Assignment 1: MapReduce with Hadoop

Yao Song 301266041

Siyong Zhu 200136298

1 Word Count, Your First Hadoop Program

Writing and running the code

Question1

- *What will the output pairs look like*

➢ The output pair of the Map phase is a series of tuples, i.e. (“We”,1), (“all”, 1)…

- *What will be the types of keys and values of the input and output pairs in the Map phase?*

➢ The Mapper<LongWritable, Text, Text, IntWritable>, the 4 types correspond to input keys, values, output keys, values.

Question2

- *What will the input pairs look like?*

➢ The input pairs to Reducer are tuples like (“We”,[1,1,1,1,1,1,1]), (“all”,[1,1,1,1])…

- *What will be the types of keys and values of the input and output pairs in the Reduce phase?*

➢ The Reducer<Text, IntWritable, Text, IntWritable>, the 4 types correspond to input keys, values, output keys, values.

Question9

- *What is the most frequent word?*

➢ A

Question10

- *Why don’t we have a replication factor for directories?*

➢ Directories, i.e. file system and the metadata are stored on the namenode of the cluster. File blocks are replicated on different datanodes.

Question11

- *What command would you use to show the size of that file, in megabytes? How would you display its last kilobyte of text? How would you display its last five lines in an efficient manner?*

➢ Show file size in megabytes: **hdfs dfs -du –h *filepath***

- Display last kilobytes of the file to stdout: **hdfs dfs -tail *filepath***
- Display the last five lines: **hdfs dfs –tail *filapath* | tail -5**

Question12

- *How many Map and Reduce tasks did running Word Count on gutenberg-100M.txt produce? What is the link between the input size, the number of Map tasks, and the size of a block on HDFS?*

➢ 1 map and 1 reduce task.

If the size of the block on HDFS is fixed, default value is 134217728bytes, i.e. 128 MB, larger input size means more Map tasks.

If the block size is bigger than the input file size, there will be only one mapper. If the block size is smaller than the input file size, the input file will be split into multiple files such that

number_of_files * block_size >= input_file_size

Question13

- *How does changing the value of that parameter impact performance? Why?*

➢ The following table shows the running time for different input files on two different block sizes. It is clear that smaller block size will increase the split number and decrease the total running time for these running cases.

Input filesize(MB)	Split size(MB)	Time to run(min)	Number of splits	Time to run(ms)
100	128	1min48s	1	108295
100	64	1min23s	2	83005
100	32	1min11s	4	71091
100	16	1min03s	7	63150
200	128	2min25s	2	145865
200	64	1min55s	4	115720
500	128	3min50s	4	230390
500	64	3min11s	8	191911

2 MapReduce for Parallelizing Computations

Question1

- *How can we pass a different seed to initialize random numbers to each Map task, in order to make sure that no two Map tasks will work on the same values? What other parameter will we have to pass to each Map task? What will be the type of the keys and values of the input of Map tasks? What will they represent?*

➢ We can first write the seed value to a file for each Map task. Then ask the Map task to

read the seed value from that file. We also need to pass the value of number of samples to each Map task. The keys and values type of the Map tasks are both LongWritable. The first LongWritable is for the seed value, the second one is for the number of samples value.

Question2

- *What will the types of the keys and values output by Map tasks? What will they represent?*

➢ The Map tasks output types are IntWritable for both keys and values. The first one represent the sampled number. The second one is simply 1.

Question3

- *The Reduce task sums the results. What will the types of the keys and the values of the Reduce task be? What will they represent?*

➢ Reducer<IntWritable, IntWritable, IntWritable, IntWritable>
The types of the reduce task are all IntWritable. The meaning is as follows:
Reducer<sampled_value, [1,1,1,1,1,1,1,….], sampled_value, sampled_times_of_the_value>

Question9

- *How many accurate digits does your program find?*

➢ By running with the parameters 10 and 100000, the result I get is 2.717694. So I have 3 accurate digits.

Question10

- *How long is the Reducer phase? Does using a Combiner phase speed things up? Why?*

➢ Below is a table summarizing our experiment result:

Parameter1	Parameter2	With combiner(s)	Without combiner(s)
10	1000	31.079	31.344
10	10,000	30.691	30.334
10	100,000	30.471	30.544
10	1000,000	32.682	32.777
10	1000,0000	41.571	42.983

There is not any obvious performance improvement from the above results.

Question11

- *Does using another random generator improve the accuracy of the result?*

➢ From my experiments, it does not improve the accuracy of the result.

Java random	2.7180571	2.7176591	2.71808384	2.71838167
MersenneTwister	2.71835589	2.71826557	2.7183322	2.71837951

3 NCDC WEATHER DATA

Question1

- *What is wrong with Jeff's proposal?*

➢ Some temperature values in the dataset are missing. If we have one day with no max temperature value, there will be an invalid key value for the mapper output.

Question2

- *What will the key and values output by the Map tasks be? What types will they be?*

➢ Some temperature values in the dataset are missing. If we have one day with no max temperature value, there will be an invalid key value for the mapper output.

Question3

- *Can the Mapper produce a key/value pair from a single input? How can we solve this issue?*

➢ Some temperature values in the dataset are missing. If we have one day with no max temperature value, there will be an invalid key value for the mapper output.

Question4

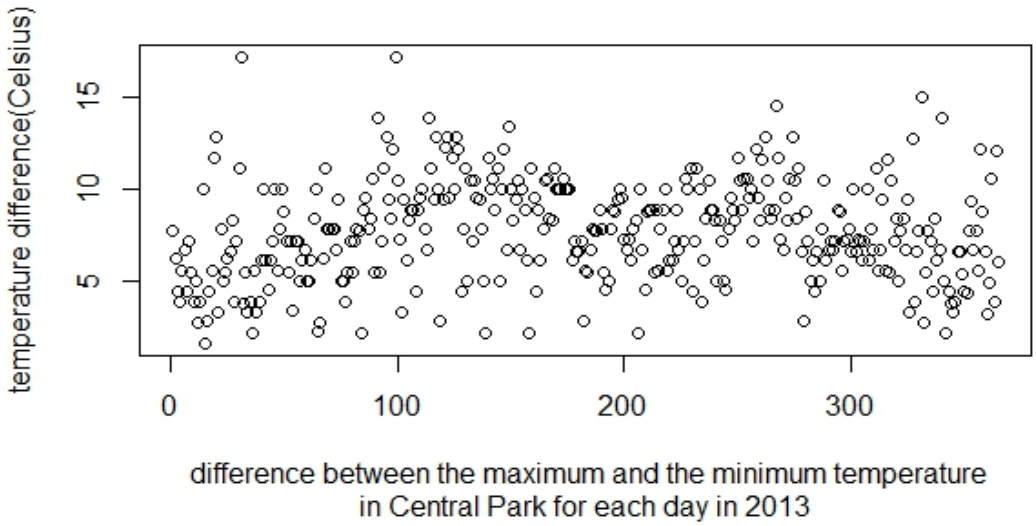
- *By following this approach, what work will be left to the Reduce task?*

➢ The reduce task will do nothing. It will just output the result from the map task. The default reducer simply writes its input to its output.

Question6

- *Plot the results.*

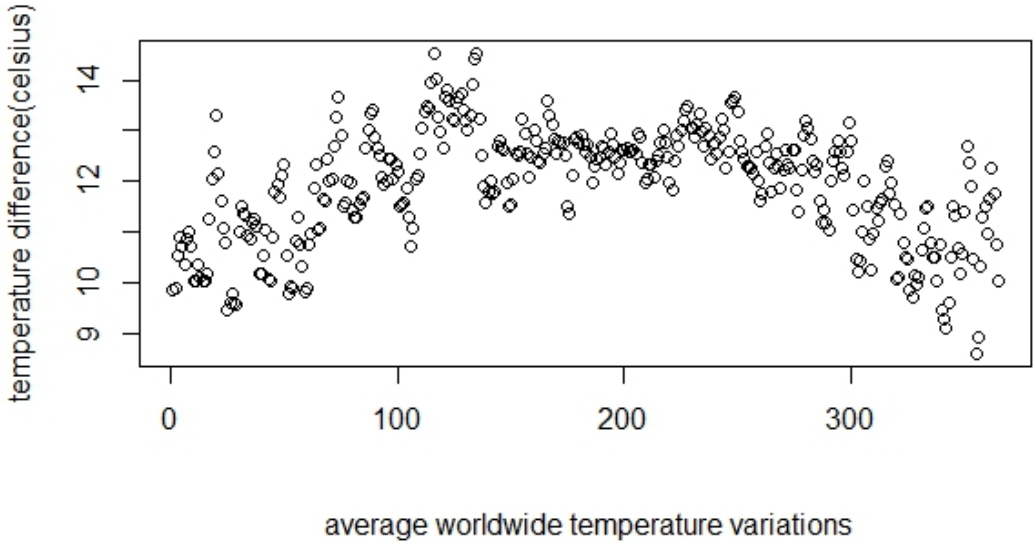
➢ See the figure below.



Question7

- *Plot the results.*

➢ See the figure below.



Question8

- *Reduce the split size by a factor of ten, Use the diff command to compare the results you get with the ones from the previous question. Are there differences? Why?*

➢ From the experiment result, if the split size is divided by 10, there is no difference in result. But if the split size is divided by 20 or a bigger number, there will be numerous differences in result. The reason is the records from the same day may be partitioned into different file splits, which results in wrong result.

Question9

- *Custom recordreader?*

➢ NCDCRecordReader creates custom map input key/value pairs. Key is still the offset. Value is all the lines that corresponding to exact one station in one day, instead of multiple lines for the same station and same day. File Splitter will check previous record from previous split to make sure it only starts from a new record that has been read from previous split.

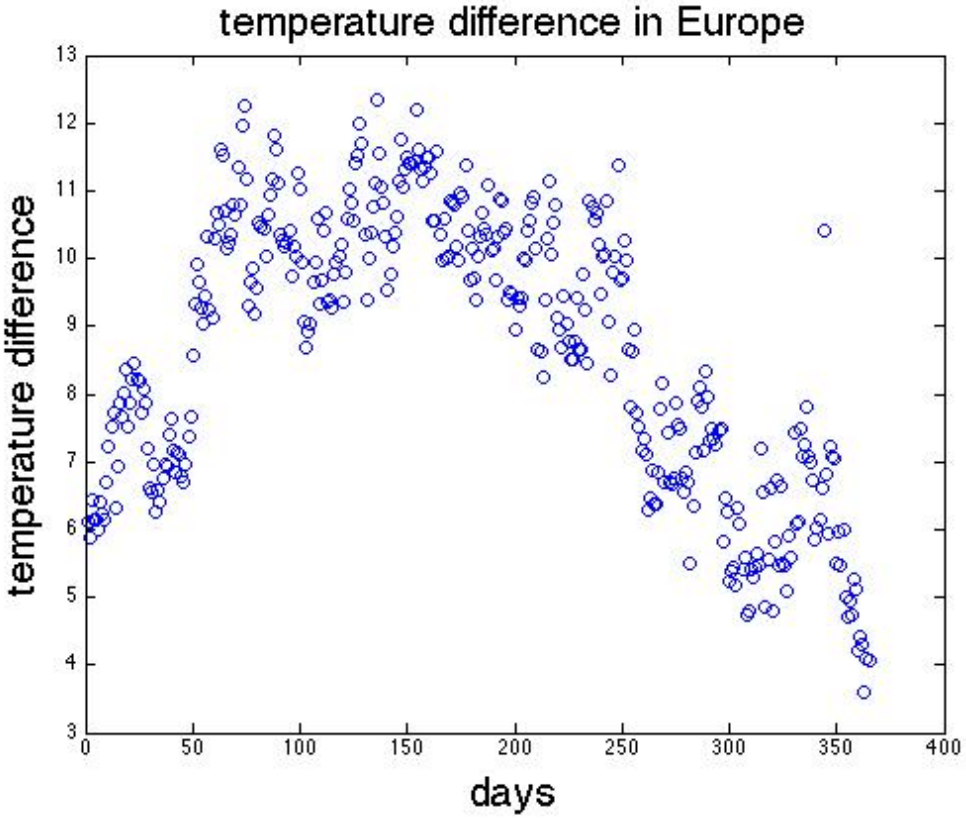
Using the custom RecordReader, we try to calculate the average temperature difference with two different split size: default 8 splits, and 10 times smaller size with 78 splits. The final results are the same. Results can be seen in current_o and current_n folders in the provided zip file (detail from readme)

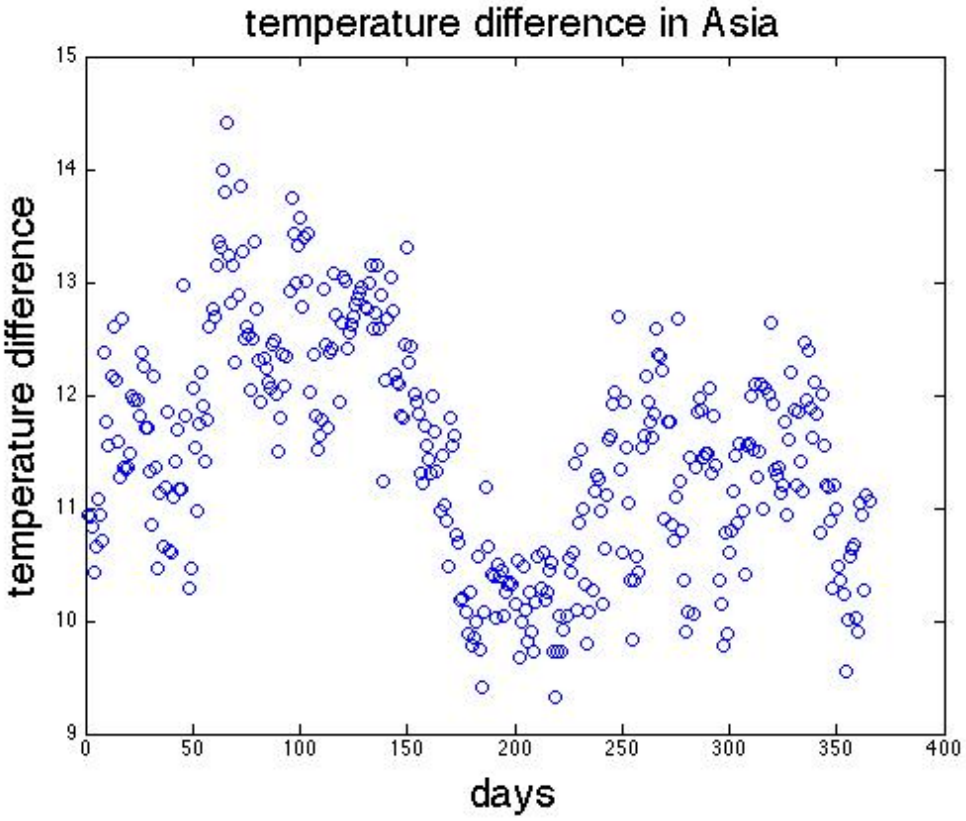
Question10

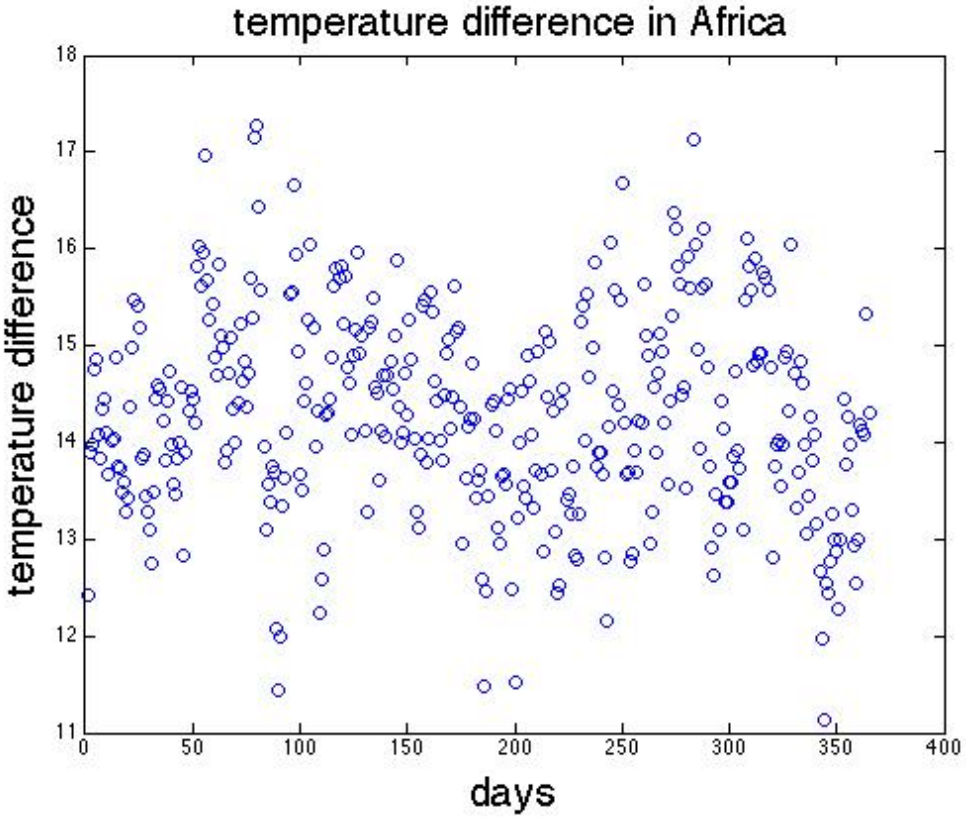
- *Rare the keys we've been using until now satisfactory?*

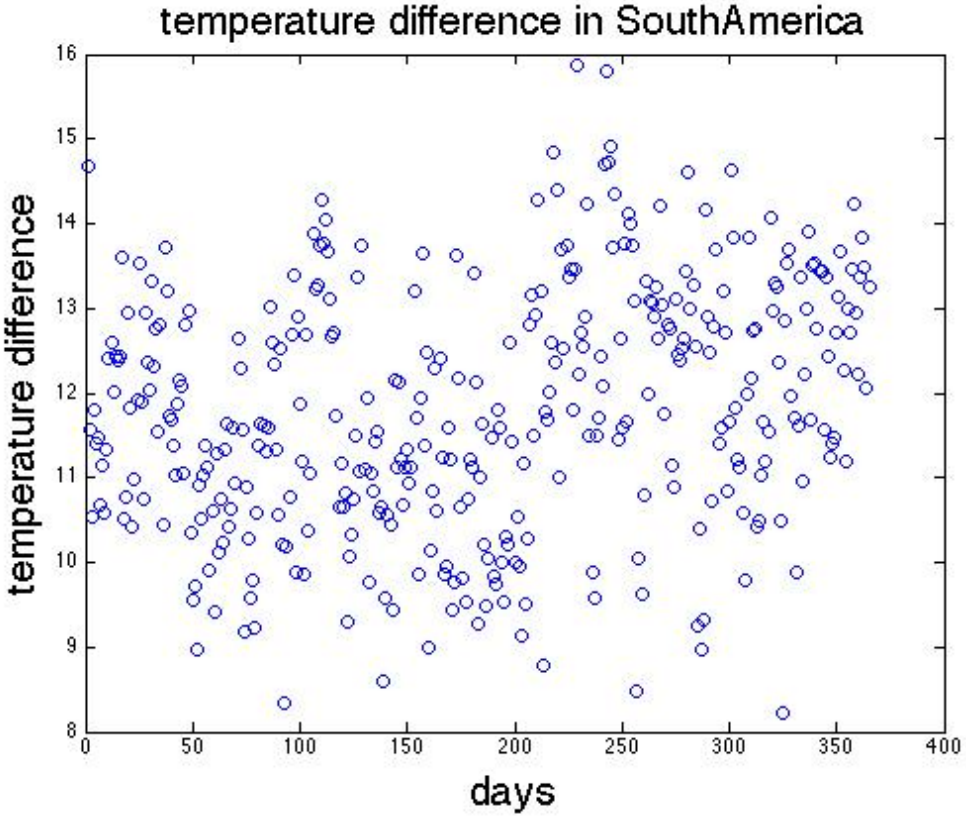
➢ No, our key is only date so far. But we need more information to indicate the continent. Thus our key need to contain both station location and date as key.

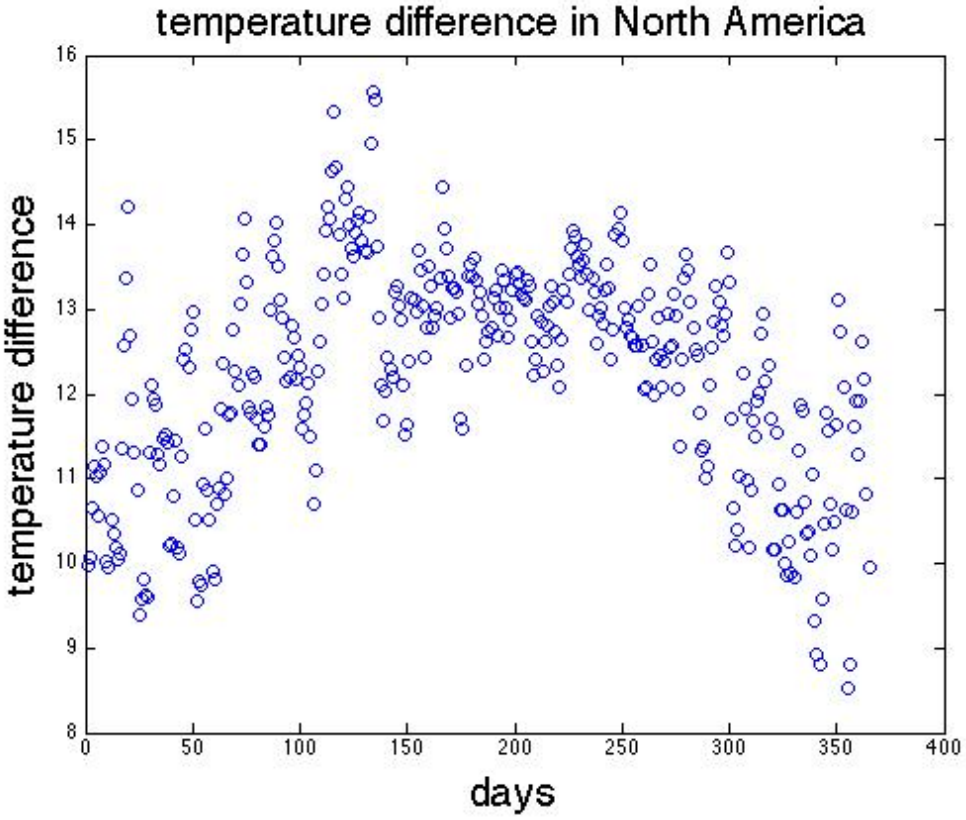
Question12

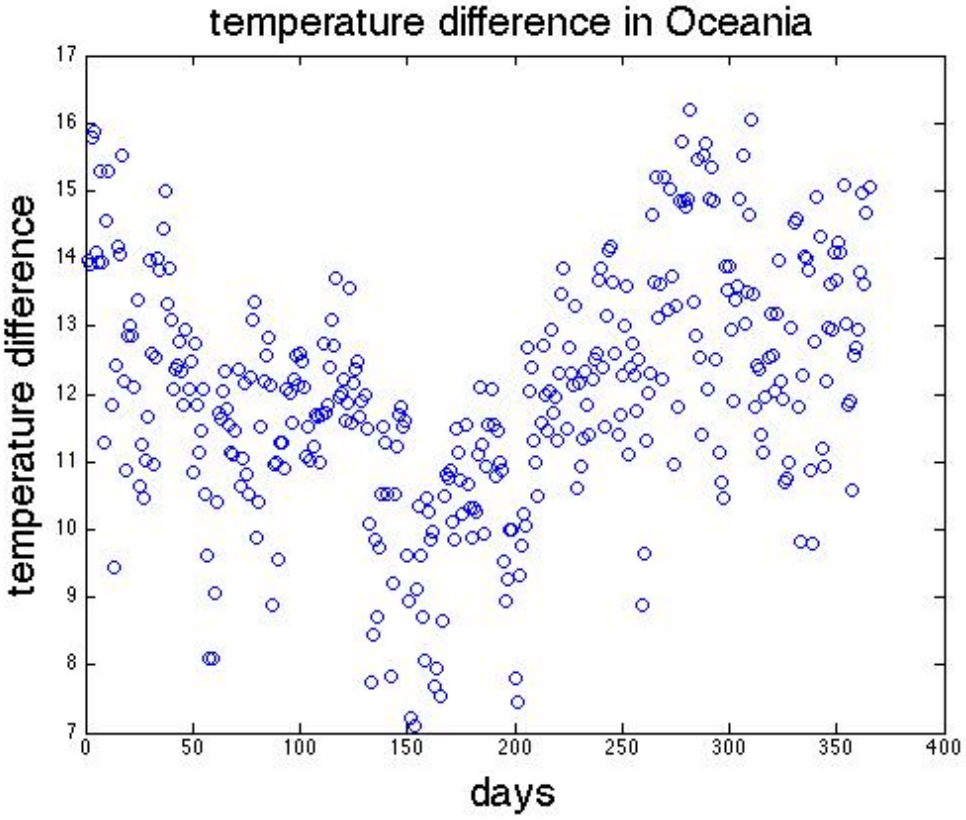

temperature difference in Europe
temperature difference
days
13
12
11
10
9
8
7
6
5
4
3
0
50
100
150
200
250
300
350
400


temperature difference in Asia
temperature difference
days
15
14
13
12
11
10
9
0
50
100
150
200
250
300
350
400


temperature difference in Africa
temperature difference
days
18
17
16
15
14
13
12
11
0
50
100
150
200
250
300
350
400


temperature difference in SouthAmerica
temperature difference
days
16
15
14
13
12
11
10
9
8
0
50
100
150
200
250
300
350
400


temperature difference in North America
temperature difference
days
16
15
14
13
12
11
10
9
8
0
50
100
150
200
250
300
350
400



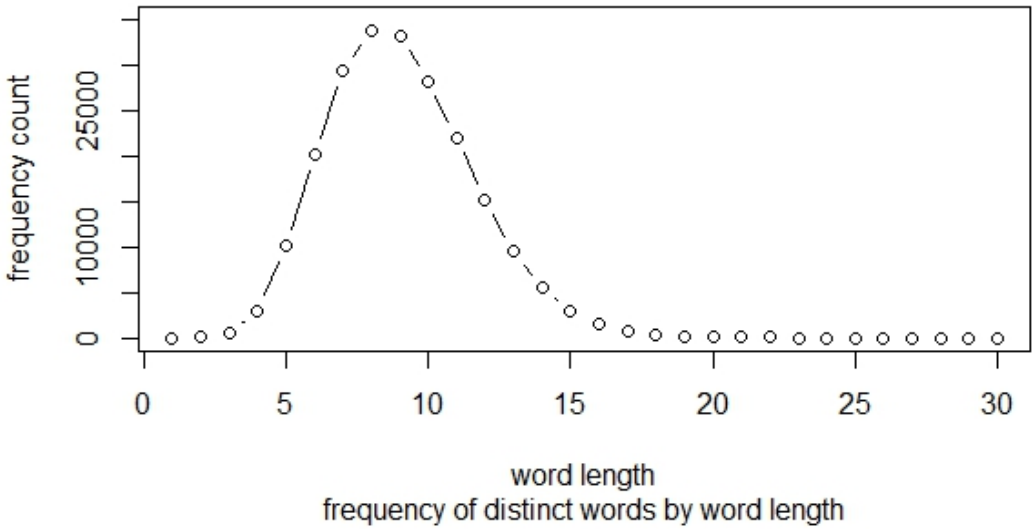
4 Back to counting

Question1

- *Counts the number of words of each length*

➢ See the figure below. The shape is similar to the one in the paper.

However, our text input has a smaller vocabulary than that of the paper.



Question2

- *calculates the frequency of each letter in a file*

➢ See the figure below.