MAT 259_HW2

Sila Dey
siladitya_dey@cs.ucsb.edu

Question:
I wanted to find out what the most popular keywords from 2005-2013 were, and how they varied across months. More concretely, I was curious if irrespective of the year there is a discernible relationship in the most popular keywords in one month with that in another month. One of the ways of determining the most popular keywords in a particular month is to query the most popular item being lent out that specific month, and then populate the keywords associated with this title. Of course, there needs to be a level of cleanup of data since words like ‘of’, ‘the’, ‘and’, numbers, etc don’t lend too much insight nor help in gaining any insight regarding a trend.

Data:
I used the tip that Prof Legrady provided last time about not using the spl2.inraw table, and instead used the other tables to query from. As is evident, I populated the Item/Bib number as an identifier to the title, and soon enough came to a conclusion that sorting by item number is not a good way to go as multiple item numbers can map to a single bib number. So instead I decided to sort via the bibliography number, and retrieved the most popular titles in a particular month. To decide the most lent out items, I had to determine the number of loans of the title and for this I restricted my query to count the number of times a specific title was lent out in a given month from 2005-2013.

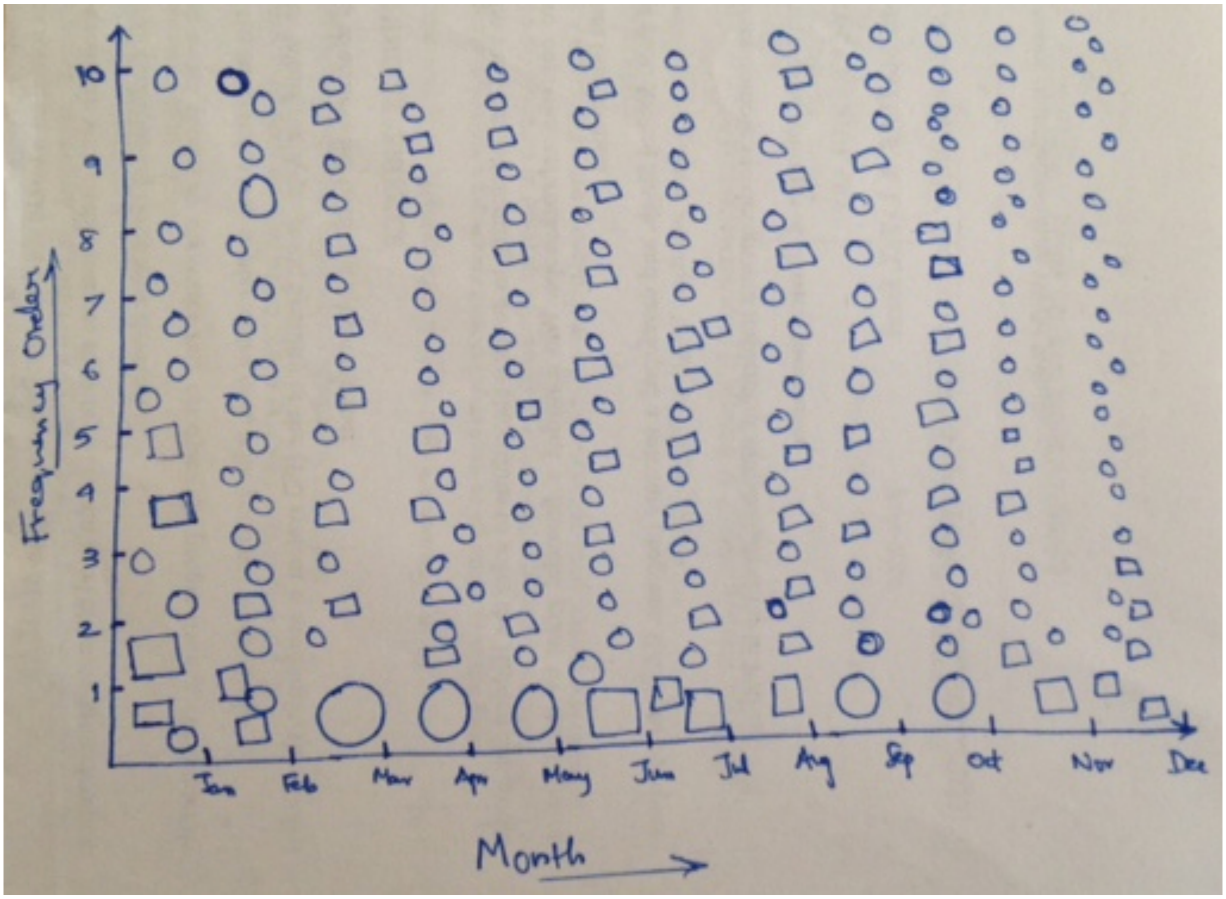
Query:

```
SELECT t1.item AS Item, t1.bib AS BibNo,
UPPER(SUBSTR(t5.kind,3)) AS Media_Type, t2.title AS Title,
COUNT(t1.o) AS Num_Loans,
t3.kywds AS Keywords
FROM spl2.activity AS t1
INNER JOIN spl2.title AS t2 ON t2.bib = t1.bib
INNER JOIN ( SELECT bib, GROUP_CONCAT(LOWER(keyword)) AS kywds
FROM spl2.keyword
WHERE LOWER(keyword) NOT REGEXP 'theIandIforInotIanyI^([0-9]+)I^([a-z]){1}$I^([a-
z]){2}$I '
GROUP BY bib) AS t3 ON t3.bib = t1.bib
INNER JOIN spl2.kind AS t5 ON t5.item = t1.item
WHERE t1.item > 0 AND YEAR(t1.i)>=2005 AND YEAR(t1.o)>=2005 AND
YEAR(t1.o)<2014
AND LOWER(t2.title) NOT REGEXP '^uncatalog+' AND MONTH(t1.o)=X
GROUP BY t1.bib
ORDER BY Num_Loans DESC
LIMIT 20
```

where **X** is a number from 1-12, denoting the month.

Processing Time for each query is between 120~140s.

Visualization:


Frequency Order
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month