Übung 7

Alexander Hinneburg

Aufgabe 7.1

```
sed "1,1d" vorlesung_interesse.txt|sort -n -k2,2 -t" " >vi_sorted.txt
gawk -v FS=" " -f transactions.awk <vi_sorted.txt > transactions.txt
```

```
NR==1{
sid=$2;
x=sprintf("%d",$1);
}
NR>1{
if ($2==sid)
 {
 x=sprintf("%s %d",x,$1);
 }
else
 {
 print x;
 x=sprintf("%d",$1);
 sid=$2;
 }
}
```

Aufgabe 7.1

Minsupport = 5

perl.exe miner.pl 5 <transactions.txt > freqsets.txt

```
gawk -F" " '{if (NF==4) print $0}'  <freqsets.txt |
    sort -n -k1,1 -t" " |tail -n 3 |
    gawk -v FS=" " -f freqsets.awk
```

```
{
  tmp="tmp_" NR ".txt";
  for (i=2;i<=NF;i++) {
    print $i >> tmp;
  }
}
```

Item-Mengen Laenge 3

```
join -t " " tmp_1.txt vorlesung_id_ohneKopf.txt > res_3_3.txt
join -t " " tmp_2.txt vorlesung_id_ohneKopf.txt > res_3_2.txt
join -t " " tmp_3.txt vorlesung_id_ohneKopf.txt > res_3_1.txt
```

- 58 mal
  - 1502 'Informatik 1'
  - 1545 'Mathematische Grundlagen der Informatik'
  - 1551 'B1: Lineare Algebra und Analytische Geometrie‘
- 26 mal
  - 1485 'Analysis I'
  - 1502 'Informatik 1'
  - 1541 'Lineare Algebra I‘
- 25 mal
  - 1490 'Modul M4: Numerische Mathematik'
  - 1492 'Praxis des Programmierens'
  - 1521 'Rechnerorganisation und -architektur (Informatik III)'

Item-Mengen Laenge 4

```
gawk -F" " '{if (NF==5) print $0}'  <freqsets.txt | sort -n -k1,1 -t" " |tail -n 3 | gawk -v FS=" " -f freqsets.awk
join -t " " tmp_1.txt vorlesung_id_ohneKopf.txt > res_4_3.txt
join -t " " tmp_2.txt vorlesung_id_ohneKopf.txt > res_4_2.txt
join -t " " tmp_3.txt vorlesung_id_ohneKopf.txt > res_4_1.txt
```

- 18 mal
  - 1674 'Allgemeine Botanik (Ihl KlÃ¶sgen)'
  - 1675 'Zellbiologie (KlÃ¶sgen)'
  - 1676 'Zoologie (Moritz)'
  - 1677 'Grundlagen der physikalischen Chemie (WeiÃ
- 18 mal
  - 1551 'B1: Lineare Algebra und Analytische Geometrie'
  - 1675 'Zellbiologie (KlÃ¶sgen)'
  - 1676 'Zoologie (Moritz)'
  - 1677 'Grundlagen der physikalischen Chemie (WeiÃ
- 18 mal
  - 1551 'B1: Lineare Algebra und Analytische Geometrie'
  - 1674 'Allgemeine Botanik (Ihl KlÃ¶sgen)'
  - 1676 'Zoologie (Moritz)'
  - 1677 'Grundlagen der physikalischen Chemie (WeiÃ

Item-Mengen Laenge 5

```
gawk -F" " '{if (NF==6) print $0}'  <freqsets.txt | sort -n -k1,1 -t" " |tail -n 3 | gawk -v FS=" " -f freqsets.awk
join -t " " tmp_1.txt vorlesung_id_ohneKopf.txt > res_5_3.txt
join -t " " tmp_2.txt vorlesung_id_ohneKopf.txt > res_5_2.txt
join -t " " tmp_3.txt vorlesung_id_ohneKopf.txt > res_5_1.txt
```

- 18 mal
  - 1551 'B1: Lineare Algebra und Analytische Geometrie'
  - 1674 'Allgemeine Botanik (Ihl KlÃ¶sgen)'
  - 1675 'Zellbiologie (KlÃ¶sgen)'
  - 1676 'Zoologie (Moritz)'
  - 1677 'Grundlagen der physikalischen Chemie (WeiÃ
- 18 mal
  - 1502 'Informatik 1'
  - 1674 'Allgemeine Botanik (Ihl KlÃ¶sgen)'
  - 1675 'Zellbiologie (KlÃ¶sgen)'
  - 1676 'Zoologie (Moritz)'
  - 1677 'Grundlagen der physikalischen Chemie (WeiÃ
- 18 mal
  - 1502 'Informatik 1'
  - 1551 'B1: Lineare Algebra und Analytische Geometrie'
  - 1675 'Zellbiologie (KlÃ¶sgen)'
  - 1676 'Zoologie (Moritz)'
  - 1677 'Grundlagen der physikalischen Chemie (WeiÃ

Item-Mengen Laenge 6

- 18 mal
  - 1502 'Informatik 1'
  - 1551 'B1: Lineare Algebra und Analytische Geometrie'
  - 1674 'Allgemeine Botanik (Ihl KlÃ¶sgen)'
  - 1675 'Zellbiologie (KlÃ¶sgen)'
  - 1676 'Zoologie (Moritz)'
  - 1677 'Grundlagen der physikalischen Chemie (WeiÃ
- 12 mal
  - 1490 'Modul M4: Numerische Mathematik'
  - 1492 'Praxis des Programmierens'
  - 1639 'Datenbanken I'
  - 1678 'Genetik (Siegemundt)'
  - 1679 'Biochemie (Ulbrich)'
  - 1681 'Bioorganische Chemie (Csuk/WeiÃ
- 6 mal
  - 1490 'Modul M4: Numerische Mathematik'
  - 1502 'Informatik 1'
  - 1639 'Datenbanken I'
  - 1678 'Genetik (Siegemundt)'
  - 1679 'Biochemie (Ulbrich)'
  - 1681 'Bioorganische Chemie (Csuk/WeiÃ

Aufgabe 7.2, Init

```
create table w ( wid integer, it integer,
     sl real, sw real, pl real, pw real);
k=4

insert into w
select wid, iter, avg(sl), avg(sw), avg(pl), avg(pw)
from (
  select floor(dbms_random.value(1,4+1)) as wid,
    0 as iter,
    sepallength as sl, sepalwidth as sw, petallength as pl, petalwidth as pw
    from iris
) a
group by a.wid, a.iter;
```

Aufgabe 7.2, Init

```
 WID       ITER    AVG(SL)    AVG(SW)    AVG(PL)    AVG(PW)
---------- ---------- ---------- ---------- ---------- ----------
         1          0 5.794E+000 3.159E+000 3.582E+000 1.174E+000
         2          0 5.869E+000 2.983E+000 3.776E+000 1.171E+000
         3          0 5.894E+000 3.037E+000  4.02E+000 1.326E+000
         4          0 5.813E+000 3.054E+000 3.659E+000 1.136E+000
```

```
select avg(sl), avg(sw), avg(pl), avg(pw)
from (
  select sepallength as sl, sepalwidth as sw, petallength as pl, petalwidth as pw
    from iris
) a;
   AVG(SL)    AVG(SW)    AVG(PL)    AVG(PW)
---------- ---------- ---------- ----------
5.843E+000 3.054E+000 3.759E+000 1.199E+000
```

Beispiel, Datenpunkt 150

```
SQL>            select pid, wid,
              sqrt((i.sepallength-w.sl)*(i.sepallength-w.sl)+
                   (i.sepalwidth-w.sw)*(i.sepalwidth-w.sw)+
                   (i.petallength-w.pl)*(i.petallength-w.pl)+
                   (i.petalwidth-w.pw)*(i.petalwidth-w.pw)) as dist
            from iris i,w
            where it=0 and pid=150
      PID       WID       DIST
--------- ---------- ----------
      150          1 1.779E+000
      150          2 1.409E+000
      150          3 9.911E-001
      150          4   1.7E+000
```

Beispiel, Datenpunkt 150

```
SQL>            select pid, min(dist) as min_dist
          from (
            select pid, wid,
               sqrt((i.sepallength-w.sl)*(i.sepallength-w.sl)+
                    (i.sepalwidth-w.sw)*(i.sepalwidth-w.sw)+
                    (i.petallength-w.pl)*(i.petallength-w.pl)+
                    (i.petalwidth-w.pw)*(i.petalwidth-w.pw)) as dist
            from iris i,w
            where it=0 and pid=150
            ) a
          group by a.pid
       PID   MIN_DIST
---------- ----------
       150 9.911E-001
```

Beispiel, Datenpunkt 150

```
# Back Join
SQL> select b.pid, c.wid
      from (
         select pid, min(dist) as min_dist
         from (
           select pid, wid,
              sqrt((i.sepallength-w.sl)*(i.sepallength-w.sl)+
                   (i.sepalwidth-w.sw)*(i.sepalwidth-w.sw)+
                   (i.petallength-w.pl)*(i.petallength-w.pl)+
                   (i.petalwidth-w.pw)*(i.petalwidth-w.pw)) as dist
           from iris i,w where it=0 and pid=150
         ) a
         group by a.pid
       ) b,
       ( select pid, wid,
            sqrt((i.sepallength-w.sl)*(i.sepallength-w.sl)+
                 (i.sepalwidth-w.sw)*(i.sepalwidth-w.sw)+
                 (i.petallength-w.pl)*(i.petallength-w.pl)+
                 (i.petalwidth-w.pw)*(i.petalwidth-w.pw)) as dist
          from iris i,w  where it=0 and pid=150
       ) c  where b.pid=c.pid and b.min_dist=c.dist;
      PID        WID
---------- ----------
      150          3
```

 # Schritt 2

```
select wid, 1 as it, sl, sw, pl, pw
from (
select d.wid,
avg(sepallength) as sl, avg(sepalwidth) as sw, avg(petallength) as pl, avg(petalwidth) as pw
from ( select b.pid, min(c.wid) as wid
       from (
          select pid, min(dist) as min_dist
          from (
            select pid, wid,
               sqrt((i.sepallength-w.sl)*(i.sepallength-w.sl)+
                    (i.sepalwidth-w.sw)*(i.sepalwidth-w.sw)+
                    (i.petallength-w.pl)*(i.petallength-w.pl)+
                    (i.petalwidth-w.pw)*(i.petalwidth-w.pw)) as dist
            from iris i,w
            where it=0
            ) a
          group by a.pid
        ) b,
        ( select pid, wid,
               sqrt((i.sepallength-w.sl)*(i.sepallength-w.sl)+
                    (i.sepalwidth-w.sw)*(i.sepalwidth-w.sw)+
                    (i.petallength-w.pl)*(i.petallength-w.pl)+
                    (i.petalwidth-w.pw)*(i.petalwidth-w.pw)) as dist
           from iris i,w
          where it=0
        ) c
       where b.pid=c.pid and b.min_dist=c.dist
       group by b.pid
     )d, iris i
where d.pid = i.pid
group by d.wid
)
```