1.

(a) Precision is meaningless unless compared to the level of the Recall desired by the user. The lawyers and the paralegals who were to use the System for litigation support Stipulated that they must be able to retrive at least 75 percent of the all the document relevant to a given request for information. So it is very necessary for lawyers and paralegals to revise the queries as many times as they want to ensure they can retrieve at least 75 percent of all the documents.

(b)

(i) True, the Recall and precision of Lawyer 1 is 22.7% and 76.0% respectively, and the Recall and precision result of Lawyer 2 is 18% and 81.4%. Although there is Some difference between the results for each lawyer, the variance is not statistically significant at 5% level, the experiment the results were independent of the particular user involved.

(ii) False, the paper has hypothesized that if the values of Recall and Precision for the requests where significantly different from the overall mean values we might be able to infer something about the requesting procedure. But unforturotely, the values for recall and precision for the substantially revised queries (23.9% and 62.1% respectively) did not indicate a statistically significant difference.

(iii) False, Although there is a marked improvement in the lawyers' average Recall for all five information requests. the improment is not statistically significant at the 5% level. So we can't say that lawyers can get significantly better precision and recall than paralegals

(C) They chose subsets of the overall document collection using the idea of stratified sampling. They choose subsets of unretrived databases Randomly and remove the duplicate document among subsets and keep only one. The sample frames organized by the processed subsets from the unretrived database.

(d) Full-text retrieval can only work well only on unrealistically small databases, but in practice the databases are so large that this method does work. One of reasons is very simple: Words and phrases have many close synonyms in documents, we can't find them by searching the limited key words.

2. (a) $tf_{max} = 5$ $W_{t1.D} = \frac{2}{5} \cdot idf = \frac{2}{5} \log_2 \frac{10}{1} = 1.33$

$W_{t2.D} = 0$

$W_{t3.D} = \frac{1}{5} \cdot idf = \frac{1}{5} \log_2 \frac{10}{3} = 0.35$

$W_{t4.D} = 1 \cdot \log_2 \frac{10}{2} = 2.32$

$W_{t5.D} = \frac{2}{5} \cdot \log_2 1 = 0$

$\therefore$ the document vector is $(1.33, 0, 0.35, 2.32, 0)$

(b) $Q = <1, 0, 0, 1, 0>$

$sim(D,Q) = \sum_{k=1}^{5} d_k q_k = 1.33 + 2.32 = 3.65$

$|Q| = \sqrt{\sum_{k=1}^{5} q_k^2} = \sqrt{2} = 1.414$

$|D| = \sqrt{\sum_{k=1}^{5} D_k^2} = 2.70$

$cosSim(D.Q) = \frac{3.65}{1.414 \times 2.70} = 0.96$