Qn III. Consider the following T-D matrix defining 6 documents defined in terms of 4 keywords.

	D1	D2	D3	D4	D5	D6
Bush	5	15	7	9	7	0
Kalahari	5	7	1	0	1	0
Iraq	1	0	7	4	6	0
Saddam	0	1	6	4	0	4

We decide to reduce the noise and dimensionality of this data through SVD analysis The SVD of this T-D matrix, according to MATLAB is: tf x ff x df^T where tf, ff and df are are given by:

$$\begin{matrix} 0.8817 & 0.1969 & -0.0444 & -0.4264 \\ 0.2887 & 0.4928 & 0.1190 & 0.8122 \\ 0.3033 & -0.6652 & -0.5674 & 0.3790 \\ 0.2173 & -0.5253 & 0.8136 & 0.1222 \end{matrix} \quad \mathbf{tf}$$

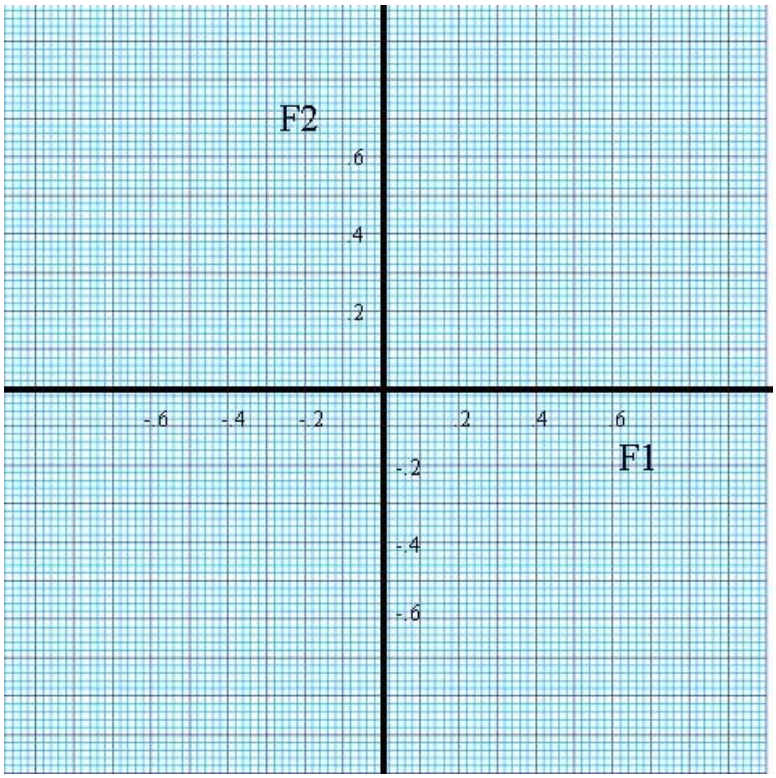
$$\begin{matrix} 23.33 & 0 & 0 & 0 & 0 & 0 \\ 0 & 9.76 & 0 & 0 & 0 & 0 \\ 0 & 0 & 5.03 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3.27 & 0 & 0 \end{matrix} \quad \mathbf{ff}$$

$$\begin{matrix} 0.2638 & 0.6627 & 0.4237 & 0.4293 & 0.3549 & 0.0373 \\ 0.2850 & 0.6018 & -0.6079 & -0.3061 & -0.2171 & -0.2151 \\ -0.0385 & 0.1948 & 0.1425 & 0.1162 & -0.7138 & 0.6460 \\ 0.7038 & -0.1795 & 0.3700 & -0.5590 & 0.0308 & 0.1491 \\ 0.5557 & -0.3294 & -0.1526 & 0.6077 & -0.3198 & -0.2965 \\ -0.2090 & 0.1411 & 0.5201 & -0.1635 & -0.4629 & -0.6519 \end{matrix} \quad \mathbf{df}$$

(1) [3pt] Suppose we are willing to sacrifice upto a maximum of 10% of the total variance in the data, then what is the least number of dimensions we need to keep? Explain how you arrived at your answer.

(2) [4pt] Suppose we decided to just keep top two most important dimensions after the LSI analysis. Draw a bounding box around the parts of **tf, ff** and **df** matrices above that will be retained after this decision. [You answer this question by directly marking the matrices above]

(3) [6pt] Suppose the two most important dimensions after LSI are called *f1* and *f2* respectively. Plot the six documents as points in the factor space (use the plot below). (It is okay if you put the points in the rough place they will come; no need to spoil your eyesight counting all the small grid lines). *Comment on the way the documents appear in the plot—is their placement related in any rational way to their similarity you would intuitively attach to them?*



(4) [5] What is the vector space similarity between D5 and D6 *before* and *after* the LSI transformation (assume, in the latter case, that we are using the top two dimensions). Is the change intuitively justified?

(5) Suppose I have the query q="Saddam". Compute the similarity of q to document D5 both in the original space and in the reduced 2-D LSI space. Comment on the results.