Final Competition Project

Theory of Computation

2014

Instructor: Kun-Ta Chuang, email: ktchuang@mail.ncku.edu.tw.

Important Note:

Please remember to send your project before 7/3 11:59pm.

You are not allowed to revise the file after the deadline.

Please upload your program complying with these rules:

1. We will pull your project, TocFinal, from your github account.
2. The competition can only be written by JAVA
3. Note that your code will be ran at ubuntu 12.04.2 LTS
    1. 140.116.86.90 port:10122
    2. java version 1.7.0_51
4. Please put “all” your source files at the "Final" directory (you should create the directory at first). Note that TA will execute a clean build and remove all .class files before execution.
    - Please put the main function in "TocFinal.java”.
    - Make sure your code can be successfully complied by the ANT build file located in /home/TOC/ANT/final/build.xml.
    - Execute the make process by ant, and you should find a “TocFinal.jar” created. Try to execute “java –jar TocFinal.jar” and check if the result follows your expectation.
    - Usage: “ant -buildfile /home/TOC/ANT/final/build.xml build -Darg 學號” for creating the TocFinal.jar file in the final directory.
5. You program must accept three arguments, which are “Data_Source_URL” and “top-k” and “L-combination”. If arguments are wrongly given, please show messages and reject the execution.
6. Usage: " ant -buildfile /home/TOC/ANT/final/build.xml build -Darg 學號” for creating the TocFinal.jar file in the final directory.
    - "java –jar TocFinal.jar Data_Source_URL top-k L-combination " for running your code.
7. Write your basic information in the start of the source code, including your name, your student id number and brief description of your code.
8. If you have questions, please email to the TA at bohengorz@gmail.com.

Final Project:

給定一個Json格式的data source url，裡頭包含了10,000筆以上房屋交易資料，請你從此text file的資料中找出”以出現次數來說”， 前k個L-combination(2<=L<=4)。請注意程式需自行判斷data source

url的交易資料欄位個數(4<=欄位數<=10)。

ex: 欄位數為9時，L-combination(L=2)為(鄉鎮市區,交易月份)、(有無管理組織,車位類別)…等,一共有36($C_2^9 = 36$)種欄位組合。

另:

假設資料集有三個欄位，分別為 C1, C2, C3。
C1 的值包含 a 和 b；C2 的值包含 c 和 d；C3 的值包含 e 和 f。
假設此資料集共有 8 筆資料，如下:
a,c,e
a,d,f
b,c,f
a,d,e
a,c,e
a,d,e
b,c,e
a,c,e

其 Top-k(k=3)的 L-combination(L=2)
統計結果如下:
a,c;2　　(出現次數)
a,e;4
c,e;3
a,d;3
a,f;1
d,f;1
b,c;2
b,f;1
c,f;1
d,e;2
b,e;1
最後 Output 為
a, e;4
c, e;3
a, d;3

換言之，假設前k個(k=3) L-combination(L=2)為(鄉鎮市區:文山區,交易年月:10212;600)、(主要用途:住家用,有無管理組織:有;580)、(都市土地使用分區:住,車位類別:null;500)，代表在此份資料中，沒有其他出現次數>500的2-combination

以下為房屋交易資料之範例檔：

data scheme： **http://www.datagarage.io/datasets/ktchuang/realprice/ZgfNqTFZOL**

data set including 10,000 transactions：

http://www.datagarage.io/api/5386c065e7259bb37d9270e5

Goal: 請試著在最快的時間內找出正確答案，答案只要1個L-combination不同,即為錯誤

注意:

Grading Policy

1. 不可執行或邏輯不正確或沒交或抄襲 – 總成績0分 (如印出一行"老師對不起,寫不出來,即為邏輯不符合project要求")
2. 程式可執行，邏輯正確，但答案不正確 – 總成績3分
3. 答案正確 – 總成績7分
4. 答案正確且比助教的baseline程式速度還快 – 總成績13分
5. 答案正確且為前30%執行結束的組 –總成績17分
6. 答案正確且為前3組最快執行結束的組 – 總成績25分

(如有抵觸，則priority以rule 1 > rule 2 > rule 3 > rule 4 > rule 5 > rule 6)

Arguments:

1. Data_Source_URL: Specify the source data

   For example, **http://www.datagarage.io/api/5386c065e7259bb37d9270e5**
2. Top-k value, for example, k=1000
3. Length of item combination L, 2<=L<=4.

Running Examples:

Input: java –jar TocProj.jar Data_Source_URL 3 2

Output:

格式為

attribute1:value1,attribute2:value2;count(筆數)

範例如下：(不是正確答案，we will announce the answer/execution time of TA program later.)

鄉鎮市區:文山區,交易年月:10212;600
主要用途:住家用,有無管理組織:有;580
都市土地使用分區:住,車位類別:null;500

Note:

1. Output的組合有一定的順序,請依照這些欄位出現的順序來排列,例如：
   鄉鎮市區:大安區,交易年月:10211;600 (正確)
   (車位類別:null,有無管理組織:有;500) (錯誤),此順序錯誤,有無管理組織應該在前面,所以更正後,(有無管理組織:有,車位類別:null;500),這樣的Output結果才是正確的。
2. Example: 現在要取k=3,統計的結果Top-1的有600筆,Top-2的有560筆,Top-3的有500筆,但同樣還有另一個組合也是500筆,這時候Output的結果應該要印出四筆結果,而非三筆,例如：

   鄉鎮市區:文山區,交易年月:10212;600
   主要用途:住家用,有無管理組織:有;580
   都市土地使用分區:住,車位類別:null;500
   鄉鎮市區:大同區,主要建材:鋼筋混凝土造;500

3. Output的結果請依照count(筆數)由大到小顯示。例如：

   鄉鎮市區:文山區,交易年月:10212;600
   主要用途:住家用,有無管理組織:有;580
   都市土地使用分區:住,車位類別:null;500
   (正確)

   都市土地使用分區:住,車位類別:null;500
   鄉鎮市區:文山區,交易年月:10212;600
   主要用途:住家用,有無管理組織:有;580
   (錯誤，沒有依照 count 由大到小排序)

4. 當combination的出現次數相同,且需要Output出來時,Output的順序,依照組合出現的優先順序而定,ex:
   統計完的結果(a,e)與(b,c)出現次數皆為10次,且是Top-k(k=3)的L-combination(L=2),需要Output出來,此時因為資料的順序為(a,b,e)、(a,d,e)、(b,c,f)…,所以Output的順序為先Output出(a,e),再Output(b,c),因為a,e在第一筆資料即出現,(b,c)在第三筆資料時才出現
5. JVM Heap Size的限制為1G memory

Finally:

When you submit your code to the server, please check if your code is also correct in the Linux environment.

When you cannot compile your code by our ant build name and you know the reason (for example, you use some external system library), please inform TA for asking the instruction.

Good luck!