CSE 586 Midterm Exam

Olcay Taner YILDIZ

I. Question

Consider these documents:

Doc1: breakthrough drug for schizophrenia
Doc2: new schizophrenia drug
Doc3: new approach for treatment of schizophrenia
Doc4: new hopes for schizophrenia patients

1) Draw the term-document incidence matrix fort his document collection.
2) Draw the inverted index representation for this collection.

II. Question

How should the Boolean query x AND NOT y be handled? Why is naive evaluation of this query normally very expensive? Write out a postings merge algorithm that evaluates this query efficiently.

III. Question

Shown below is a portion of a positional index in the format: term: doc1: <position1, position2, . . . >; doc2: <position1, position2, . . . >; etc.

```
angels: 2: <36,174,252,651>;
4: <12,22,102,432>; 7: <17>;
fools: 2: <1,17,74,222>;
4: <8,78,108,458>; 7: <3,13,23,193>;
fear: 2: <87,704,722,901>;
4: <13,43,113,433>; 7: <18,328,528>;
in: 2: <3,37,76,444,851>;
4: <10,20,110,470,500>; 7: <5,15,25,195>;
rush: 2: <2,66,194,321,702>;
4: <9,69,149,429,569>; 7: <4,14,404>;
to: 2: <47,86,234,999>;
4: <14,24,774,944>; 7: <199,319,599,709>;
tread: 2: <57,94,333>;
4: <15,35,155>; 7: <20,320>;
where: 2: <67,124,393,1001>;
4: <11,41,101,421,431>; 7: <16,36,736>;
```

Which document(s) if any match each of the following queries, where each expression within quotes is a phrase query?

1) ”fools rush in”
2) ”fools rush in” AND ”angels fear to tread”

IV. Question

If $|s_i|$ denotes the length of string s_i, show that the edit distance between s_1 and s_2 is never more than $\max\{|s_1|, |s_2|\}$.

V. Question

Compute variable byte codes for the postings list 777, 17743, 294068, 31251336. Use gaps instead of postings where appropriate. Give the solution for variable bytes as a sequence of 8-bit blocks.