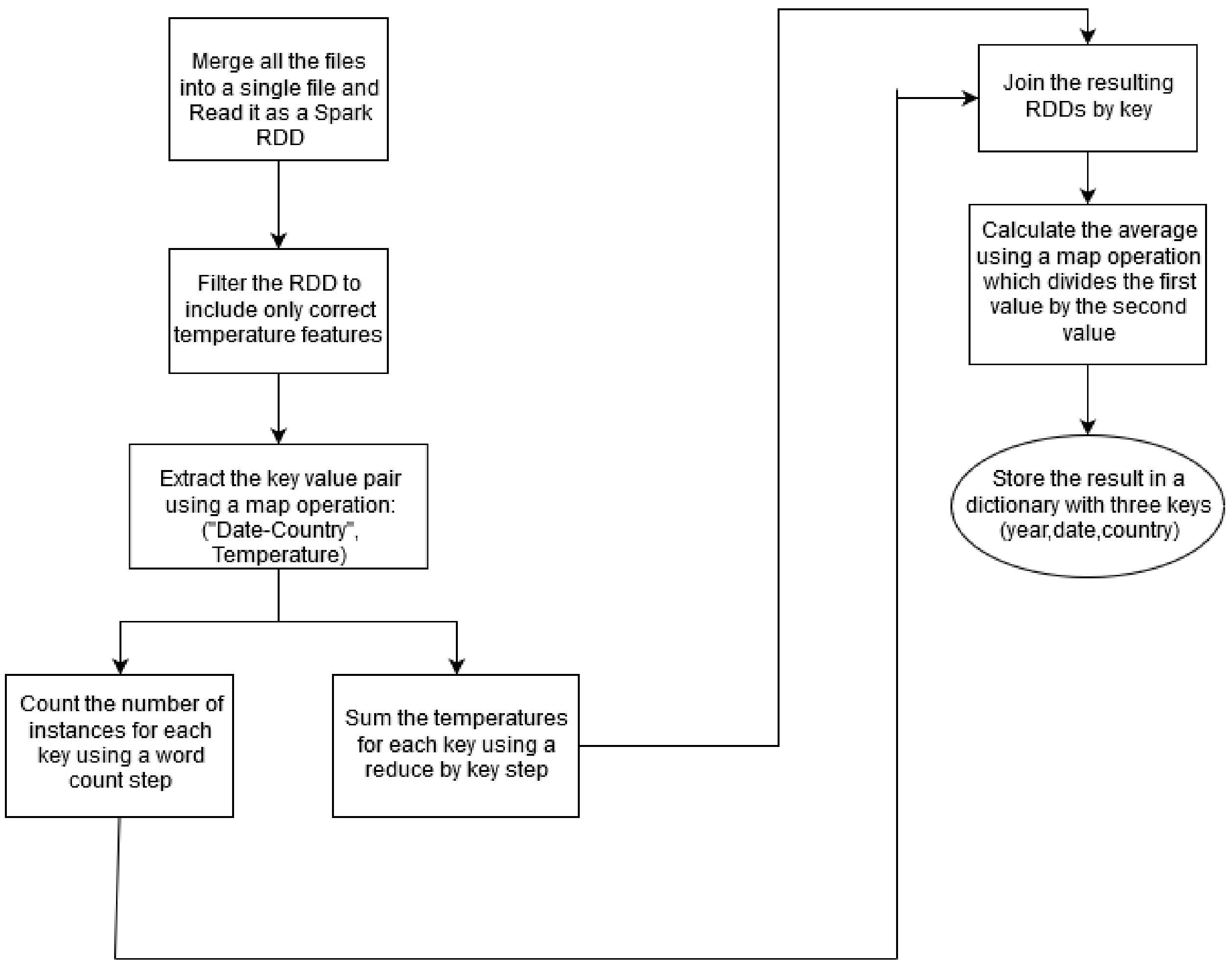

Merge all the files into a single file and Read it as a Spark RDD
Filter the RDD to include only correct temperature features
Extract the key value pair using a map operation: ("Date-Country", Temperature)
Count the number of instances for each key using a word count step
Sum the temperatures for each key using a reduce by key step
Join the resulting RDDs by key
Calculate the average using a map operation which divides the first value by the second value
Store the result in a dictionary with three keys (year,date,country)

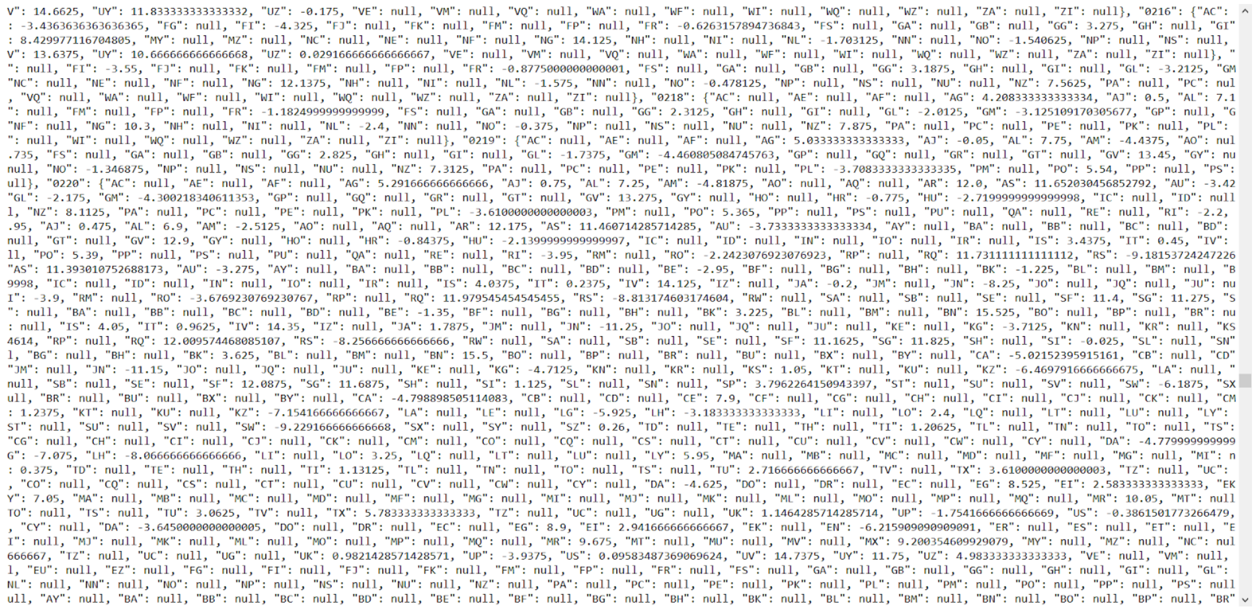
Figure 4. Flowchart of the final analysis pipeline

The reduction in the runtime achieved by the final pseudocode could not be achieved without running on the cluster. Merging all the files into a single RDD cannot be done as the data is too large to fit in the memory of the local machine and spark fails to manage. The estimated runtime on the cluster is calculated by an experiment we made on approximately 15% of the data which took about 9 minutes on the cluster.

Code

```
import os
dir = '/content/dataset'
files = os.listdir(dir)
dict = {year:{date:{country:float('nan') for country in countries} for dat
e in dates} for year in years}
counter = 1
data = sc.emptyRDD()
for file in files:
  if '.csv' in file:
    print(file)
    print("Counter: "+str(counter))
    counter = counter+1
    year = file.split('.')[0]
    data = data.union(sc.textFile('/content/dataset/'+file).map(lambda x:
x.split(',')[0:4]))
T = data.filter(lambda x: x[2][0:2]=='TM' and x[3]<=700)
T = T.map(lambda x: (x[1]+x[0][0:2],int(x[3])))
counts = T.map(lambda x: (x[0],1)).reduceByKey(lambda a,b:a+b)
T = T.reduceByKey(lambda a,b: a+b)
T = T.join(counts)
T = T.map(lambda x: (x[0],x[1][0]/x[1][1]/20))
temp_summary = T.collect()
for summary in temp_summary:
  length = len(summary[0])
  date = summary[0][4:length-2]
  year = summary[0][0:4]
  country = summary[0][length-2:length]
  temp_dict[year][date][country]=summary[1]
```

Output of the analysis



```
V": 14.6625, "UY": 11.833333333333332, "UZ": -0.175, "VE": null, "VM": null, "VQ": null, "WA": null, "WF": null, "WI": null, "WQ": null, "WZ": null, "ZA": null, "ZI": null}, "0216": {"AC":
: -3.4363636363636365, "FG": null, "FI": -4.325, "FJ": null, "FK": null, "FM": null, "FP": null, "FR": -0.6263157894736843, "FS": null, "GA": null, "GB": null, "GG": 3.275, "GH": null, "GI"
: 8.429977116704805, "MY": null, "MZ": null, "NC": null, "NE": null, "NF": null, "NG": 14.125, "NH": null, "NI": null, "NL": -1.703125, "NN": null, "NO": -1.540625, "NP": null, "NS": null,
V": 13.6375, "UY": 10.666666666666668, "UZ": 0.029166666666666667, "VE": null, "VM": null, "VQ": null, "WA": null, "WF": null, "WI": null, "WQ": null, "WZ": null, "ZA": null, "ZI": null}, "
": null, "FI": -3.55, "FJ": null, "FK": null, "FM": null, "FP": null, "FR": -0.8775000000000001, "FS": null, "GA": null, "GB": null, "GG": 3.1875, "GH": null, "GI": null, "GL": -3.2125, "GM
"NC": null, "NE": null, "NF": null, "NG": 12.1375, "NH": null, "NI": null, "NL": -1.575, "NN": null, "NO": -0.478125, "NP": null, "NS": null, "NU": null, "NZ": 7.5625, "PA": null, "PC": nul
, "VQ": null, "WA": null, "WF": null, "WI": null, "WQ": null, "WZ": null, "ZA": null, "ZI": null}, "0218": {"AC": null, "AE": null, "AF": null, "AG": 4.208333333333334, "AJ": 0.5, "AL": 7.1
": null, "FM": null, "FP": null, "FR": -1.1824999999999999, "FS": null, "GA": null, "GB": null, "GG": 2.3125, "GH": null, "GI": null, "GL": -2.0125, "GM": -3.125109170305677, "GP": null, "G
"NF": null, "NG": 10.3, "NH": null, "NI": null, "NL": -2.4, "NN": null, "NO": -0.375, "NP": null, "NS": null, "NU": null, "NZ": 7.875, "PA": null, "PC": null, "PE": null, "PK": null, "PL":
": null, "WI": null, "WQ": null, "WZ": null, "ZA": null, "ZI": null}, "0219": {"AC": null, "AE": null, "AF": null, "AG": 5.033333333333333, "AJ": -0.05, "AL": 7.75, "AM": -4.4375, "AO": nul
.735, "FS": null, "GA": null, "GB": null, "GG": 2.825, "GH": null, "GI": null, "GL": -1.7375, "GM": -4.460805084745763, "GP": null, "GQ": null, "GR": null, "GT": null, "GV": 13.45, "GY": nu
null, "NO": -1.346875, "NP": null, "NS": null, "NU": null, "NZ": 7.3125, "PA": null, "PC": null, "PE": null, "PK": null, "PL": -3.7083333333333335, "PM": null, "PO": 5.54, "PP": null, "PS":
ull}, "0220": {"AC": null, "AE": null, "AF": null, "AG": 5.291666666666666, "AJ": 0.75, "AL": 7.25, "AM": -4.81875, "AO": null, "AQ": null, "AR": 12.0, "AS": 11.652030456852792, "AU": -3.42
"GL": -2.175, "GM": -4.300218340611353, "GP": null, "GQ": null, "GR": null, "GT": null, "GV": 13.275, "GY": null, "HO": null, "HR": -0.775, "HU": -2.7199999999999998, "IC": null, "ID": null
l, "NZ": 8.1125, "PA": null, "PC": null, "PE": null, "PK": null, "PL": -3.6100000000000003, "PM": null, "PO": 5.365, "PP": null, "PS": null, "PU": null, "QA": null, "RE": null, "RI": -2.2,
.95, "AJ": 0.475, "AL": 6.9, "AM": -2.5125, "AO": null, "AQ": null, "AR": 12.175, "AS": 11.460714285714285, "AU": -3.7333333333333334, "AY": null, "BA": null, "BB": null, "BC": null, "BD":
null, "GT": null, "GV": 12.9, "GY": null, "HO": null, "HR": -0.84375, "HU": -2.1399999999999997, "IC": null, "ID": null, "IN": null, "IO": null, "IR": null, "IS": 3.4375, "IT": 0.45, "IV":
ll, "PO": 5.39, "PP": null, "PS": null, "PU": null, "QA": null, "RE": null, "RI": -3.95, "RM": null, "RO": -2.2423076923076923, "RP": null, "RQ": 11.731111111111112, "RS": -9.18153724247226
"AS": 11.393010752688173, "AU": -3.275, "AY": null, "BA": null, "BB": null, "BC": null, "BD": null, "BE": -2.95, "BF": null, "BG": null, "BH": null, "BK": -1.225, "BL": null, "BM": null, "B
9998, "IC": null, "ID": null, "IN": null, "IO": null, "IR": null, "IS": 4.0375, "IT": 0.2375, "IV": 14.125, "IZ": null, "JA": -0.2, "JM": null, "JN": -8.25, "JO": null, "JQ": null, "JU": nu
I": -3.9, "RM": null, "RO": -3.6769230769230767, "RP": null, "RQ": 11.979545454545455, "RS": -8.813174603174604, "RW": null, "SA": null, "SB": null, "SE": null, "SF": 11.4, "SG": 11.275, "S
": null, "BA": null, "BB": null, "BC": null, "BD": null, "BE": -1.35, "BF": null, "BG": null, "BH": null, "BK": 3.225, "BL": null, "BM": null, "BN": 15.525, "BO": null, "BP": null, "BR": nu
: null, "IS": 4.05, "IT": 0.9625, "IV": 14.35, "IZ": null, "JA": 1.7875, "JM": null, "JN": -11.25, "JO": null, "JQ": null, "JU": null, "KE": null, "KG": -3.7125, "KN": null, "KR": null, "KS
4614, "RP": null, "RQ": 12.009574468085107, "RS": -8.256666666666666, "RW": null, "SA": null, "SB": null, "SE": null, "SF": 11.1625, "SG": 11.825, "SH": null, "SI": -0.025, "SL": null, "SN"
l, "BG": null, "BH": null, "BK": 3.625, "BL": null, "BM": null, "BN": 15.5, "BO": null, "BP": null, "BR": null, "BU": null, "BX": null, "BY": null, "CA": -5.02152395915161, "CB": null, "CD"
"JM": null, "JN": -11.15, "JO": null, "JQ": null, "JU": null, "KE": null, "KG": -4.7125, "KN": null, "KR": null, "KS": 1.05, "KT": null, "KU": null, "KZ": -6.4697916666666675, "LA": null, "
null, "SB": null, "SE": null, "SF": 12.0875, "SG": 11.6875, "SH": null, "SI": 1.125, "SL": null, "SN": null, "SP": 3.7962264150943397, "ST": null, "SU": null, "SV": null, "SW": -6.1875, "SX
ull, "BR": null, "BU": null, "BX": null, "BY": null, "CA": -4.798898505114083, "CB": null, "CD": null, "CE": 7.9, "CF": null, "CG": null, "CH": null, "CI": null, "CJ": null, "CK": null, "CM
: 1.2375, "KT": null, "KU": null, "KZ": -7.154166666666667, "LA": null, "LE": null, "LG": -5.925, "LH": -3.183333333333333, "LI": null, "LO": 2.4, "LQ": null, "LT": null, "LU": null, "LY":
ST": null, "SU": null, "SV": null, "SW": -9.229166666666668, "SX": null, "SY": null, "SZ": 0.26, "TD": null, "TE": null, "TH": null, "TI": 1.20625, "TL": null, "TN": null, "TO": null, "TS":
"CG": null, "CH": null, "CI": null, "CJ": null, "CK": null, "CM": null, "CO": null, "CQ": null, "CS": null, "CT": null, "CU": null, "CV": null, "CW": null, "CY": null, "DA": -4.779999999999
G": -7.075, "LH": -8.066666666666666, "LI": null, "LO": 3.25, "LQ": null, "LT": null, "LU": null, "LY": 5.95, "MA": null, "MB": null, "MC": null, "MD": null, "MF": null, "MG": null, "MI": n
: 0.375, "TD": null, "TE": null, "TH": null, "TI": 1.13125, "TL": null, "TN": null, "TO": null, "TS": null, "TU": 2.716666666666667, "TV": null, "TX": 3.6100000000000003, "TZ": null, "UC":
, "CO": null, "CQ": null, "CS": null, "CT": null, "CU": null, "CV": null, "CW": null, "CY": null, "DA": -4.625, "DO": null, "DR": null, "EC": null, "EG": 8.525, "EI": 2.583333333333333, "EK
Y": 7.05, "MA": null, "MB": null, "MC": null, "MD": null, "MF": null, "MG": null, "MI": null, "MJ": null, "MK": null, "ML": null, "MO": null, "MP": null, "MQ": null, "MR": 10.05, "MT": null
TO": null, "TS": null, "TU": 3.0625, "TV": null, "TX": 5.783333333333333, "TZ": null, "UC": null, "UG": null, "UK": 1.1464285714285714, "UP": -1.7541666666666669, "US": -0.3861501773266479,
, "CY": null, "DA": -3.6450000000000005, "DO": null, "DR": null, "EC": null, "EG": 8.9, "EI": 2.941666666666667, "EK": null, "EN": -6.215909090909091, "ER": null, "ES": null, "ET": null, "E
I": null, "MJ": null, "MK": null, "ML": null, "MO": null, "MP": null, "MQ": null, "MR": 9.675, "MT": null, "MU": null, "MV": null, "MX": 9.200354609929079, "MY": null, "MZ": null, "NC": nul
666667, "TZ": null, "UC": null, "UG": null, "UK": 0.9821428571428571, "UP": -3.9375, "US": 0.09583487369069624, "UV": 14.7375, "UY": 11.75, "UZ": 4.983333333333333, "VE": null, "VM": null,
l, "EU": null, "EZ": null, "FG": null, "FI": null, "FJ": null, "FK": null, "FM": null, "FP": null, "FR": null, "FS": null, "GA": null, "GB": null, "GG": null, "GH": null, "GI": null, "GL":
NL": null, "NN": null, "NO": null, "NP": null, "NS": null, "NU": null, "NZ": null, "PA": null, "PC": null, "PE": null, "PK": null, "PL": null, "PM": null, "PO": null, "PP": null, "PS": null
ull, "AY": null, "BA": null, "BB": null, "BC": null, "BD": null, "BE": null, "BF": null, "BG": null, "BH": null, "BK": null, "BL": null, "BM": null, "BN": null, "BO": null, "BP": null, "BR"
```

Results and Evaluation

The organization of this section will be based on the approach used for generating the graph and its results.

First approach

The first approach is to calculate the average value of the temperature for each year and then plot the temperature anomaly where the reference is the minimum value. The result of using this approach is illustrated in figure 5. The blue dots are actual values and the red line is lowess regression of the data.

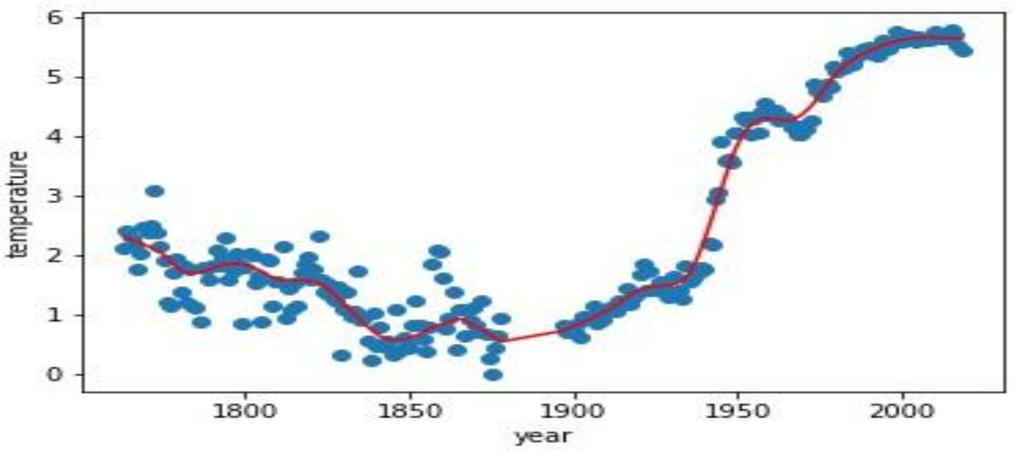


Figure 5. temperature anomaly over the years using average values.

As you can see, the graph follows the trend of the graph generated by NASA; however, the exact temperature anomalies are totally different.

Second approach

The second approach is to calculate the maximum value of the temperature for each year and then plot the temperature anomaly where the reference is the minimum value. The result of using this approach is illustrated in figure 6. The blue dots are actual values and the red line is lowess regression of the data.

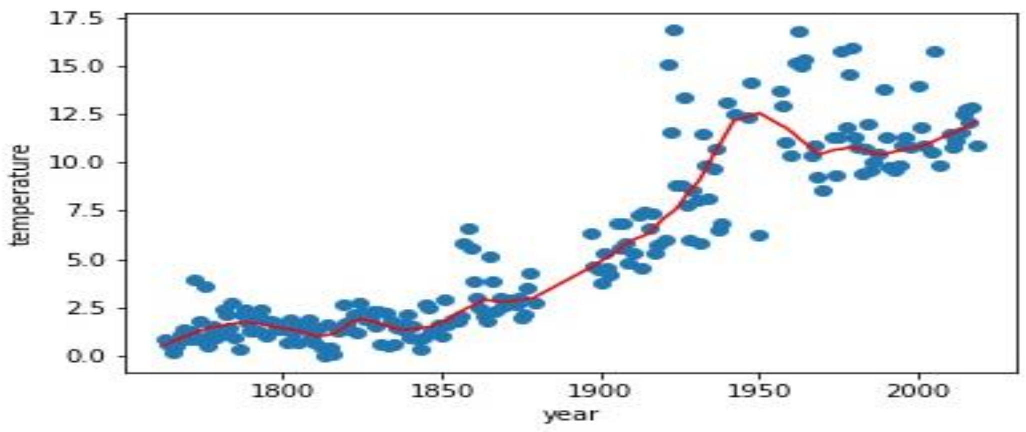


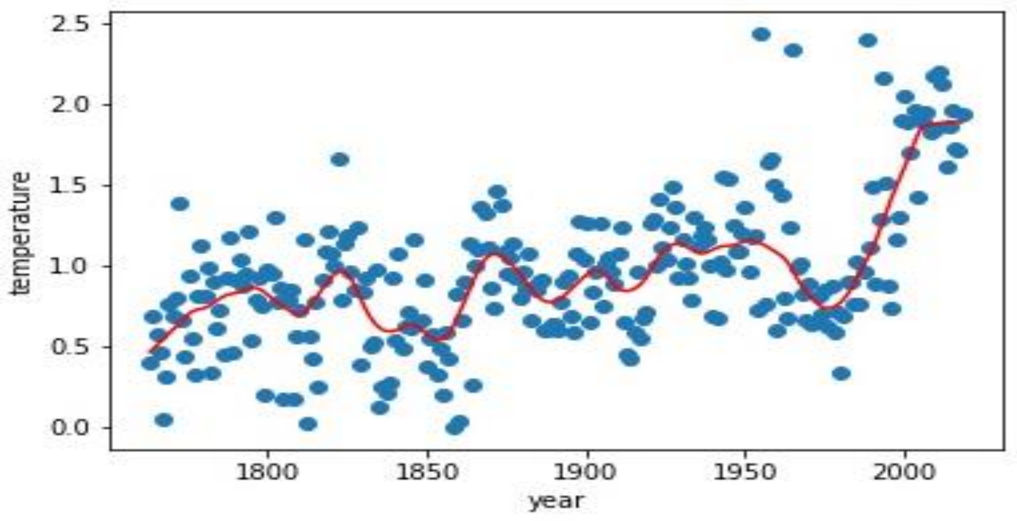
Figure 6. temperature anomaly over the years using maximum values.

Third approach

The third approach is to compare the daily temperature for each country with the temperature on the same day in the previous year. Afterwards, we find the average change in temperature per year per country; from which, we find the average change in temperature per year. This approach failed as there are rarely readings from the same country on the same day from one year to another.

Fourth approach

We focus on one of the countries and calculate the mean value of the temperature for each year and then calculate the temperature where the reference is the minimum value. We chose Italy as it is present in most of the files and more importantly at the very beginning. The result of using this approach is illustrated in figure 7. The blue dots are actual values and the red line is lowess regression of the data.



As you can see the result is quite puzzling for reasons we will discuss in the next section.

Sources of error

- The readings increase as time progresses which makes it hard to compare a computation done the observations from a certain year to another.
- The set on which the statistics are calculated changes from one year to another due to the addition of new countries non-present in previous years or the absence of countries that were present in previous years. For example, a certain year might have readings mostly from cold countries which makes the statistics lower than other years which have balanced readings.
- The dataset is non-uniform in every possible way. For example, in figure 7 we stick to computations on one country. For this country, some years have readings only in winter and other years have readings only in summer which makes the averages non-representative of the actual averages.

Enhancements and future work

There are many sources of non-uniformity in the dataset which makes inferring proper and reliable results very difficult. One approach is to fill the gaps in the data using interpolation methods; however, this approach is reliable for small-time windows like for example a couple of missing readings in a window of 10 days; however, it cannot deal with large time gaps. Another drawback of this approach is that undergoing interpolation for the whole data is very complex and it requires a lot of computational resources.