Due: 2019/3/15

Homework 2

Student Number:
Name:

Problem 1. (20 points) Consider the following fragment of a positional index with the format:
word: document:<position, position, . . .>; document: <position, . . .>
Gates: 1: <3>; 2:<6>; 3: <2,17>; 4: <1>;
IBM: 4: <3>; 7: <14>;
Microsoft: 1: <1>; 2: <1,21>; 3: <3>; 5 :<16,22,51>;

The /k operator, word1 /k word2 finds occurrences of word1 within k words of word2 (on either side), where k is a positive integer argument. Thus k = 1 demands that word1 be adjacent to word2.

a. Describe the set of documents that satisfy the query Gates /2 Microsoft.

b. Describe each set of values for k for which the query Gates /k Microsoft returns a different set of documents as the answer.

Problem 2. (30 points) Given two strings S_1 and S_2, write down the pseudo-code of computing the edit distance between them.

Problem 3. (30 points) If you wanted to search for *s*ng* in a permuterm wildcard index, what key(s) would one do the lookup on?

Problem 4. (20 points) Write the pseudo code showing the details of computing the Jaccard coefficient while scanning the posting of the k-gram index. (Page 49 in the slide)