Assignment 2
Signature Indexes - Testing

Last updated: **Monday 12th April 10:32pm**
Most recent changes are shown in red;
older changes are shown in brown.
A changelog is at the end of the file.
Hopefully, this changelog will be very short.

intro   data   notes   examples   changelog

Testing Assignment 2

The goal of Assignment 2 is to implement signature-based indexing that allows users to answer partial-match retrieval queries without need to scan the entire data file.

There are three important aspects to solving this problem well:

- your code finds the correct matching tuples for a given query

- your code achieves this by reading as little data as possible

- your code inserts and queries data efficiently

These are in priority order: correctness is more important than minimising pge reads, which in turn is more important than time taken.

Test Data

In order to avoid issues with different random number generators on different machine, we supply pre-made data files. Since these are quite large, it is best *not* to copy them to your working directory for this assignment, unless you're working on your own machine, in which case you can grab all of the data files in ass2-data.zip.

The data files are all available in the directory `/web/cs9315/21T1/assignments/ass2/testing/`.

The individual data files were generated as

- `data1` generated via `./gendata 2000 3 1234567 13`

- `data2` generated by `./gendata 5000 5 1234321 29`

- `data3` generated by `./gendata 10000 4 7654321 23`

- `data4` generated by `./gendata 20000 5 4321234 7`

- `data5` generated by `./gendata 50000 3 7654567 3`

You will probably not get the same results if you generate the data on your own machine.

Notes on Expected Results

There is no debate over what are the expected result tuples for each query below. However, the values for query statistics may vary from person to person depending on precisely how they implemented codeword building and numbers of bits set.

However, there are some statistics values that everyone should get the same answer for, regardless of how they implemented codewords and signatures.

Reminder: you get some statistics on the *relation* via the `./stats` command, e.g.

```
$ ./stats R
Global Info:
Dynamic:
  #items:  tuples: 2000  tsigs: 2000  psigs: 18  bsigs: 3336
  #pages:  tuples: 18  tsigs: 2  psigs: 2  bsigs: 3
Static:
  tups    #attrs: 3  size: 35 bytes  max/page: 116
  sigs    simc  bits/attr: 6           (alternatively: catc and no bits/attr)
  tsigs   size: 32 bits (4 bytes)  max/page: 1023
  psigs   size: 3336 bits (417 bytes)  max/page: 9
  bsigs   size: 24 bits (3 bytes)   max/page: 1364
```

We'll use the above to explain some determined values.

If you ask a query that doesn't use any of the signatures, the statistics should look like:

```
Query Stats:
# sig pages read:     0
# signatures read:    0
# data pages read:    18
# tuples examined:    2000
# false match pages: ??
```

You read no signature pages or signatures. You read all of the data pages. You examine all of the tuples. The number of pages that do not contain matching tuples depends on the query. If you ask a query with a single result, then the number of pages without matching tuples would be 17 (#data-pages - 1) in the above database.

If you ask a query using the tuple signatures, then you should see statistics like:

```
Query Stats:
# sig pages read:     2     (there are 2 pages of tuple signatures)
# signatures read:    2000  (there are 2000 tuple signatures)
# data pages read:    X
# tuples examined:    Y
# false match pages: Z
```

You will always read all of the 2000 tuple signatures from the 2 pages in the `.tsig`. The values of X, Y and Z depend on (a) whether you're using SIMC or CATC, (b) the query, (c) the false match probability, and (d) how you built your codewords.

If you ask a query using page signatures, on the above database, then you should see statistics like:

```
Query Stats:
# sig pages read:     2     (there are 2 pages of page signatures)
# signatures read:    18    (there are 18 page signatures)
# data pages read:    X
# tuples examined:    Y
# false match pages: Z
```

As for the previous example, the values of X, Y and Z are affected by several factors. You can infer some values from their relationship with other values. For example, if you read 1 data page, then you'll read 116 tuples (the number of tuples per page). If you happen to look at the last page in the data file, and it's not full, then the number may be less than this.

The above examples show that you will not necessarily get the same numbers for x, y and z as shown in the examples below.

Testing Examples

We will use just data file `data3` for these examples (otherwise the output would be too long). We generate two relations, one using SIMC, one using CATC, and then run 5 queries on each relation. We run each query with no indexing, tuple signature indexing, page signature indexing and bit-sliced indexing. We make the data files capable of holding more tuples than we actually insert, to avoid overflow issues.

You can easily devise your own tests. You can check correctness by running the `grep` command on the appropraite data file. You can check the sanity of the query stats by looking at the relation stats and reasoning from there. If you constantly read all (or nearly all) of the data pages when using the signature indexes, then you can infer that your signature filtering is not working properly.

The queries are

- `'7663852,?,?,?'` which has one matching tuple
- `'7664096,PjZZsBYoEYAMzgpCgRKg,?,?'` which has one matching tuple
- `'?,?,a3-242,a4-242'` which has 10 matching tuples
- `'8765432,?,?,?'` which has zero matching tuples
- `'7664096,tRzgWRUlUEdoYPZjofYr,?,?'` which has zero matching tuples

Using `data3` with SIMC indexing:

```
$ rm -f R.bsig R.data R.info R.psig R.tsig

$ ./create R simc 10100 4 1000

$ ./insert R
real    0m1.218s
user    0m0.574s
sys         0m0.643s

$ ./stats R
Global Info:
Dynamic:
  #items:  tuples: 10000  tsigs: 10000  psigs: 104  bsigs: 5584
  #pages:  tuples: 104  tsigs: 20  psigs: 21  bsigs: 20
Static:
  tups    #attrs: 4  size: 42 bytes  max/page: 97
  sigs    simc  bits/attr: 9
  tsigs   size: 64 bits (8 bytes)  max/page: 511
  psigs   size: 5584 bits (698 bytes)  max/page: 5
  bsigs   size: 112 bits (14 bytes)  max/page: 292

$ ./select R '7663852,?,?,?'
7663852,ivUryYVexVPJQFjhHxea,a3-069,a4-235
Query Stats:
# sig pages read:     0
# signatures read:    0
# data pages read:    104
# tuples examined:    10000
# false match pages: 103

$ ./select R '7663852,?,?,?' t
```

```
7663852,ivUryYVexVPJQFjhHxea,a3-069,a4-235
Query Stats:
# sig pages read:      20
# signatures read:     10000
# data pages read:     6
# tuples examined:     582
# false match pages: 5

$ ./select R '7663852,?,?,?' p
7663852,ivUryYVexVPJQFjhHxea,a3-069,a4-235
Query Stats:
# sig pages read:      21
# signatures read:     104
# data pages read:     1
# tuples examined:     97
# false match pages: 0

$ ./select R '7663852,?,?,?' b
7663852,ivUryYVexVPJQFjhHxea,a3-069,a4-235
Query Stats:
# sig pages read:      7
# signatures read:     9
# data pages read:     1
# tuples examined:     97
# false match pages: 0

$ ./select R '7664096,PjZZsBYoEYAMzgpCgRKg,?,?'
7664096,PjZZsBYoEYAMzgpCgRKg,a3-064,a4-147
Query Stats:
# sig pages read:      0
# signatures read:     0
# data pages read:     104
# tuples examined:     10000
# false match pages: 103

$ ./select R '7664096,PjZZsBYoEYAMzgpCgRKg,?,?' t
7664096,PjZZsBYoEYAMzgpCgRKg,a3-064,a4-147
Query Stats:
# sig pages read:      20
# signatures read:     10000
# data pages read:     1
# tuples examined:     97
# false match pages: 0

$ ./select R '7664096,PjZZsBYoEYAMzgpCgRKg,?,?' p
7664096,PjZZsBYoEYAMzgpCgRKg,a3-064,a4-147
Query Stats:
# sig pages read:      21
# signatures read:     104
# data pages read:     1
# tuples examined:     97
# false match pages: 0

$ ./select R '7664096,PjZZsBYoEYAMzgpCgRKg,?,?' b
7664096,PjZZsBYoEYAMzgpCgRKg,a3-064,a4-147
Query Stats:
# sig pages read:      12
```

```
# signatures read:    18
# data pages read:    1
# tuples examined:    97
# false match pages: 0


$ ./select R '?,?,a3-242,a4-242'
7654563,kGdSycHkVRpBAXbPMwIq,a3-242,a4-242
7655559,jLsUNyUuDtJxeUirkvXJ,a3-242,a4-242
7656555,VngPhvobyzwypmiERBNG,a3-242,a4-242
7657551,RuFFnNlYxQHYbYatRfNZ,a3-242,a4-242
7658547,bsIHFFHQiGXiRIZLRGdt,a3-242,a4-242
7659543,jWMPpkazbXZEsmZfuADc,a3-242,a4-242
7660539,ecSinUcJGqXsvSRleEda,a3-242,a4-242
7661535,CruzuxQXYSqCctqvDzFW,a3-242,a4-242
7662531,lRibeHIcdoIPurlWusYu,a3-242,a4-242
7663527,DxQAAntFsMFLfUXzeGGh,a3-242,a4-242
Query Stats:
# sig pages read:     0
# signatures read:    0
# data pages read:    104
# tuples examined:    10000
# false match pages: 94


$ ./select R '?,?,a3-242,a4-242' t
7654563,kGdSycHkVRpBAXbPMwIq,a3-242,a4-242
7655559,jLsUNyUuDtJxeUirkvXJ,a3-242,a4-242
7656555,VngPhvobyzwypmiERBNG,a3-242,a4-242
7657551,RuFFnNlYxQHYbYatRfNZ,a3-242,a4-242
7658547,bsIHFFHQiGXiRIZLRGdt,a3-242,a4-242
7659543,jWMPpkazbXZEsmZfuADc,a3-242,a4-242
7660539,ecSinUcJGqXsvSRleEda,a3-242,a4-242
7661535,CruzuxQXYSqCctqvDzFW,a3-242,a4-242
7662531,lRibeHIcdoIPurlWusYu,a3-242,a4-242
7663527,DxQAAntFsMFLfUXzeGGh,a3-242,a4-242
Query Stats:
# sig pages read:     20
# signatures read:    10000
# data pages read:    10
# tuples examined:    970
# false match pages: 0


$ ./select R '?,?,a3-242,a4-242' p
7654563,kGdSycHkVRpBAXbPMwIq,a3-242,a4-242
7655559,jLsUNyUuDtJxeUirkvXJ,a3-242,a4-242
7656555,VngPhvobyzwypmiERBNG,a3-242,a4-242
7657551,RuFFnNlYxQHYbYatRfNZ,a3-242,a4-242
7658547,bsIHFFHQiGXiRIZLRGdt,a3-242,a4-242
7659543,jWMPpkazbXZEsmZfuADc,a3-242,a4-242
7660539,ecSinUcJGqXsvSRleEda,a3-242,a4-242
7661535,CruzuxQXYSqCctqvDzFW,a3-242,a4-242
7662531,lRibeHIcdoIPurlWusYu,a3-242,a4-242
7663527,DxQAAntFsMFLfUXzeGGh,a3-242,a4-242
Query Stats:
# sig pages read:     21
# signatures read:    104
# data pages read:    12
# tuples examined:    1164
```

```
# false match pages: 2

$ ./select R '?,?,a3-242,a4-242' b
7654563,kGdSycHkVRpBAXbPMwIq,a3-242,a4-242
7655559,jLsUNyUuDtJxeUirkvXJ,a3-242,a4-242
7656555,VngPhvobyzwypmiERBNG,a3-242,a4-242
7657551,RuFFnNlYxQHYbYatRfNZ,a3-242,a4-242
7658547,bsIHFFHQiGXiRIZLRGdt,a3-242,a4-242
7659543,jWMPpkazbXZEsmZfuADc,a3-242,a4-242
7660539,ecSinUcJGqXsvSRleEda,a3-242,a4-242
7661535,CruzuxQXYSqCctqvDzFW,a3-242,a4-242
7662531,lRibeHIcdoIPurlWusYu,a3-242,a4-242
7663527,DxQAAntFsMFLfUXzeGGh,a3-242,a4-242
Query Stats:
# sig pages read:      11
# signatures read:     18
# data pages read:     12
# tuples examined:     1164
# false match pages: 2

$ ./select R '8765432,?,?,?'
Query Stats:
# sig pages read:      0
# signatures read:     0
# data pages read:     104
# tuples examined:     10000
# false match pages: 104

$ ./select R '8765432,?,?,?' t
Query Stats:
# sig pages read:      20
# signatures read:     10000
# data pages read:     3
# tuples examined:     291
# false match pages: 3

$ ./select R '8765432,?,?,?' p
Query Stats:
# sig pages read:      21
# signatures read:     104
# data pages read:     0
# tuples examined:     0
# false match pages: 0

$ ./select R '8765432,?,?,?' b
Query Stats:
# sig pages read:      8
# signatures read:     9
# data pages read:     0
# tuples examined:     0
# false match pages: 0

$ ./select R '7664096,tRzgWRUlUEdoYPZjofYr,?,?'
Query Stats:
# sig pages read:      0
# signatures read:     0
# data pages read:     104
```

```
# tuples examined:    10000
# false match pages: 104

$ ./select R '7664096,tRzgWRUlUEdoYPZjofYr,?,?' t
Query Stats:
# sig pages read:     20
# signatures read:    10000
# data pages read:    0
# tuples examined:    0
# false match pages: 0

$ ./select R '7664096,tRzgWRUlUEdoYPZjofYr,?,?' p
Query Stats:
# sig pages read:     21
# signatures read:    104
# data pages read:    0
# tuples examined:    0
# false match pages: 0

$ ./select R '7664096,tRzgWRUlUEdoYPZjofYr,?,?' b
Query Stats:
# sig pages read:     11
# signatures read:    18
# data pages read:    0
# tuples examined:    0
# false match pages: 0
```

Using `data3` with CATC indexing:

```
$ rm -f R.bsig R.data R.info R.psig R.tsig

$ ./create R catc 10100 4 1000

$ ./insert R
real   0m1.268s
user   0m0.753s
sys        0m0.513s

$ ./stats R
Global Info:
Dynamic:
  #items:  tuples: 10000  tsigs: 10000  psigs: 104  bsigs: 5584
  #pages:  tuples: 104  tsigs: 20  psigs: 21  bsigs: 20
Static:
  tups   #attrs: 4  size: 42 bytes  max/page: 97
  sigs   catc
  tsigs  size: 64 bits (8 bytes)  max/page: 511
  psigs  size: 5584 bits (698 bytes)  max/page: 5
  bsigs  size: 112 bits (14 bytes)  max/page: 292

$ ./select R '7663852,?,?,?'
7663852,ivUryYVexVPJQFjhHxea,a3-069,a4-235
Query Stats:
# sig pages read:     0
# signatures read:    0
```

```
# data pages read:    104
# tuples examined:    10000
# false match pages: 103

$ ./select R '7663852,?,?,?' t
7663852,ivUryYVexVPJQFjhHxea,a3-069,a4-235
Query Stats:
# sig pages read:     20
# signatures read:    10000
# data pages read:    2
# tuples examined:    194
# false match pages: 1

$ ./select R '7663852,?,?,?' p
7663852,ivUryYVexVPJQFjhHxea,a3-069,a4-235
Query Stats:
# sig pages read:     21
# signatures read:    104
# data pages read:    1
# tuples examined:    97
# false match pages: 0

$ ./select R '7663852,?,?,?' b
7663852,ivUryYVexVPJQFjhHxea,a3-069,a4-235
Query Stats:
# sig pages read:     4
# signatures read:    7
# data pages read:    1
# tuples examined:    97
# false match pages: 0

$ ./select R '7664096,PjZZsBYoEYAMzgpCgRKg,?,?'
7664096,PjZZsBYoEYAMzgpCgRKg,a3-064,a4-147
Query Stats:
# sig pages read:     0
# signatures read:    0
# data pages read:    104
# tuples examined:    10000
# false match pages: 103

$ ./select R '7664096,PjZZsBYoEYAMzgpCgRKg,?,?' t
7664096,PjZZsBYoEYAMzgpCgRKg,a3-064,a4-147
Query Stats:
# sig pages read:     20
# signatures read:    10000
# data pages read:    1
# tuples examined:    97
# false match pages: 0

$ ./select R '7664096,PjZZsBYoEYAMzgpCgRKg,?,?' p
7664096,PjZZsBYoEYAMzgpCgRKg,a3-064,a4-147
Query Stats:
# sig pages read:     21
# signatures read:    104
# data pages read:    1
# tuples examined:    97
# false match pages: 0
```

```
$ ./select R '7664096,PjZZsBYoEYAMzgpCgRKg,?,?' b
7664096,PjZZsBYoEYAMzgpCgRKg,a3-064,a4-147
Query Stats:
# sig pages read:       8
# signatures read:      14
# data pages read:      1
# tuples examined:      97
# false match pages: 0


$ ./select R '?,?,a3-242,a4-242'
7654563,kGdSycHkVRpBAXbPMwIq,a3-242,a4-242
7655559,jLsUNyUuDtJxeUirkvXJ,a3-242,a4-242
7656555,VngPhvobyzwypmiERBNG,a3-242,a4-242
7657551,RuFFnNlYxQHYbYatRfNZ,a3-242,a4-242
7658547,bsIHFFHQiGXiRIZLRGdt,a3-242,a4-242
7659543,jWMPpkazbXZEsmZfuADc,a3-242,a4-242
7660539,ecSinUcJGqXsvSRleEda,a3-242,a4-242
7661535,CruzuxQXYSqCctqvDzFW,a3-242,a4-242
7662531,lRibeHIcdoIPurlWusYu,a3-242,a4-242
7663527,DxQAAntFsMFLfUXzeGGh,a3-242,a4-242
Query Stats:
# sig pages read:       0
# signatures read:      0
# data pages read:      104
# tuples examined:      10000
# false match pages: 94


$ ./select R '?,?,a3-242,a4-242' t
7654563,kGdSycHkVRpBAXbPMwIq,a3-242,a4-242
7655559,jLsUNyUuDtJxeUirkvXJ,a3-242,a4-242
7656555,VngPhvobyzwypmiERBNG,a3-242,a4-242
7657551,RuFFnNlYxQHYbYatRfNZ,a3-242,a4-242
7658547,bsIHFFHQiGXiRIZLRGdt,a3-242,a4-242
7659543,jWMPpkazbXZEsmZfuADc,a3-242,a4-242
7660539,ecSinUcJGqXsvSRleEda,a3-242,a4-242
7661535,CruzuxQXYSqCctqvDzFW,a3-242,a4-242
7662531,lRibeHIcdoIPurlWusYu,a3-242,a4-242
7663527,DxQAAntFsMFLfUXzeGGh,a3-242,a4-242
Query Stats:
# sig pages read:       20
# signatures read:      10000
# data pages read:      10
# tuples examined:      970
# false match pages: 0


$ ./select R '?,?,a3-242,a4-242' p
7654563,kGdSycHkVRpBAXbPMwIq,a3-242,a4-242
7655559,jLsUNyUuDtJxeUirkvXJ,a3-242,a4-242
7656555,VngPhvobyzwypmiERBNG,a3-242,a4-242
7657551,RuFFnNlYxQHYbYatRfNZ,a3-242,a4-242
7658547,bsIHFFHQiGXiRIZLRGdt,a3-242,a4-242
7659543,jWMPpkazbXZEsmZfuADc,a3-242,a4-242
7660539,ecSinUcJGqXsvSRleEda,a3-242,a4-242
7661535,CruzuxQXYSqCctqvDzFW,a3-242,a4-242
7662531,lRibeHIcdoIPurlWusYu,a3-242,a4-242
7663527,DxQAAntFsMFLfUXzeGGh,a3-242,a4-242
```

```
Query Stats:
# sig pages read:     21
# signatures read:    104
# data pages read:    12
# tuples examined:    1164
# false match pages: 2


$ ./select R '?,?,a3-242,a4-242' b
7654563,kGdSycHkVRpBAXbPMwIq,a3-242,a4-242
7655559,jLsUNyUuDtJxeUirkvXJ,a3-242,a4-242
7656555,VngPhvobyzwypmiERBNG,a3-242,a4-242
7657551,RuFFnNlYxQHYbYatRfNZ,a3-242,a4-242
7658547,bsIHFFHQiGXiRIZLRGdt,a3-242,a4-242
7659543,jWMPpkazbXZEsmZfuADc,a3-242,a4-242
7660539,ecSinUcJGqXsvSRleEda,a3-242,a4-242
7661535,CruzuxQXYSqCctqvDzFW,a3-242,a4-242
7662531,lRibeHIcdoIPurlWusYu,a3-242,a4-242
7663527,DxQAAntFsMFLfUXzeGGh,a3-242,a4-242
Query Stats:
# sig pages read:     8
# signatures read:    14
# data pages read:    12
# tuples examined:    1164
# false match pages: 2


$ ./select R '8765432,?,?,?'
Query Stats:
# sig pages read:     0
# signatures read:    0
# data pages read:    104
# tuples examined:    10000
# false match pages: 104


$ ./select R '8765432,?,?,?' t
Query Stats:
# sig pages read:     20
# signatures read:    10000
# data pages read:    0
# tuples examined:    0
# false match pages: 0


$ ./select R '8765432,?,?,?' p
Query Stats:
# sig pages read:     21
# signatures read:    104
# data pages read:    0
# tuples examined:    0
# false match pages: 0


$ ./select R '8765432,?,?,?' b
Query Stats:
# sig pages read:     4
# signatures read:    7
# data pages read:    0
# tuples examined:    0
# false match pages: 0
```

```
$ ./select R '7664096,tRzgWRUlUEdoYPZjofYr,?,?'
Query Stats:
# sig pages read:      0
# signatures read:     0
# data pages read:     104
# tuples examined:     10000
# false match pages: 104

$ ./select R '7664096,tRzgWRUlUEdoYPZjofYr,?,?' t
Query Stats:
# sig pages read:      20
# signatures read:     10000
# data pages read:     0
# tuples examined:     0
# false match pages: 0

$ ./select R '7664096,tRzgWRUlUEdoYPZjofYr,?,?' p
Query Stats:
# sig pages read:      21
# signatures read:     104
# data pages read:     0
# tuples examined:     0
# false match pages: 0

$ ./select R '7664096,tRzgWRUlUEdoYPZjofYr,?,?' b
Query Stats:
# sig pages read:      8
# signatures read:     14
# data pages read:     0
# tuples examined:     0
# false match pages: 0
```

ChangeLog

- **v1.0** (2021-04-12 21:00:00+10:00)
  - released Assignment 2 Testing

Have fun, *jas*