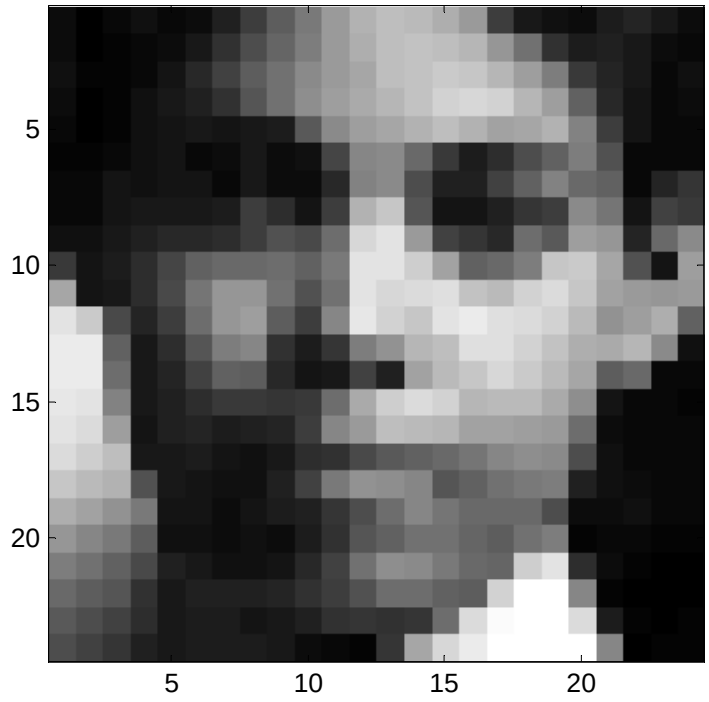
Problem1

a. Load the data and try one image.

```
i = 6;
X = load('data/faces.txt');
img = reshape(X(i,:),[24 24]);
imagesc(img); axis square; colormap gray;
```



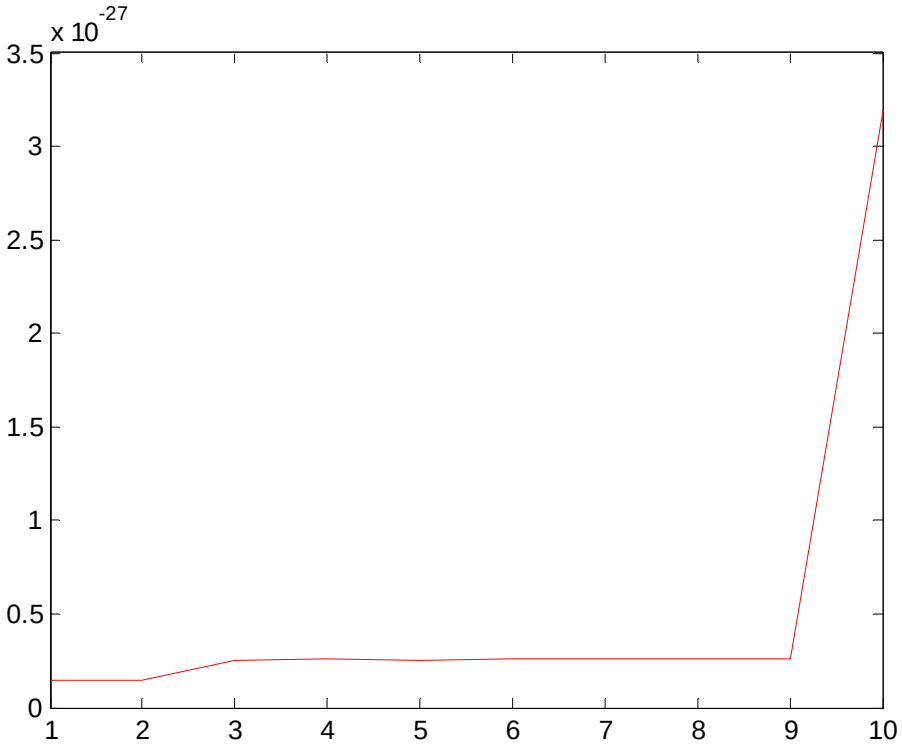
Compute the mean and subtract it to get make data zero-mean.

```
[n m] = size(X);
mn = mean(X,1);
X0 = X - repmat(mn,[n,1]);
```

b. [U,S,V] = svds(X0,10);

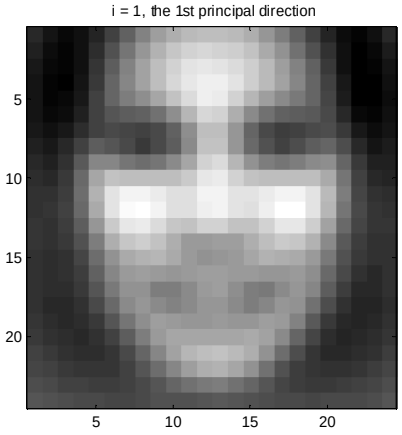
c.

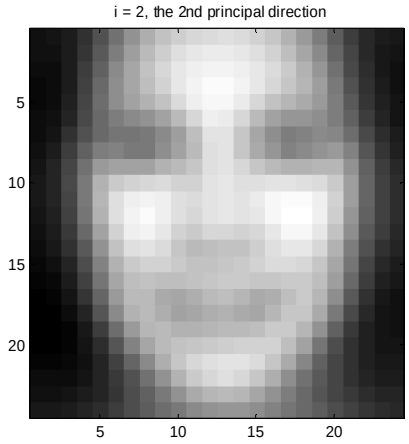
```
W = U*S;
K = [1 2 3 4 5 6 7 8 9 10];
MSE = 0*K;
for k=1:length(K)
  Xapp = W(:,1:K(k))*V(:,1:K(k))';
  MSE(k) = mean(mean(X0-Xapp).^2);
end
figure;
plot(K,MSE,'r-');
```

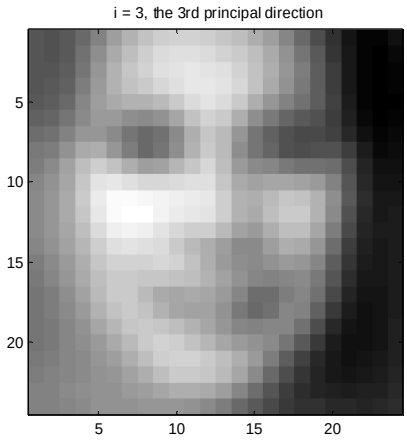


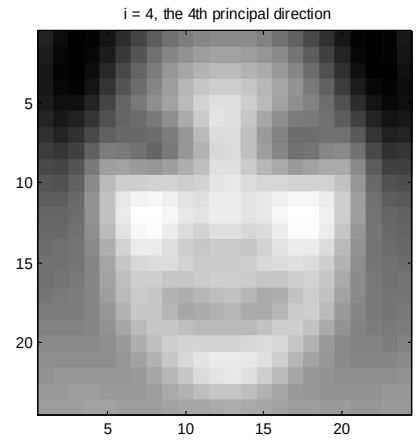
d.

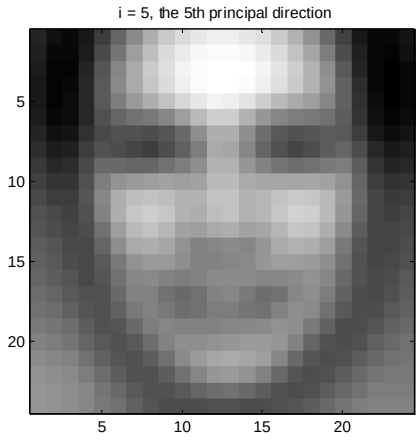
```
[c r] = size(W);
mn = mean(X,1);
m = repmat(mn,[r,1]);
X1 = m;

for j=1:r
   o = 2*median(abs(W(:,j)));
   X1(j,:) = X1(j,:) + o*V(:,j)';
end

I = [1 2 3 4 5 6 7 8 9 10] %i = 1, the 1st principal direction
For i=1:10
   img = reshape(X1(I(i),:),[24 24]);
   imagesc(img); axis square; colormap gray;
```

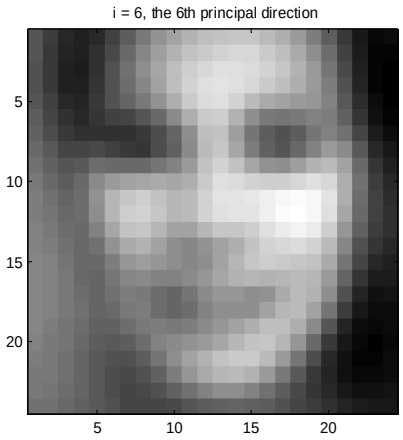


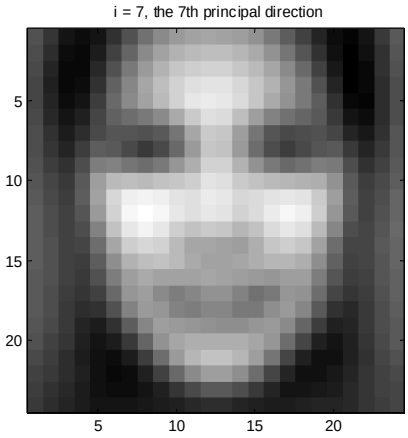


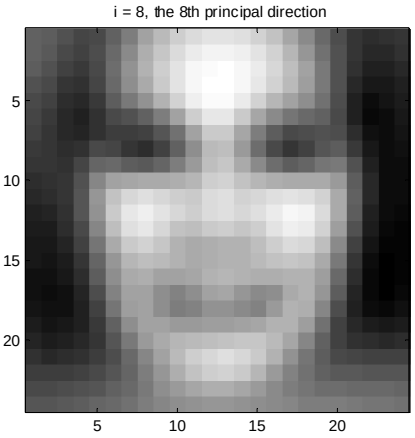




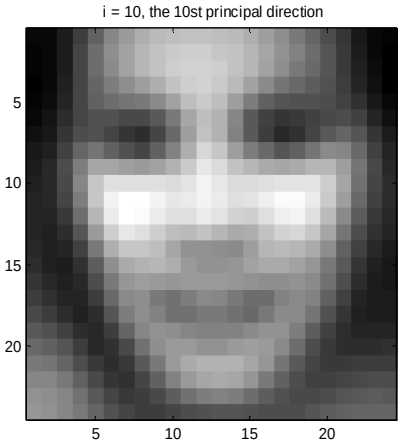

i = 5, the 5th principal direction
5
10
15
20
5
10
15
20


i = 6, the 6th principal direction
5
10
15
20
5
10
15
20


i = 7, the 7th principal direction
5
10
15
20
5
10
15
20

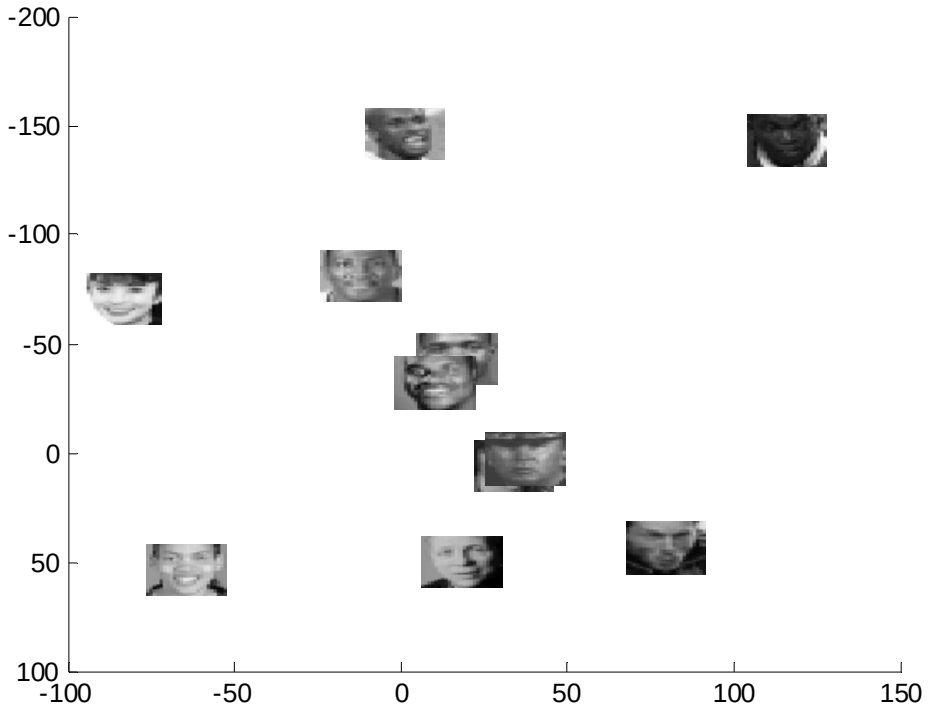

i = 8, the 8th principal direction
5
10
15
20
5
10
15
20


i = 9, the 9th principal direction
5
10
15
20
5
10
15
20


i = 10, the 10st principal direction
5
10
15
20
5
10
15
20

e.

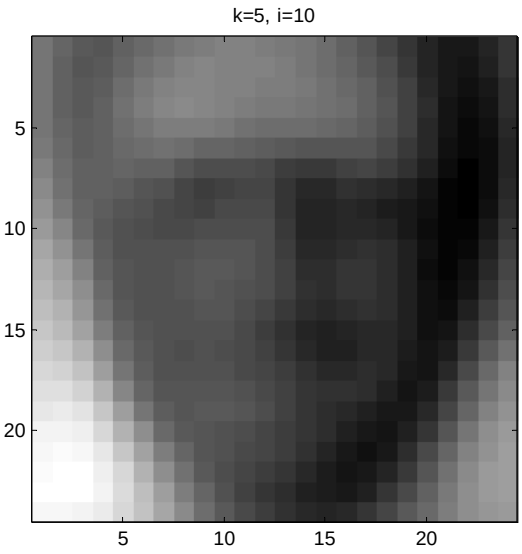
```
idx = [15 16 17 18 19 20 21 22 23 24 25];
figure; hold on; axis ij; colormap(gray);
range = max(W(idx,1:2)) - min(W(idx,1:2));
scale = [200 200]./range;
%imagesc(W(17,1)*scale(1),W(17,2)*scale(2),reshape(X(17,:),24,24));

for i=1:length(idx)
   imagesc(W(idx(i),1)*scale(1),W(idx(i),2)*scale(2),reshape(X(idx(i),:),24,24));
end
```

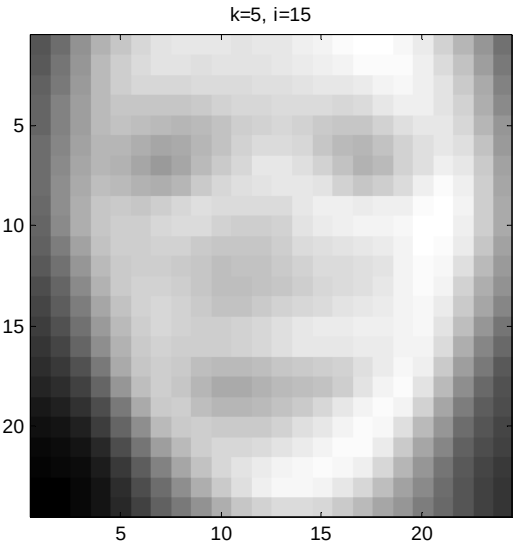


f. we pick up 10th and 15th image

For k = 5,

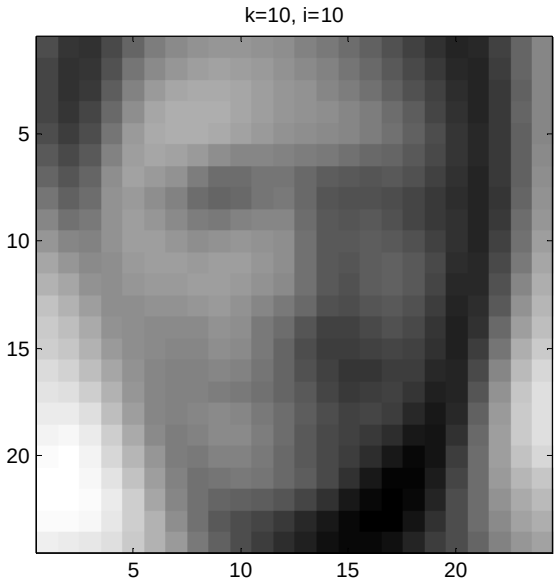
```
[U,S,V] = svds(X0,5);
W = U*S;

i = 10; and i = 15;
Xi = W(i,:)*V';

img = reshape(Xi,[24 24]);
imagesc(img); axis square; colormap gray;
```

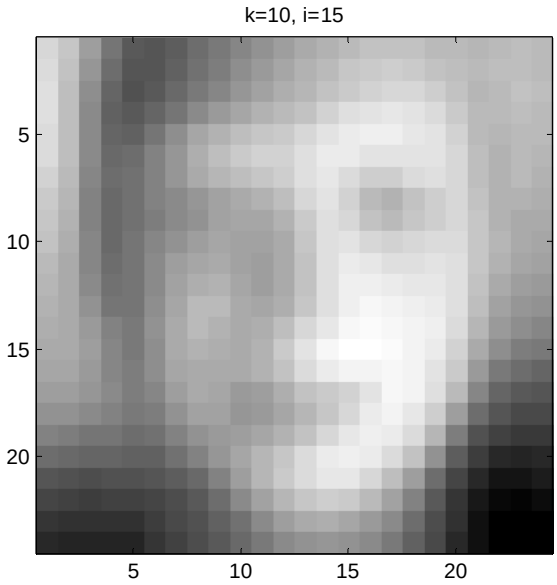




For k = 10,

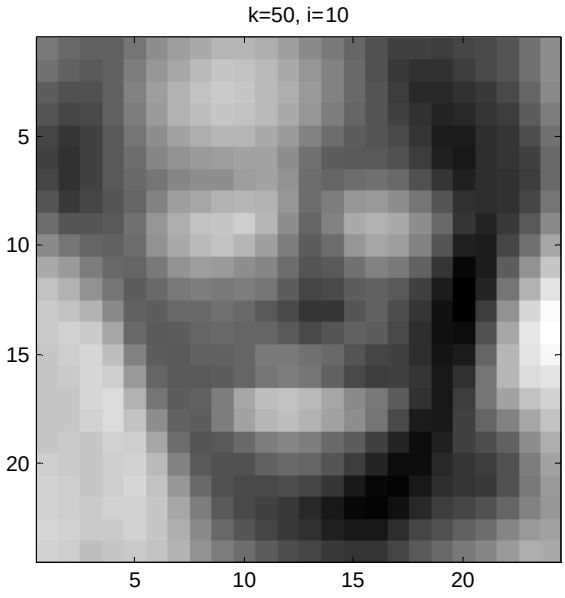
```
[U,S,V] = svds(X0,10);
W = U*S;

i = 10; and i = 15;
Xi = W(i,:)*V';

img = reshape(Xi,[24 24]);
imagesc(img); axis square; colormap gray;
```

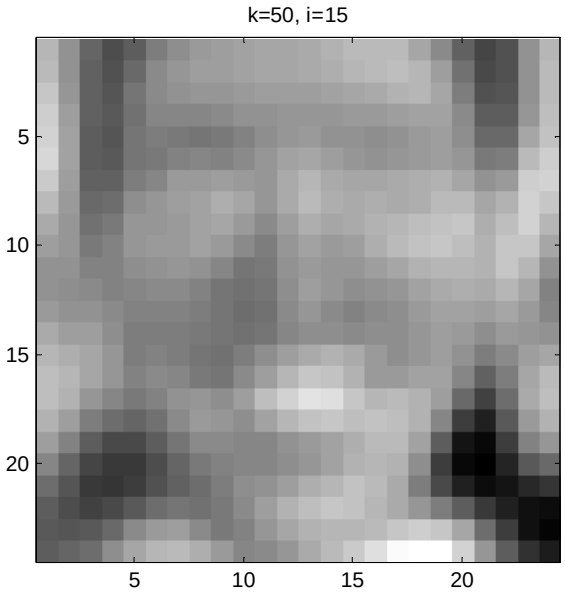




For k = 50,

```
[U,S,V] = svds(X0,50);
W = U*S;

i = 10; and i = 15;
Xi = W(i,:)*V';

img = reshape(Xi,[24 24]);
imagesc(img); axis square; colormap gray;
```





Problem2

a. (1)Load the data

```
[vocab] = textread('vocab.txt','%s');
[did,wid,cnt] = textread('docword.txt','%d%d%d','headerlines',3);

X = sparse(did,wid,cnt);
D = max(did);
W = max(wid);
N = sum(cnt);

Xn = X./repmat(sum(X,2),[1,W]);
```

(2). Find SVD (T = 8)

```
[U,S,V] = svds(Xn,8);
```

b. For the 10 “most positive” words in each topic:

```
[r c] = size(V);

for i=1:c
   [sorted,order] = sort(V(:,i)',2,'descend');
   fprintf('Topic %d\n', i);
   fprintf('%s ',vocab{order(1:10)});
   fprintf('\n');
end
```

Topic 1 (Hard to tell)

keyword wordage adv02 1stld pdf 0101 exp 2takes belatedly microchips

Topic 2 (y2k + 2000)

city 2000 game times team season y2k york millennium 000

Topic 3 (Sport)

game season team games coach players league play giants yards

Topic 4 (Politics - Russia)

putin yeltsin tutsi hutu rwanda russia burundi political russian ethnic

Topic 5 (Politics - Russia)

tutsi hutu rwanda burundi ethnic africa experts group 1994 van

Topic 6 (Technology - y2k)

y2k computer system additional computers systems koskinen problem accounts transferor

Topic 7 (Politics - Europe)

drug american marijuana europe boot algeria americans political states policy

Topic 8 (Hard to tell)

y2k koskinen saturday problems problem reported officials federal 2000 unit

For the 10 “most negative” words in each topic

```
[r c] = size(V);

for i=1:c
   [sorted,order] = sort(V(:,i)',2,'ascend');
   fprintf('Topic %d\n', i);
   fprintf('%s ',vocab{order(1:10)});
   fprintf('\n');
end
```

Topic 1 (Hard to tell)

test end houston city 2000 game team millennium season times

Topic 2 (Hard to tell)

test houston keyword wordage adv02 1stld 0101 exp pdf 2takes

Topic 3 (Hard to tell)

y2k yeltsin putin government russia 2000 system country power computer

Topic 4 (2000)

times square 2000 y2k millennium city midnight fireworks computer night

Topic 5 (Russian)

yeltsin putin russia russian government president chechnya power kremlin roosevelt

Topic 6 (Celebration)

square times fireworks millennium city yeltsin russian night police feet

Topic 7 (Politics)

tutsi yeltsin hutu times rwanda russian burundi square putin ethnic

Topic 8 (Hard to tell)

additional accounts transferor participations trust account rating addition times computer

c. I choose 6th topic and positive sign. And the topic is (Technology - y2k).

```
t = 6;
W = U*S;
[sorted,order] = sort(W(:,t)',2,'descend');
fprintf('Topic %d\n', t);
fprintf('%d ',order(1:3));
fprintf('\n');

array = order(1:3);
for i=1:3
   fprintf('%d\n',i);
   fprintf('Doc %d\n',array(i));
   fname = sprintf('example1/20000101.%04d.txt',array(i));
   txt = textread(fname,'%s',10,'whitespace','\r\n');
   fprintf('%s\n',txt{:});
end
```

```
Topic 6
37 38 166
1
Doc 37          (Technology)
At times more words, not fewer, are needed to decode
particularly turgid passages. Here's one example cited by Lutz,
with a suggested revision.
Before: When business processes are automated, employees are
careful not to fall into the trap of applying new technology to
old, inefficient work procedures. Instead, a needs assessment is
completed to identify system requirements, and then automated
systems are designed to accomplish these goals.
After: Computers don't improve the way you do business, if you
simply do business the same old way using the computers. So, before
2
Doc 38          (Technology)
At times more words, not fewer, are needed to decode
particularly turgid passages. Here's one example cited by Lutz,
with a suggested revision.
Before: When business processes are automated, employees are
careful not to fall into the trap of applying new technology to
old, inefficient work procedures. Instead, a needs assessment is
completed to identify system requirements, and then automated
systems are designed to accomplish these goals.
After: Computers don't improve the way you do business, if you
simply do business the same old way using the computers. So, before
3
Doc 166         (Technology - y2k)
As the world glided smoothly into the new century,
the United States reported Saturday only a smattering of minor Y2K
```

glitches, but officials cautioned that problems could still crop up, especially starting Monday.
John Koskinen, chief of the White House's Y2K command center, said Monday will be a crucial milestone because it will be the first day of normal business operations following the long holiday weekend.
Otherwise, he offered generally upbeat news about the nation's progress. ``We have not been able to find anything of great

These three documents 36, 37 and 166 have the largest magnitude coefficient in topic 6 direction. As the experiment shows, they have the same topic with topic6. So the topic we did in problem b is a good description of the document.