CSC 370 — Database Systems
Summer 2015
Assignment No. 4
Version 1.1 (Jun 9, 2015)

Note 1 **This assignment is to be done individually**

Note 2 Working with other people is prohibited.

- Due date: June 12, 2015, 9:30.
- This assignment is worth 1% of your total course mark.
- Summit in paper your queries, and their corresponding relational algebra.
- Submit electronically the SQL queries in a single **text** file.

Objectives

After completing this assignment, you will have experience:

- Use aggregation.

Your task, should you choose to accept it

1. Answer the following questions, both in relational algebra, and SQL. **nRelational algebra queries should match SQL**. For SQL queries provide the query and the result. One query per question. Your query should only use the information provided in the question. **You cannot use the LIMIT keyword in your queries.**

   1. Let us analyze TV shows. In this case, we want to list the TV-shows with episodes that have ratings in at least 4 different seasons. To limit the number of tuples, let us narrow our search to those TV shows that have an average rank of their episodes > 8 (only the episodes) and have an average number of votes for their episodes of 1,000 votes or more. List the id of their production (episodeof), their average rank, their average number of votes, the number of episodes with ratings and the number of different seasons with ratings. Order the result by average rank descending first and in case of collisions by average number of votes descending second. Hint: use count(distinct attr) to count the number of different values of a given attribute.

      Solution:

      Create a table that contains the summary of the episodes of a series

      $$Summ = \gamma^{episodeof}_{avg(votes) \rightarrow avotes, count(distinct\ season) \rightarrow cseasons, count(*) \rightarrow cepisodes, avg(rank) \rightarrow arank}(E \bowtie R)$$

      Then select those with $avotes >= 1,000$ and $arank >= 8$ then project:

      $$\Pi_{episodeof,arank,avotes,cepisodes,cseasons}\sigma_{avotes>=1000\ and\ cseasons=4\ and\ arank>=8}Summ$$

```
WITH summ AS (
    SELECT episodeof, avg(votes) as avotes, count(distinct season) as cseasons,
           count(*) as cepisodes, avg(rank) as arank
    FROM episodes NATURAL JOIN ratings
    GROUP BY episodeof)
SELECT episodeof, arank, avotes, cepisodes, cseasons
FROM summ
WHERE avotes >= 1000 AND cseasons >= 4 AND arank >= 8.0
ORDER BY arank desc;
```

```
| episodeof                         |            arank |                avotes | cepisodes | cseasons |
|-----------------------------------+------------------+-----------------------+-----------+----------|
| "Person of Interest" (2011)       | 8.99186046511628 | 1168.7441860465116279 |        86 |        4 |
| "House M.D." (2004)               | 8.68579545454545 | 1178.9659090909090909 |       176 |        8 |
| "Supernatural" (2005)             | 8.65446009389671 | 1557.0000000000000000 |       213 |       10 |
| "Lost" (2004)                     | 8.60762711864407 | 2280.9067796610169492 |       118 |        6 |
| "Dexter" (2006)                   |          8.58125 | 2316.4895833333333333 |        96 |        8 |
| "Prison Break" (2005)             | 8.58024691358025 | 1177.5308641975308642 |        81 |        4 |
| "The Sopranos" (1999)             | 8.57674418604651 | 1673.6162790697674419 |        86 |        6 |
| "Game of Thrones" (2011)          |           8.5275 | 9614.2250000000000000 |        40 |        4 |
| "Friends" (1994)                  | 8.50635593220339 | 1089.3347457627118644 |       236 |       10 |
| "The Walking Dead" (2010)         | 8.46119402985075 | 5075.4029850746268657 |        67 |        5 |
| "Homeland" (2011)                 | 8.41041666666667 | 1670.0000000000000000 |        48 |        4 |
| "Fringe" (2008)                   |            8.393 | 1049.4300000000000000 |       100 |        5 |
| "Breaking Bad" (2008)             | 8.37903225806452 | 8616.6935483870967742 |        62 |        5 |
| "Community" (2009)                | 8.35544554455445 | 1039.3267326732673267 |       101 |        6 |
| "The Big Bang Theory" (2007)      | 8.14972067039107 | 1323.4189944134078212 |       179 |        8 |
| "How I Met Your Mother" (2005)    | 8.12884615384615 | 1446.8846153846153846 |       208 |        9 |
| "Buffy the Vampire Slayer" (1997) | 8.09103448275862 | 1138.9724137931034483 |       145 |        7 |
| "South Park" (1997)               | 8.08560311284046 | 1079.7470817120622568 |       257 |       18 |
| "American Horror Story" (2011)    | 8.06666666666667 | 1656.3137254901960784 |        51 |        4 |
| "Doctor Who" (2005)               | 8.04887218045113 | 2421.2556390977443609 |       133 |        8 |
| "The X Files" (1993)              | 8.03681592039801 | 1252.2985074626865672 |       201 |        9 |
(21 rows)
```

2. Of the movies with at least 50,000 votes, list the one(s) with the highest rank. List its title, year, rank, and votes.

   Solution:

   Find movies with at least 50k votes and with their ratings:

$$A = \sigma_{\text{attr is NULL} \wedge votes >= 50{,}000} Ra \bowtie P$$

   Find the maximum rank:

$$M = \gamma_{max(rank) \rightarrow rank} A$$

   Join to find which movie has that rank:

$$\Pi_{title,year,rank,vote} A \bowtie M$$

```
WITH A as (select * from ratings
           natural join
           productions where attr is NULL and votes >= 50000),
     M as (select max(rank) as rank from A)
select title, year, rank, votes from A natural join M ;
```

```
| title                    | year | rank |   votes |
|--------------------------+------+------+---------|
| The Shawshank Redemption | 1994 |  9.3 | 1424596 |
(1 row)
```

3. This query is restricted to movies with a rank of at least 8 and at least 50,000 votes. Find the **pid** of persons who have been in at least 10 of these movies. List their **pid**, number of such movies, and average rating of such movies. Order by average rank.

Solution:
We can start by joining Productions, Ratings and Roles

$$A = P \bowtie R \bowtie Ra$$

Select only those with 50k votes or more and rank $>=$ 8:

$$B = \sigma_{votes>=50000\ AND\ rank>=8}A$$

Now compute the count, sum, and average:

$$C = \gamma^{pid}_{count(id)\rightarrow count, avg(rank)\rightarrow avg}B$$

and narrow it to those with at least 10 movies

$$\sigma_{count>=8}C$$

and since we don't care about order-by in relational algebra, we are done.

```
WITH A as (SELECT * from productions
                NATURAL JOIN
             ratings
                NATURAL JOIN
           roles),
     B as (SELECT * from A
           WHERE attr IS NULL and
                votes >= 50000 and
                rank >= 8),
     C as (select pid, count(id) as count, avg(rank) as avg FROM B
          GROUP BY pid)
SELECT * from C where count >= 10
ORDER by avg
```

```
| pid                | count |              avg |
|--------------------+-------+------------------|
| Tovey, Arthur      |    10 |             8.27 |
| Lynn, Sherry (I)   |    11 | 8.27272727272727 |
| Jackson, Samuel L. |    10 |             8.28 |
| Ratzenberger, John |    11 | 8.29090909090909 |
| Flowers, Bess      |    12 | 8.30833333333333 |
| De Niro, Robert    |    10 |             8.36 |
(6 rows)
```

4. For movies with at least 50,000 votes, and rank of at least 7.1 (the median for the rank of movies with at least 50,000 episodes): list the person (or persons) that has appeared in the most of such movies, the id of the movie, their billing, and their character. Result should contain pid, id, billing, and character.

   Solution:
   First we compute A, a big table with the roles of movies that have had at least 50,000 votes:

$$A = \sigma_{attr\ is\ NULL\ AND\ votes>=50{,}000\ AND\ rank>=7.1}(P \bowtie R \bowtie Ra)$$

the number of times that a person has appeared in these movies:

$$B = \gamma^{pid}_{count(id) \rightarrow count} A$$

Now we find the maximum count. Note that we rename the attribute to count (same as in table B):

$$M = \gamma_{max(count) \rightarrow count} B$$

Now we can simply join M with B and A:

$$\Pi_{pid,id,billing,character}(M \bowtie B \bowtie A)$$

```
WITH
  A as (
     SELECT * from
     roles NATURAL JOIN productions NATURAL JOIN ratings
     WHERE attr is NULL and votes >= 50000
  ),
  B as (
     SELECT pid, count(id) as count
     FROM A
     GROUP BY pid
  ),
  M as (
     SELECT max(count) as count FROM B
  )
  SELECT pid, id, billing, character
  FROM M NATURAL JOIN B NATURAL JOIN A
```

```
| pid             | id                                 | billing | character                 |
|-----------------+------------------------------------+---------+---------------------------|
| De Niro, Robert | A Bronx Tale (1993)                |       1 | Lorenzo                   |
| De Niro, Robert | American Hustle (2013)             |         | Victor Tellegio           |
| De Niro, Robert | Angel Heart (1987)                 |       2 | Louis Cyphre              |
| De Niro, Robert | Awakenings (1990)                  |       1 | Leonard Lowe              |
| De Niro, Robert | Brazil (1985)                      |       2 | Harry Tuttle              |
| De Niro, Robert | Cape Fear (1991)                   |       1 | Max Cady                  |
| De Niro, Robert | Casino (1995)                      |       1 | Sam ’Ace’ Rothstein       |
| De Niro, Robert | Fahrenheit 9/11 (2004)             |         | Himself                   |
| De Niro, Robert | Goodfellas (1990)                  |       1 | James Conway              |
| De Niro, Robert | Heat (1995)                        |       2 | Neil McCauley             |
| De Niro, Robert | Jackie Brown (1997)                |       6 | Louis Gara                |
| De Niro, Robert | Limitless (2011/I)                 |       2 | Carl Van Loon             |
| De Niro, Robert | Mean Streets (1973)                |       1 | Johnny Boy                |
| De Niro, Robert | Men of Honor (2000)                |       1 | Master Chief Billy Sunday |
| De Niro, Robert | Once Upon a Time in America (1984) |       1 | David ’Noodles’ Aaronson  |
| De Niro, Robert | Raging Bull (1980)                 |       1 | Jake La Motta             |
| De Niro, Robert | Ronin (1998)                       |       1 | Sam                       |
| De Niro, Robert | Silver Linings Playbook (2012)     |       3 | Pat Sr.                   |
| De Niro, Robert | Sleepers (1996)                    |       3 | Father Bobby              |
| De Niro, Robert | Stardust (2007)                    |      34 | Captain Shakespeare       |
| De Niro, Robert | Taxi Driver (1976)                 |      10 | Travis Bickle             |
| De Niro, Robert | The Deer Hunter (1978)             |       1 | Michael                   |
| De Niro, Robert | The Godfather: Part II (1974)      |       4 | Vito Corleone             |
| De Niro, Robert | The Untouchables (1987)            |       5 | Al Capone                 |
| De Niro, Robert | Wag the Dog (1997)                 |       2 | Conrad Brean              |
(25 rows)
```

5. For this question consider only movies with at least 50,000 votes. Movies with at least 50,000 votes and a rank > 8 are usually very good; let us call these movies *good movies* Some directors are really good, others are lucky. For every director who has directed at least 10 movies (regardless rating), but had directed at least one *good movies*, display his/her pid, the total number of movies made, the percentage of movies with rank > 8, the number of movies with rank > 8, their average

ranking, the number of movies below this ranking, and their average, and compute the difference between the average of the good ones minus the average of the rest. Order by the difference (descending) first, and in the case of same difference, by percentage of *good movies* (descending). Your result should contain 8 columns: pid, number of movies made (total), total of good ones (goodones), their average rank (avggoodones), the number of other movies (rest), their average rank (avgrest) and the difference of agvgoodones minus avgrest. Note the formatting of the output, you can achieve it by using the `to_char` function in postgresql.

Solution:

First let us find movies (with their rank and director) with at least 50k votes

$$M = \Pi_{id,pid,rank}\sigma_{votes>=50{,}000\ and\ attr\ IS\ NULL}(P \bowtie Ra \bowtie D)$$

Now we can now do the aggregation of good ones

$$G = \gamma^{pid}_{count(id)\rightarrow goodones, avg(rank)\rightarrow avggoodones}\sigma_{rank>=8}M$$

And now let us do the rest of the movies directed by the directors in G, including the aggregation:

$$R = \gamma^{pid}_{count(id)\rightarrow rest, avg(rank)\rightarrow avgrest}\sigma_{rank<8}(M \bowtie (\Pi_{pid}G)))$$

Finally join R and G, where their total number of movies is at least 5. *List* is:

$$\begin{aligned} List =& pid, goodones * 100.0/(rest + goodones) \rightarrow prop,\\ & rest + goodones \rightarrow total,\\ & goodones, avggoodones, rest, avgrest,\\ & avggoodones - avgrest \rightarrow diff \end{aligned}$$

$$\Pi_{List}\sigma_{goodones+rest>=10}(G \bowtie R)$$

```
WITH M AS (SELECT id, pid, rank
            FROM Productions NATURAL JOIN Directors NATURAL JOIN Ratings
           WHERE attr is NULL and votes >= 50000),
     G AS (SELECT pid, count(*) as goodones, avg(rank) as avggoodones
            FROM M where rank > 8
            GROUP BY pid),
     A AS (SELECT pid, count(*) as rest, avg(rank) as avgrest
            from M NATURAL JOIN (select pid from G) as rip
            WHERE rank <= 8
            GROUP by pid)
     SELECT pid, to_char(goodones*100.0/(rest+goodones), '99D9') || '%' as prop,
            rest+goodones as total,
            goodones, to_char(avggoodones, '99D9') as avggoodones,
            rest, to_char(avgrest, '99D9') as avgrest,
            to_char(avggoodones-avgrest, '99D9') as diff
     FROM A NATURAL JOIN G
     WHERE (goodones + rest) >= 10
     ORDER BY avggoodones-avgrest DESC, prop desc;
```

```
| pid                    |  prop | total | goodones | avggoodones | rest | avgrest | diff |
|------------------------+-------+-------+----------+-------------+------+---------+------|
| Hitchcock, Alfred (I)  | 90.0% |    10 |        9 |         8.3 |    1 |     7.8 |   .5 |
| Miyazaki, Hayao        | 75.0% |     8 |        6 |         8.3 |    2 |     7.8 |   .5 |
| Kubrick, Stanley       | 72.7% |    11 |        8 |         8.3 |    3 |     7.7 |   .7 |
| Fincher, David         | 30.0% |    10 |        3 |         8.6 |    7 |     7.5 |  1.1 |
| Jackson, Peter (I)     | 30.0% |     10 |         3 |          8.8 |     7 |      7.5 |  1.3 |
| Lynch, David (I)       | 12.5% |      8 |         1 |          8.2 |     7 |      7.5 |   .7 |
| Nolan, Christopher (I) | 77.8% |      9 |         7 |          8.6 |     2 |      7.4 |  1.2 |
| Tarantino, Quentin     | 54.5% |     11 |         6 |          8.4 |     5 |      7.4 |  1.0 |
| Scorsese, Martin (I)   | 53.8% |     13 |         7 |          8.3 |     6 |      7.4 |  1.0 |
| Gilliam, Terry         | 25.0% |      8 |         2 |          8.2 |     6 |      7.3 |   .9 |
| Eastwood, Clint        | 23.1% |     13 |         3 |          8.2 |    10 |      7.3 |   .9 |
| Spielberg, Steven      | 20.8% |     24 |         5 |          8.5 |    19 |      7.3 |  1.2 |
| Coen, Ethan            | 17.6% |     17 |         3 |          8.2 |    14 |      7.3 |   .9 |
| Coen, Joel             | 17.6% |     17 |         3 |          8.2 |    14 |      7.3 |   .9 |
| Allen, Woody           | 11.1% |      9 |         1 |          8.1 |     8 |      7.3 |   .8 |
| Zemeckis, Robert       | 14.3% |     14 |         2 |          8.7 |    12 |      7.1 |  1.6 |
| Howard, Ron (I)        | 18.2% |     11 |         2 |          8.2 |     9 |      7.0 |  1.2 |
| Scott, Ridley          | 18.8% |     16 |         3 |          8.4 |    13 |      6.9 |  1.5 |
| Verbinski, Gore        | 12.5% |      8 |         1 |          8.1 |     7 |      6.9 |  1.2 |
| Stone, Oliver (I)      |  9.1% |     11 |         1 |          8.1 |    10 |      6.8 |  1.3 |
| Rodriguez, Robert (I)  |  9.1% |     11 |         1 |          8.1 |    10 |      6.8 |  1.3 |
| De Palma, Brian        | 12.5% |      8 |         1 |          8.3 |     7 |      6.8 |  1.5 |
| Petersen, Wolfgang     | 12.5% |      8 |         1 |          8.4 |     7 |      6.7 |  1.7 |
| Shyamalan, M. Night    | 12.5% |      8 |         1 |          8.2 |     7 |      5.8 |  2.4 |
(24 rows)
```

What to submit

In paper submit your the Relational Algebra, the SQL queries and the results (you can limit your results to the first 10 tuples). Electronically, submit your SQL queries.