Homework 2

30 + 3 = 3/

Que. 1

1.

$$\begin{bmatrix} 15 & 7 & 6 \\ 7 & 9 & 8 \\ 6 & 8 & 22 \end{bmatrix}$$

Because correlation between terms is symmetric

2.

$$C_{k_1k_2} = \frac{7}{9+15-7} = \frac{7}{17} = 0.41$$

$$C_{k_2k_3} = \frac{8}{9+22-8} = \frac{8}{23} = 0.35$$

So K_1 is most correlated with K_2.

But using un normalize clusters.

K_3 is most correlated with K_2.

Q. 3

Scalar clusters.

$$S_{K_1K_2} = \frac{\cos\theta\,(\vec{K_1},\vec{K_2})}{}$$

$$= \frac{\vec{K_1}\cdot\vec{K_2}}{|\vec{K_1}|\,|\vec{K_2}|}$$

For K_1 vector can be found as

$$C_{K_1K_1} = \frac{15}{15+15-15} = 1$$

$C_{K_1K_2} = 0.41$ (from prev. que.)

$$C_{K_1K_3} = \frac{6}{15+22-6} = \frac{6}{31} = 0.19$$

$$\vec{K_1} = (1, 0.41, 0.19)$$

For $\vec{K_2}$

$C_{K_1K_2} = 0.41$

$C_{K_2K_2} = 1$

$C_{K_2K_3} = 0.35$ (from prev. que.)

$$S_{k_1k_2} = \frac{0.41 + 0.41 + 0.19 \times 0.35}{(1.09)(1.14)}$$

$$= \frac{0.82 + 0.0665}{1.09 \times 1.14}$$

$$\boxed{S_{k_1k_2} = 0.71}$$

Homework 2

Question 2 (Bush / Saddam & LSI)

1.

F-F =

23.3340	0	0	0
0	9.7667	0	0
0	0	5.0379	0
0	0	0	3.2793
0	0	0	0
0	0	0	0

If we remove last one then

$$\text{loss} = 1 - \frac{\sum_{i=1}^{3} \lambda_i^2}{\sum_{i=1}^{4} \lambda_i^2}$$ (∵ here FF contai singular values & we need eigen values

$= 1 - \frac{665.14}{675.90}$

$= 1 - 0.98$

$= 0.02$ (2% loss)

We can take only two singular values then

$$\text{Loss} = 1 - \frac{\sum_{i=1}^{2} \lambda_i^2}{\sum_{i=1}^{4} \lambda_i^2} = 1 - \frac{639.74}{675}$$

$= 0.05$ (5% loss)

We can keep only one singular v

$$loss = 1 - \frac{\cancel{\xi}\lambda_1^2}{\sum_{i=1}^{4} \lambda_i^2} = 1 - \frac{544.}{675}$$

$= 0.19$

(19 % loss

So we can't remove third,

~~So at max~~

So minimum 2 dimensions we need to keep to have loss < 10%. (5% in this case).

2.

As we keep only two most important dimensions.

$$DF = \begin{matrix} -0.2638 & 0.285 \\ -0.6627 & 0.6018 \\ -0.4237 & -0.6079 \\ -0.4293 & -0.3061 \\ -0.3549 & -0.2171 \\ -0.0373 & -0.2151 \end{matrix}$$

FF

$$\begin{matrix} 23.334 & 0 \\ 0 & 9.7667 \end{matrix}$$

TF'

$$\begin{matrix} -0.8817 & -0.2887 & -0.3033 & -0.2173 \\ 0.1969 & 0.4928 & -0.6652 & -0.5253 \\ -0.0444 & 0.119 & -0.5674 & 0.8136 \\ -0.4264 & 0.8122 & 0.379 & 0.1222 \end{matrix}$$

Now remove 2 rows so

TF'_k

$$\begin{matrix} -0.8817 & -0.2887 & -0.3033 & -0.2173 \\ 0.1969 & 0.4928 & -0.6652 & -0.525 \end{matrix}$$

To get vector for d_1 we need to multiply first row of DF with FF and then with TF'

$DF(d_1) * FF * TF'$

$= [-6.155 \quad 2.7835] \; TF'$

$= [5.9749 \quad 3.1487 \quad 0.0152 \quad -0.1247]$

Yes, it is similar to original d_1.

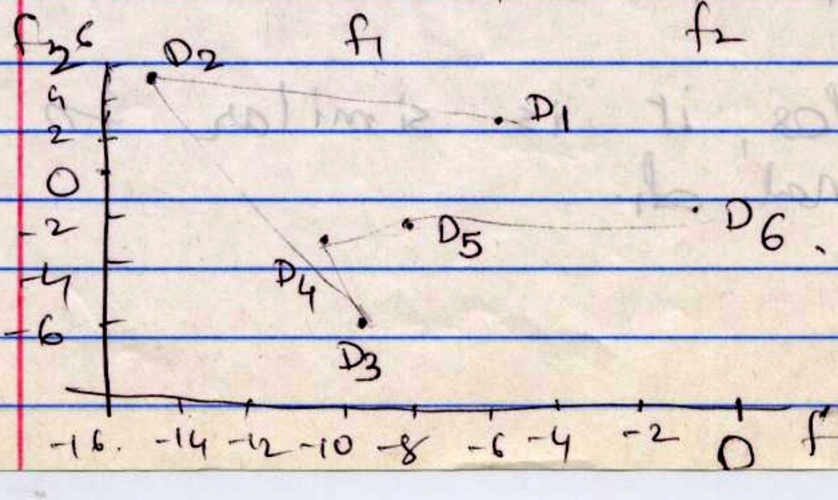
(3)

After keeping two important dimensions.

We get
$DF^{*}FF$

$$= \begin{bmatrix} -0.2638 & 0.285 \\ -0.6627 & 0.6018 \\ -0.4237 & -0.6079 \\ -0.4293 & -0.3061 \\ -0.3549 & -0.2171 \\ -0.0373 & -0.2151 \end{bmatrix} \begin{bmatrix} 23.334 & 0 \\ 0 & 9.760 \end{bmatrix}$$

$$= \begin{array}{c} \begin{bmatrix} -6.1555 & 2.7835 \\ -15.4634 & 5.8776 \\ -9.8866 & -5.\cancel{8776}\,9372 \\ -10.0173 & -2.9896 \\ -8.2812 & -2.1204 \\ -0.8704 & -2.1508 \end{bmatrix} \\ \quad f_1 \qquad\qquad f_2 \end{array}$$



(Kalahari, Bush)

Here D_1 & D_2 are connected.

And

D_3, D_4 and D_4, D_5, D_6 are connected. (Bush, Saddam)

4) In vector space

D_5	7	1	6	0
D_6	0	0	0	4

For D_5 $\omega_{11} = 7/7 * \ln(6/5)$
$= 0.18$

$\omega_{12} = 1/7 * \ln(6/4)$
$= 0.06$

$\omega_{13} = 6/7 * \ln(6/4)$
$= 0.36$

$\omega_{14} = 0$

$D_5 = (0.18 \quad 0.06 \quad 0.36 \quad 0)$

For D_6.

$w_{21} = 0$
$w_{22} = 0$
$w_{23} = 0$
$w_{24} = 4/4 * \ln(6/4)$
$= 0.42$

$D_6 = (0 \quad 0 \quad 0 \quad 0.42)$

$\mathrm{Sim}(D_5, D_6) = \dfrac{D_5 \cdot D_6}{|D_5|\,|D_6|}$

$= \dfrac{0}{}$

(Reduced 2D LSI from previous prob)

$D5 = -8.2812 \quad -2.1204$
$D6 = -0.8704 \quad -2.1008$

$\mathrm{SimLSI}(D_5, D_6) = \dfrac{D_5 \cdot D_6}{|D_5|\,|D_6|} = \dfrac{7.21 + 4.45}{(8.55)(2.27)}$

$= \dfrac{11.66}{8.55 \times 2.27}$

$\boxed{= 0.6}$

Yes. Because in D_3 & D_4 when Term Iraq occurs Saddam also occurs. So Saddam & Iraq have high correlation.

So D_5 contains Iraq & D_6 contains Saddam so they are correlated.

5) In original space.

$q = (0,0,0,1)$ &
$D_5 = (0.18, 0.06, 0.36, 0)$ from ex. 4

$$\text{So } sim(q, D_5) = \frac{q \cdot D_5}{|q|\,|D_5|}$$
$$= \frac{0+0+0+0}{|q|\,|D_5|}$$
$$= 0.$$

In reduced 2-D LSI

DF* FF for D_5

$= (-8.2812 \quad -2.1204)$.

for $q^* TF$

$q = (0,0,0,1)$ & $TF = \begin{bmatrix} -0.8817 & 0.1969 \\ -0.2887 & 0.4428 \\ -0.3033 & -0.6652 \\ -0.2173 & -0.5253 \end{bmatrix}$

$q^* TF$

$= [-0.2173 \quad -0.5253]$

$$\text{Sosim}(D_5, q) = \frac{D_5 \cdot q}{|D_5||q|}$$

$$= \frac{+8.2812 * 0.2173 + 2.1204 * 0.5253}{(8.55)(0.57)}$$

$$= 0.6$$

It shows high similarity between D_5 & q though D_5 doesn't contain word "Saddam" Because D_3 & D_4 Contain Saddam & Iraq both, so Saddam & Iraq are highly correlate

~~As~~ D_5 contain Iraq & query Contains Saddam so D_5 & query are similar.

Q4) For the database/regression example, we have

TFIDF Matrix: (for the d-t matrix)

2.5300	14.5600	4.6000	0	0	2.0700
3.3700	6.9300	2.5540	0	1.1000	0
0.1300	11.0900	2.5500	0	0	0
0.6300	4.8500	1.0200	0	0	0
4.5300	21.4900	10.2200	0	1.0700	0
0.2100	0	0	12.4700	2.4900	11.0900
0	0	0.5100	22.1800	4.2800	0
0.3200	0	0	15.2500	1.4300	1.3900
0.1100	0	0	23.5600	9.6300	17.3300
0.6300	0	0	11.7800	1.4300	15.9400

```
>> [U S V]= svd (TFIDFmatrix )
```

U =

-0.0312	0.4807	0.0831	0.1714	-0.2121	0.2958	-0.3434	-0.4567	0.4672	-0.2335
-0.0106	0.2449	-0.0154	-0.2083	0.3813	0.7720	0.0953	0.0172	-0.3784	0.0481
-0.0063	0.3451	-0.0220	0.1507	-0.7235	0.0051	0.0172	0.1089	-0.5503	0.1382
-0.0029	0.1530	-0.0090	0.0592	-0.2076	0.2290	0.6468	0.3988	0.5256	0.1545
-0.0203	0.7517	-0.0545	-0.1209	0.3857	-0.4889	0.0470	0.1577	-0.0291	0.0365
- 0.3587	-0.0109	0.2195	0.2337	0.0501	-0.0459	0.3319	0.0427	-0.1920	-0.7842
-0.4296	-0.0234	- 0.6934	0.0610	0.0146	-0.0548	0.2975	-0.4685	-0.0348	0.1341
-0.3027	-0.0188	-0.3974	0.3861	0.0928	0.1402	-0.4722	0.5838	0.0813	-0.0477
-0.6632	- 0.0201	0.2094	-0.6468	-0.2010	-0.0069	-0.1699	0.1123	0.1070	0.0651
-0.3921	-0.0065	0.5086	0.5106	0.2062	-0.0154	0.0396	-0.1434	-0.0505	0.5093

S =

45.9164	0	0	0	0	0
0	32.1346	0	0	0	0
0	0	15.8521	0	0	0
0	0	0	4.6348	0	0
0	0	0	0	3.3288	0
0	0	0	0	0	1.9763
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0

V =

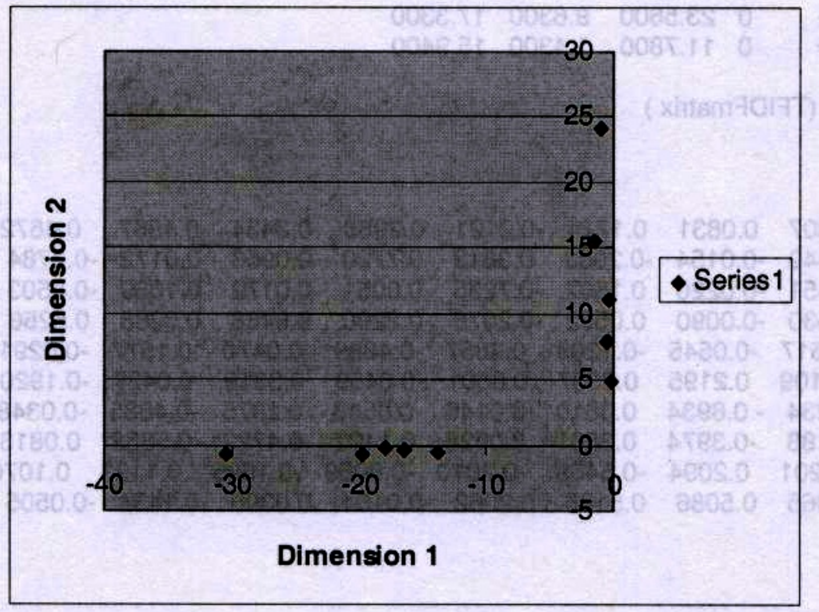
-0.0153	0.1734	0.0104	-0.0725	0.7267	0.6605
-0.0229	0.9155	-0.0224	0.0888	-0.3563	0.1613
-0.0134	0.3592	-0.0399	-0.1087	0.5682	-0.7311
-0.8464	-0.0464	-0.4907	0.2010	0.0174	0.0032
-0.2210	0.0223	-0.0203	-0.9627	-0.1467	0.0451
-0.4836	0.0123	0.8698	0.0891	0.0146	-0.0344

Values for taking the top-2 dimensions:

-1.4346	15.4459
-0.4873	7.8706

-0.2898	11.0911
-0.1342	4.9157
-0.9344	24.1541
-16.4712	-0.3505
-19.7249	-0.7509
-13.9002	-0.6034
-30.4513	-0.6467
-18.005	-0.2096

Plot for the above example:



LSI is scale sensitive. so on using tf idf weights as input to LSI the eigen vectors change.