INFO-533 Homework 1
(Due: February 14, 2023 in class)

Please remember to include the following at the beginning of your submitted assignment and SIGN it. Without, your assignment will not be graded. "I have done this assignment completely on my own. I have not copied it, nor have I given my solution to anyone else. I understand that if I am involved in plagiarism or cheating, I will have to sign an official form that I have cheated and that this form will be stored in my official university record. I also understand that I will receive a grade of 0 for the involved assignment and my grade will be reduced by one level (e.g., from A to A- or from B+ to B) for my first offense, and that I will receive a grade of "F" for the course for any additional offense of any kind."

1. Consider the following four documents:

   D1: new auto sales top forecasts
   D2: auto sales rise in october
   D3: increase in auto sales in october
   D4: october new auto sales rise

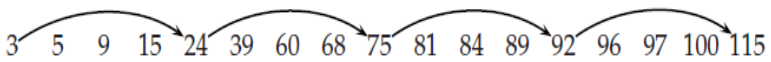
   (a) (8 points) Draw the term-document incidence matrix for this document collection (sort terms in alphabetic order).
   (b) (24 points) Manually build the inverted index for this document collection following the steps below: (i) create a sequence of (word, docID) pairs for all documents; (ii) sort the pairs first by words and then by docIDs and remove redundant pairs (if any); and (iii) create the inverted index (don't forget to include document frequency). Show the results after each step (see slides in lecture notes of Chapter 1 for appropriate formats).

2. (20 points) Recommending a query processing order for query: (tangerine OR trees) AND (marmalade OR skies) AND (kaleidoscope OR eyes), given the following postings list sizes (the number in each pair is the size): (eyes, 1800), (kaleidoscope, 970), (marmalade, 1000), (skies, 2600), (tangerine, 460), (trees, 3000). Provide a brief justification for your recommendation. [Hint: Use worst case scenario to estimate the size of each OR subquery.]

3. (14 points) Use examples to explain why stemming tends to boost recall but hurts precision.

4. (12 points) Explain why skip pointers are more useful for AND Boolean queries (e.g., "t1 AND t2") than OR queries (e.g., "t1 OR t2"?).

5. (12 points) Consider a postings intersection between this postings list:



   and the following intermediate result postings list (which hence has no skip pointers): 2 9 84 90 118. Point out where and when a skip pointer in the first postings list is followed when the two lists are being intersected (i.e., merged by AND). Explain why for each skip.

6. (10 points) Shown below is a portion of a positional index in the format: term: docID: <position1, position2, . . . >; docID: <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>;

Identify document(s) that satisfy the following phrase query: “fools rush in”.