CMPT459 Assignment 2

Huiyi Zou 301355563

Question 1

1. We can assume the measure is the number of tweets containing a token. N keywords as attributes and tweets as records. If the keyword appears k times in a tweet, assign k in that cell; otherwise, assign NULL. The COUNT(keyword) counts the number of tweets having the keyword.
   Q1: COUNT(keyword1) where keyword1=1 group by keyword2 and keyword3
   Q2: COUNT(keyword1) where keyword1=1 group by keyword2
   Q3: COUNT(keyword1) where keyword1=1 group by keyword3

   If there are 1000 different keywords in the data set, the data cube will have 1000 dimensions and the number of cuboids should be 2^{1000}.

2. The People dimension can have a concept hierarchy "person < family < human". We can query the number of cities each person visited using count(distinct city) from the photos set that grouped by person in People dimension.

3. There are dimension Text and dimension Image. Assume the Text dimension has a concept hierarchy "caption < initial letter of caption", and the Image dimension has a concept hierarchy "photo < types"
   Q1: photos grouped by initial letter of caption in the Text dimension
   Q2: photos grouped by caption in Text dimension
   Q3: photos grouped by types in Image dimension and caption in Text dimension

Question 2

Since there are 13 provinces and territories, we use 4-bit binary numbers to represent 13 different places. We can build a bitmap index using 4 bits (a nibble) per record. For n records, the bitmap index has 4n bits (n nibbles) and can be packed into $\left\lceil \frac{n}{2} \right\rceil$ bytes. From each 4-bit to the row-id: the j-th nibble of the p-th byte → row-id$= p * 2 + j$.

Example 1:
Set $total = 0$. Read a nibble each time in the bitmap. Get the row-id from the position of each nibble representing “BC”, get the Sales value of the relative row-id in table T, and update the sum by adding the Sales value.

Example 2:
Set $total = 0$. Read a nibble each time in the bitmap. Get the row-id from the position of each nibble representing “BC”, “ON”, or “NT”, get the Sales value of the relative row-id in table T, and update the sum by adding the Sales value.

Question 3

1. I compute the aggregate value of measure by different combinations of dimensions.

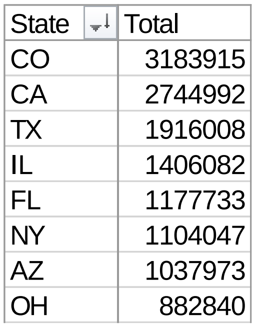
```
procedure  MapBaseTableToPostings(tuple)
    while not tuple.done() do
        record ← tuple.next()
        dim1 ← record.D1()
        dim2 ← record.D2()
        dim3 ← record.D3()
        dim4 ← record.D4()
        Emit(‘*’+‘*’+‘*’+‘*’, record.M())
        Emit(dim1+‘*’+‘*’+‘*’, record.M())
        Emit(‘*’+dim2+‘*’+‘*’, record.M())
        Emit(‘*’+‘*’+dim3+‘*’, record.M())
        Emit(‘*’+‘*’+‘*’+dim4, record.M())
        Emit(dim1+dim2+‘*’+‘*’, record.M())
        Emit(dim1+‘*’+dim3+‘*’, record.M())
        Emit(dim1+‘*’+‘*’+dim4, record.M())
        Emit(‘*’+dim2+dim3+‘*’, record.M())
        Emit(‘*’+dim2+‘*’+dim4, record.M())
        Emit(‘*’+‘*’+dim3+dim4, record.M())
        Emit(dim1+dim2+dim3+‘*’, record.M())
        Emit(dim1+dim2+‘*’+dim4, record.M())
        Emit(dim1+‘*’+dim3+dim4, record.M())
        Emit(‘*’+dim2+dim3+dim4, record.M())
        Emit(dim1+dim2+dim3+dim4, record.M())
    end while
end proceduce

procedure  ReducePostingsToAggregate(key, values)
    category ← key
    while not values.done() do
        measureResult ← aggregateFunction(values)
    end while
    Emit(category, measureResult)
end procedure
```

2. In the mappers there are 16 pairs emitted for each record. Therefore, if there are N tuples in the table T, there will be 16N key-value pairs the mappers emit.

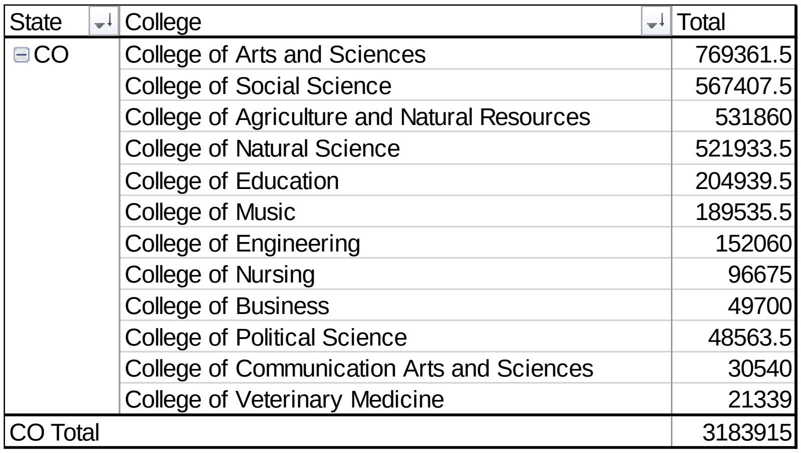
Question 4

1. The CO state has the largest total gift amount 3183915.
   I select the data in dimensions City, College, State, and Gift Amount to create a PivotTable in Microsoft Excel. I set the SUM of Gift Amount to summarize. And get the total Gift amount value in descending order grouped by dimensions.

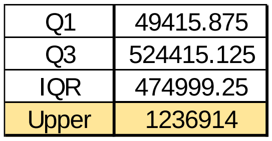


State	Total
CO	3183915
CA	2744992
TX	1916008
IL	1406082
FL	1177733
NY	1104047
AZ	1037973
OH	882840

(1) The total gift amount of state CO grouped by **College**:



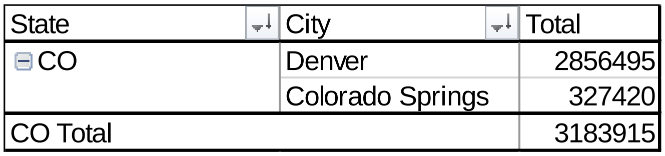
State	College	Total
⊟CO	College of Arts and Sciences	769361.5
	College of Social Science	567407.5
	College of Agriculture and Natural Resources	531860
	College of Natural Science	521933.5
	College of Education	204939.5
	College of Music	189535.5
	College of Engineering	152060
	College of Nursing	96675
	College of Business	49700
	College of Political Science	48563.5
	College of Communication Arts and Sciences	30540
	College of Veterinary Medicine	21339
CO Total		3183915



Q1	49415.875
Q3	524415.125
IQR	474999.25
Upper	1236914

I use the Interquartile Range Rule to check whether there is an exceptionally large value, where $IQR = Q3 - Q1$ and $upper\_limit = Q3 - 1.5 \times IQR$.
The amount of college with largest amount in CO state is not higher than upper limit calculated by IQR. It is not an outlier. Thus, the large total amount is not due to a college with exceptionally large total amount.

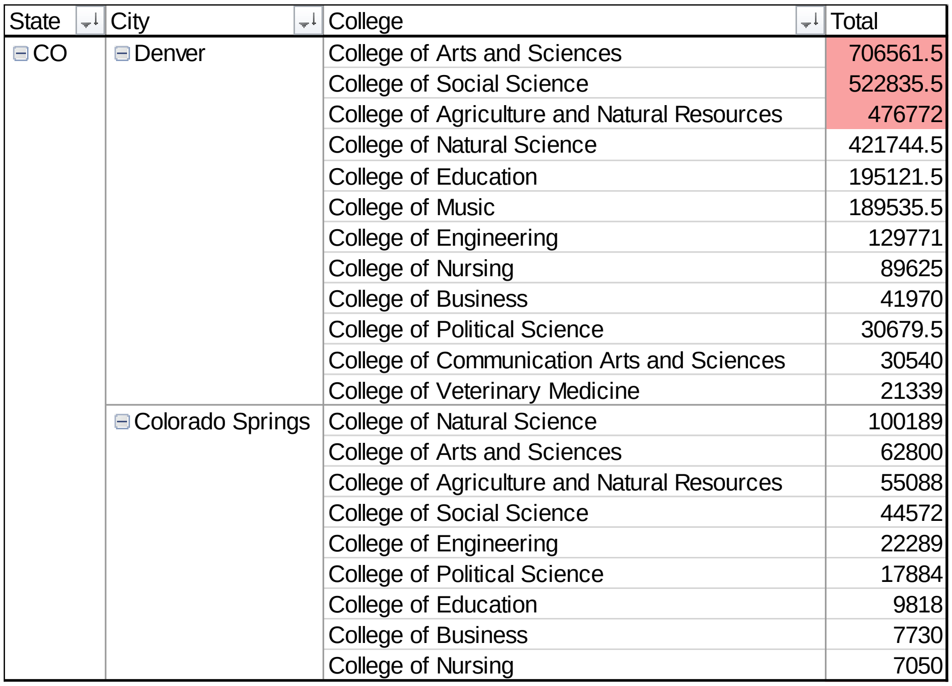
(2) The total gift amount of state CO grouped by **City**:



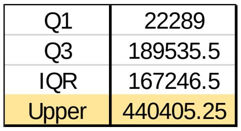
State	City	Total
⊟CO	Denver	2856495
	Colorado Springs	327420
CO Total		3183915

Since there are only two cities in the state CO, it is easy to see that the total gift amount of Denver is much larger than the total gift amount of Colorado Springs. Thus, the large total amount is due to a city with exceptionally large total amount.

(3) The total gift amount of state CO grouped by **College** and **City**:



State	City	College	Total
⊟ CO	⊟ Denver	College of Arts and Sciences	706561.5
		College of Social Science	522835.5
		College of Agriculture and Natural Resources	476772
		College of Natural Science	421744.5
		College of Education	195121.5
		College of Music	189535.5
		College of Engineering	129771
		College of Nursing	89625
		College of Business	41970
		College of Political Science	30679.5
		College of Communication Arts and Sciences	30540
		College of Veterinary Medicine	21339
	⊟ Colorado Springs	College of Natural Science	100189
		College of Arts and Sciences	62800
		College of Agriculture and Natural Resources	55088
		College of Social Science	44572
		College of Engineering	22289
		College of Political Science	17884
		College of Education	9818
		College of Business	7730
		College of Nursing	7050



Q1	22289
Q3	189535.5
IQR	167246.5
Upper	440405.25

I use the Interquartile Range Rule to check whether there is an exceptionally large value, where $IQR = Q3 - Q1$ and $upper\_limit = Q3 - 1.5 \times IQR$.
There are 3 tuples which have total amount larger than upper limit, which means that there are three outliers. Thus, the large total amount is not due to only one city and one college with exceptionally large total amount.

In conclusion, the large total amount is due to a city with exceptionally large total amount.

2. The total number of pairs that t.SUM()>=3*t'.SUM() is 151285.
   I write the result to the ‘result.txt’ file.
   The code in ‘code.py’ is built on top of the program on the web:
   https://programmersought.com/article/23211379442/;jsessionid=1F49F81E3D95A8A13C
   I use the BUC method to form the unique combination of dimensions and build a list $combination$ with 7 sub-lists for data cube. The sub-list $combination[0]$ store (*,*,*,*,*,*) with the total amount, and the other lists $combination[dim]$ store combinations with specific dimension number dim and its amount. To find the pairs, for each parent in sub-list $combination[dim]$ compare its amount with all children amount in sub-list $combination[dim+1]$. If pair (parent, child) satisfy the condition, add the pair to the list $output$. And write all pairs in $ouput$ to file.